Accelerating DSPM Scanning: How Forcepoint Overcomes API Throttling

Forcepoint

Whitepaper

forcepoint.com

Table of Contents

Introduction

Modern enterprises often face an overwhelming volume of data spread across cloud and on- premises repositories. A Data Security Posture Management (DSPM) solution must scan and classify millions of files to uncover sensitive information and risks. However, a common challenge in cloud environments (like Microsoft OneDrive and SharePoint Online) is API request throttling.

API request throttling refers to the process by which cloud providers limit how many calls can be made in a given timeframe. This throttling causes many DSPM tools to scan slowly or stall when faced with tens of millions of files. In fact, one analysis noted that for an organization of ~4,000 users, Microsoft's API limits would only allow scanning about 240 GB of data per day (roughly 2.4 million file calls per day). At that rate, scanning an environment of 100 TB "would take forever."

Forcepoint DSPM takes a different approach to solve this problem. Instead of attempting deep content inspection on every file and immediately running into API limits, Forcepoint employs a two-phase strategy: first rapidly cataloging all data with minimal API calls, then performing targeted deep scans on high-priority data. This approach, combined with an AI- driven classification engine, enables blazing-fast scanning speeds without overwhelming the APIs. The result is a DSPM solution that can handle enterprise-scale data while delivering highly accurate insights.
This whitepaper details how Forcepoint achieves these efficiencies through smart engineering and why that matters to data leaders seeking to protect their data more efficiently.

Challenges of Traditional DSPM Scanning

Scanning cloud data at scale is inherently difficult. Platforms like Microsoft 365 impose limits to protect service performance. For example, Microsoft Graph API caps the number of calls an app can make in a day based on tenant size (e.g., ~2.4 million calls per day for a mid- sized tenant). A typical "naïve" DSPM scan that opens and inspects every file's content could easily require tens of millions of API calls to cover an entire enterprise, far exceeding these limits. When those limits are hit, the cloud service will start throttling (delaying or blocking further requests), causing the scan to drag on interminably. This is why initial data discovery with some tools can take weeks or even months; the scanner must constantly pause due to rate limits.

Moreover, a broad, unfocused scan often yields an unmanageable flood of data classification results. Scanning everything indiscriminately means security teams get noise along with the signal: thousands of alerts on trivial data or false positives that obscure the truly critical risks.

Several vendors in the DSPM space have struggled with this balance. For instance, legacy scanning tools often produced many false negatives or missed data due to limited classifiers and slow, "one-size-fits-all" scanning methods. Customers of such tools report frustration not only with the slow speed but also with lack of accuracy and actionable insights. The scope is so broad that the results aren't prioritized or tuned to what the business cares about.

In summary, traditional approaches to DSPM face a double-edged problem: performance bottlenecks from API throttling and overwhelmed alerting due to overly broad scans. A new approach is needed to drastically speed up discovery while focusing on what matters most.

Forcepoint's Two-Phase High-Speed Scanning Methodology

Forcepoint DSPM addresses these challenges with a two-phase scanning methodology designed for speed and efficiency:

Phase 1

Rapid Data Cataloging: First, Forcepoint performs a high-speed discovery scan that catalogs all files and folders across the enterprise data landscape. Crucially, this phase does not download or deeply inspect every file's content. Instead, it uses optimized API calls to gather metadata file names, sizes, types, paths and other attributes to enumerate file locations and permissions. Because one API request can retrieve many files' metadata (for example, a single call can list a whole folder of hundreds of files), this approach drastically reduces the number of API calls needed compared to deep-content scanning.

The result is an extremely fast inventory: Forcepoint DSPM can scan approximately 300 files per second, or about 1 million files per hour, on a single scanning node. Over a full day, that equates to 24+ million files; indeed, real-world tests have shown on the order of ~30 million files per day handled by one node. This "cataloging" scan runs in parallel across multiple data sources (cloud apps like AWS S3, SharePoint, OneDrive, Google Drive, as well as on-premises file shares) to cover the entire environment. By the end of Phase 1, the organization attains a complete catalog of where data lives and basic information about each file achieved in a fraction of the time a traditional scan would take.

Phase 2

Targeted Deep Content Analysis: Next, Forcepoint uses the insights from the catalog to prioritize targeted content scans on the subsets of data that pose the highest risk. Rather than immediately inspecting every file, Forcepoint DSPM pinpoints which data sets are likely to contain sensitive information: for example, files in locations known to hold Personally Identifiable Information (PII), financial records subject to PCI compliance, intellectual property or any "crown jewels" as defined by the client. Those high-priority files (and only those files) are then subjected to deep content inspection using Forcepoint's AI-powered classification engine.

By focusing the intensive analysis on specific high-risk areas identified in Phase 1, Forcepoint dramatically reduces the total workload and API calls needed. This means the deeper scans can run without "eating up" all the available API headroom.

In effect, Phase 2 acts like a scalpel instead of a shotgun: it zooms in on what's important. The vast majority of files that are low-risk (e.g., system files, archives of trivial data, redundant copies) can be scanned more lightly or deferred, ensuring that the cloud APIs are not overwhelmed by millions of unnecessary content-download requests. The outcome is a much faster time-to-insight: Forcepoint DSPM quickly surfaces the risky data first, rather than trudging file-by-file equally. And because Phase 2 is guided by Phase 1's catalog, it's an informed process in which the system "knows" where the sensitive data is likely to be, making each content scan count.

This two-phase approach is akin to using a funnel for data discovery. In Phase 1, the funnel's wide mouth rapidly collects everything (with minimal cost per item), giving breadth of visibility and in Phase 2, the narrow spout concentrates on the critical items giving depth where needed. Many in the industry recognize this as the only scalable way to handle huge data volumes, but Forcepoint has operationalized it in a way that is seamless to the customer. By separating broad discovery from deep analysis, Forcepoint DSPM avoids the pitfall of "promise everything upfront" (which often leads to slow, throttled scans and shallow results). Instead, it delivers quick wins and then progressively deeper insight in a prioritized manner.

Notably, Phase 1 is not just a simple file listing; even during the rapid cataloging, Forcepoint DSPM extracts valuable security insights. The system will flag, for example, files that are likely sensitive even before content analysis by looking at metadata patterns (such as files with certain keywords in names or certain locations) and by applying lightweight detectors. This is achieved without heavy content processing.

The organization benefits from an initial Data Risk Assessment right after Phase 1, a quick overview of where the biggest piles of data are, which areas are most exposed and where obvious cleanup (like deleting ROT data or tightening permissions) can reduce risk.

With the catalog in hand, Phase 2 proceeds to perform content classification scans on the high-priority segments. These targeted scans are orchestrated to respect API limits. For example, the scanner can be scheduled during off-peak hours or tuned to a pace that stays below Microsoft's throttling thresholds. Because the volume of files for a deep scan is much smaller than the total, the API budget is sufficient to handle them. This means no throttling interruptions even as sensitive files are thoroughly inspected. In essence, Forcepoint's method sidesteps the bottleneck: it never tries to push the cloud API beyond what it's built to handle. Phase 1 uses bulk-efficient calls, and Phase 2 limits intensive calls to a manageable subset – a stark contrast to a brute-force scan that would trigger Microsoft's defenses almost immediately.

AI Mesh: Efficient and Accurate Data Classification

A cornerstone of Forcepoint DSPM's advantage is its AI Mesh technology: a sophisticated, AI-driven file classification pipeline that operates during the content analysis (Phase 2) and even in parts of Phase 1. The AI Mesh is designed to achieve high accuracy in identifying sensitive data without the heavy performance cost of traditional content scanning.

What is the AI Mesh? In simple terms, it's a network of multiple machine learning models and detectors working in concert to analyze file content. When a file is scanned, Forcepoint doesn't rely on just one method (like a regex pattern matcher or a single ML model); instead, it employs a variety of AI components each looking at the data from different angles.

The AI Mesh includes:

- "LLM-like" language models that convert file text into semantic vectors (capturing the context and meaning of the text)
- Deep neural network classifiers (e.g., for things like sentiment or tone) that evaluate those vectors
- Topic detection models (e.g., simpler Bag-of-Words approaches) that identify themes or subjects in the text
- Keyword/regex filters for specific patterns (like credit card numbers or social security numbers)
- Other evaluators (such as text complexity or presence of certain keywords like "confidential") that are custom-coded
- A Bayesian inference layer that takes all the signals from the above models and fuses them to make a final determination

This "mesh" of AI ensures that the classification is context-aware and robust. For example, it can differentiate a list of "ingredients" from a company's secret "recipe" even if both contain similar terms by understanding context and intent, something Forcepoint notes as a strength of its 50-dimensional ML classification model. In practice, that means fewer false alarms: trivial or benign content is less likely to be misclassified as sensitive simply because it contains a certain keyword, and truly sensitive content (even if it doesn't match a simple pattern) can be recognized by its context.

The elegance of this sophisticated multi-model approach is that the AI Mesh is engineered to be highly efficient – delivering superior outcomes while requiring lower resource consumption. The models it uses are relatively small and optimized. In fact, the AI models are 10x smaller than even the smallest typical large-language models, enabling the system to classify a file in roughly 200 milliseconds using a normal CPU, no GPU acceleration needed. This speed is remarkable: it means that even when we do have to inspect content deeply in Phase 2, the analysis itself does not become a bottleneck. Hundreds or thousands of files can be processed in parallel in a matter of seconds. The lightweight nature of the AI also means Forcepoint can deploy it in flexible ways (even on-premises) without requiring special hardware.

Accuracy is where the AI Mesh really shines. Forcepoint's machine learning engine continuously learns and adapts to the organization's data. It starts with a rich base: a "massive number of dimensions" in its model and even industry-specific AI sub-models (for healthcare, finance, etc.) that feed into the broader cloud AI model. On top of that, when Forcepoint DSPM is deployed for a customer, it can spin up specialized AI models just for that organization (e.g., an on-premises model component) which then train on the customer's data to further improve accuracy. This federated learning approach means the AI Mesh becomes increasingly fine-tuned to what your sensitive data looks like, whether that's a certain format of client IDs, proprietary project code names or other contextual cues unique to your business. The result is an exceptionally high detection accuracy with far fewer false positives.

In summary, the AI Mesh is the engine that powers Forcepoint's content scanning phase to be both fast and accurate. It ensures that when Phase 2 of scanning kicks in, the system isn't just blindly reading files – it's intelligently interpreting them. This yields a rich, contextual understanding of your data-at-rest. For example, it can tell the difference between a file containing random 16-digit numbers and one containing real credit card numbers tied to personal info, or distinguish a confidential design document from a generic template. This intelligence directly addresses the earlier point about other vendors having "broad scope but lack of accuracy." Forcepoint's targeted use of AI means broad scope with high accuracy is the ultimate goal for DSPM.

Overcoming API Throttling and Performance Limits

Forcepoint DSPM's architecture is explicitly designed to coexist gracefully with cloud provider limits rather than fight them. Here's how Forcepoint overcomes the throttling and performance constraints that plague other solutions:

→ Minimized API Calls via Metadata-Only Scanning

→ Adaptive Throttling Management

→ Parallel and Distributed Processing

→ Incremental and Continuous Scanning

→ No-Content-Transfer Design

Key Definitions

→ Discovery: Initial identification of files and data sources

→ Cataloging (Metadata Classification): Lightweight analysis of file metadata including path, permissions and labels

→ Content Classification: Deep inspection of file contents using regex, ML models and attribute detectors

→ Smart Scanning: A prioritization strategy that focuses deep inspection on files flagged as sensitive during cataloging

Throughput Metrics (Normalized)

Assuming 1 petabyte (PB) of data consists of 1,000,000,000 MB (1 PB = 1,000 × 1,000 × 1,000 MB), and each file averages 5MB:

→ Total Files per PB: 1,000,000,000 MB ÷ 5 MB/file = 200,000,000 files

Cataloging (Metadata Classification)

→ Peak Speed: 992 files/sec

→ Per Hour: 3.57M files/hour

→ Per Day: 85.7M files/day

→ Petabytes/day: ~0.43 PB/day

→ Days to Process 1 PB: ~2.33 days

Cataloging (Metadata Classification)

→ Peak Speed: 12.2 files/sec

→ Per Hour: 43,920 files/hour

→ Per Day: 1.05M files/day

→ Petabytes/day: ~0.0053 PB/day

→ Days to Process 1 PB: ~189 days

Standalone Classification

→ Peak Speed: 13.9 files/sec

→ Petabytes/day: ~0.006 PB/day

→ Days to process 1 PB: ~167 days

Discovery + Classification

→ Peak Speed: 10.3 files/sec

→ Petabytes/day: ~0.00445 PB/day

→ Days to Process 1 PB: ~225 days

Discovery + Classification

→ Upgrading infrastructure from t3a.2xlarge to m5a.4xlarge (8 CPUs → 16 CPUs, 32GB → 64GB RAM):

→ Cataloging Speed: 253 → 992 files/sec (3.9x increase)

→ Classification Speed: 3.6 → 12.3 files/sec (3.4x increase)

Smart Scanning Strategy

Cataloging acts as a triage layer:

→ Quickly shows overexposed or sensitive files using metadata

→ Flags these files for deep content inspection

→ Reduces overall system load by avoiding unnecessary deep scans

This approach mirrors medical triage: simple checks first, followed by deeper diagnostics only when needed.

System Resilience and Load Testing

- → Queue-based task allocation ensures Edge components only process what they can manage
- → Apache Flink backpressure dynamically regulates data flow to prevent overload
- → Edge CPU usage peaks at ~85% under load but remains stable

Bottlenecks Identified

- → ML Classification: High impact on performance
- → Dynamic Attribute Detection: High impact
- → Content Detectors: Low impact

In combination, these techniques allow Forcepoint DSPM to fly under the radar of API throttling while still achieving thorough coverage. It's a delicate balance of being aggressive in speed but polite in API usage – something Forcepoint's engineers have fine-tuned.

Forcepoint DSPM demonstrates scalable and resilient performance across discovery, cataloging and classification phases. Smart scanning significantly optimizes resource usage by focusing deep inspection on high-risk files. Hardware scaling and batching strategies further enhance throughput, making the platform suitable for large-scale enterprise environments.

Delivering Speed and Accuracy for Better Risk Outcomes

Speed is only truly valuable if the insights you gain at that speed are actionable. One of the criticisms of first-generation DSPM or data discovery tools was that they'd eventually find a lot of issues, but by the time they did, the data might have changed, or the results were so noisy that the security team didn't know where to start.

Forcepoint's fast, focused approach avoids those pitfalls and enhances accuracy and relevance of results:

→ Prioritized Risk Identification for Data Asset Classes: By focusing on sensitive data first, Forcepoint DSPM ensures that the most critical exposures are uncovered early in the process.

→ Reduction of False Positives/Negatives: As the AI Mesh's high accuracy significantly cuts down on false positives and false negatives compared to simplistic scanning methods.

→ Comprehensive Visibility with Context: One of Forcepoint DSPM's strengths is that even as it speeds up discovery, it does not sacrifice context. The platform provides a rich dashboard and reports that give a "bird's-eye view" of the data environment and risk hotspots. You not only learn what sensitive data exists, but also where it is, who has access to it and how it's shared.

→ Remediation and Workflow Integration: Speed and accuracy in finding issues are step one; step two is fixing them. Forcepoint DSPM includes built-in orchestration for remediation workflows (e.g., automatically or semi-automatically removing overly exposed links, revoking access and triggering encryption or stubbing on certain files) What's important here is that because Forcepoint has carefully identified the most important issues, those remediation actions can be laser-focused. The platform allows custom playbooks. For instance, you can set a rule such as: "If a file with PCI data is found in an open SharePoint site, immediately notify the data owner and quarantine the file while leaving a file stub." Forcepoint's accuracy ensures this won't be a trigger-happy process that disrupts benign files. In contrast, if one tried automated remediation with a high false-positive tool, it could cause chaos by locking down files that weren't truly sensitive. Thus, Forcepoint's reliable scanning enables safe automation of data governance tasks, multiplying the value of the speedy discovery.

Conclusion

For large organizations struggling with slow and inaccurate data security scans, Forcepoint's DSPM offers a transformative solution. By re-thinking the scanning process – splitting it into a rapid cataloging phase and a focused deep analysis phase – Forcepoint avoids the common traps of API throttling that bedevil other tools. This speed, however, is not a trade-off against quality; thanks to the AI Mesh technology and a risk-prioritized approach, the data classification is highly precise and tuned to what you care about most.

Forcepoint directly addresses other DSPM provider pain points: the clever use of metadata cataloging ensures minimal throttling and fast completion, and the tailored AI-driven scanning ensures high accuracy and relevancy of findings. The platform empowers security executives with a clear map of their data revealing where the most sensitive information resides, who has access to it and what risks are associated with it all within weeks of deployment. From there, robust reporting and workflow tools translate insight into action, whether it's compliance reporting or automated risk remediation.

To summarize the key advantages:

- → Unmatched Scanning Performance: Up to tens of millions of files per day per node, leveraging scalable architecture, so you can quickly get visibility across all your cloud and on-prem data stores.
- → No Throttle Headaches: Designed in accordance with cloud API limits and best practices, meaning the technology, not the limitations, dictates the timeline of your data discovery project.
- → Intelligent Focus on What Matters: A strategy that doesn't boil the ocean. You see results for critical data (PII, PCI, IP, etc.) first, enabling a risk-based approach to data security from day one.
- → AI-Powered Accuracy: An AI Mesh that continuously learns and delivers highly accurate classification, drastically cutting down false positives/negatives. This accuracy has been proven to outperform legacy methods, giving you confidence in the findings.
- → End-to-End Data Protection: Integration with Forcepoint's broader data security ecosystem (like Data Detection and Response (DDR) for real-time monitoring, and Data Loss Prevention (DLP) to protect data in motion) means DSPM doesn't operate in a silo. You can discover and then protect data seamlessly, something point solutions cannot match. Even as a standalone, DSPM provides remediation workflows and reports that drive action, not just awareness.

In conclusion, Forcepoint DSPM's innovative scanning methodology and AI technology directly address the concerns of today's CISOs and CIOs. It offers a solution for the perennial DSPM dilemma: how do we quickly find and secure the critical data among everything we have, without drowning in the process? By overcoming technical limits and smartly narrowing scope, Forcepoint delivers a faster, smarter DSPM. For the large enterprise customer evaluating solutions, the message is clear: you no longer have to settle for slow or incomplete data security insights. With Forcepoint, you can rapidly shine a light on your dark data and do so with pinpoint accuracy, strengthening your data security posture in record time.

Forcepoint

forcepoint.com/contact

About Forcepoint

Forcepoint simplifies security for global businesses and governments. Forcepoint's all-in-one, truly cloud-native platform makes it easy to adopt Zero Trust and prevent the theft or loss of sensitive data and intellectual property no matter where people are working. Based in Austin, Texas, Forcepoint creates safe, trusted environments for customers and their employees in more than 150 countries. Engage with Forcepoint on www.forcepoint.com, Twitter and LinkedIn.

[FP-Whitepaper-Accelerating DSPM Scanning-US-EN] 19Dec2025