

Building a Federated DataCatalog

Per-share scanning · GUFFI DSI · DeepScan · Federated Query

Hammerspace, Inc. | Confidential



HAMMERSPACE

Hammerspace

Highly Parallel Filesystem built for AI



HAMMERSPACE



High Performance Parallel File System for AI Workloads



Superior Performance

for Training and Inference Workloads



Scales Massively,
Predictably and Easily



Standards-based

with No Proprietary Agents
or Modified Clients



Unparalleled End-to-End

Time-to-First-Token Performance



Massive
Scale



Maximum
Throughput



Low
Latency



Global
Access



Built for
AI



TRAINING
(GPU CLUSTERS)



DATA
ANYWHERE



INFERENCE
(AI FACTORIES)

Data Orchestration for Sites and Systems



Policies to Move Data
Where You Need It -
When You Need It



Orchestrate Data
to Different Storage
Systems



**Move Data From
Remotes Sites**
Regardless of Location



Utilize Remote Data
in Other Data Centers
or Clouds



Extended Metadata Compliance Classification Data Movement



HAMMERSPACE

Hammerspace Extended Metadata

Compliance is different than Classification

Compliance – *already answered by Hammerspace*

- Is the data safe stored for both availability and durability?
- is the data local or remote? If remote, where is it stored? (Livermore, Sandia, LANL, etc)
- Is the data in-flight between sites?
- Can a user store custom metadata in tags with key/value pairs?

Classification – *handled by GUF1 and the DeepScan Engine*

- Compliance officers and auditors don't ask "find me PII"; they ask risk and posture questions that join classification with location, access, retention, and movement



GUFI – Grand Unified File Index

There are many products that can index or search against filesystem data.

- None of them provide easy hooks to examine and index extended metadata
- GUFI is 10x to 100x faster because of a distributed database structure. No single database like MySQL, Postgres, etc can scale to potentially billions of files and provide query results in 1 to 3 second range
- Index extensibility. We can add additional db in the same directory structures and GUFI will scan all of them during a query. This allows us to add any type of metadata that we wish with a very light lift.



GUFI with DSI

GUFI 0.7.0 ships the DSI plugin framework — no fork.

- A Hammerspace DSI plugin reads `?.attribute=inode_info` — instead of `statx`
- One read returns POSIX + Hammerspace extended metadata together — in a single pass
- The index gains `hs_*` columns — objectives, instances, placement, tags, classification
- Incremental scans use `?.snapdiff` — only changed files — 10–100× faster



DeepScan: A Compliance Engine on the Catalog

Pluggable deep scans riding on top of GUFU.

- **Pluggable frameworks** — PII, HIPAA, GDPR, SOX, PCI — each one a plugin
- **Findings stored in tables alongside the catalog** — in the per-directory SQLite, next to GUFU's own data
- **Hammerspace's global namespace mirrors file metadata between sites automatically** — but NOT the DeepScan indices
- **So we have to replicate the findings ourselves** — the cluster never carries them



Keeping Sites in Sync

Cross-site sync is done by custom written code that mirrors what Hammerspace already does for its own metadata.

- **Each site ships its per-cycle finding deltas to S3** — and emits one event to a coordinator. Using the S3 built-in to Hammerspace will allow for deduplication, compression, and encryption. Results in much faster transfers of deltas
- **The coordinator finds the referral-share-paired sites that are online** — and fans out a notification
- **Receiving sites pull the deltas from S3** — and apply them to their local catalog via GUFF
- **content_sha256 is the sync key** — offline sites catch up from persistent per-site queues



Hammerspace defined GUFU MCP

MCP understands extended metadata as well as data classification

- **The DataCatalog MCP is globally aware** — a query can be sent to any single site
- **Every site mirrors every other site's findings** — so any site returns the same answer
- **No per-file data ever leaves the site** — queries federate out to the sites
- **"Which sites have high-severity HIPAA findings?" or "Which sites violate GDPR"** — one federated query, answered anywhere



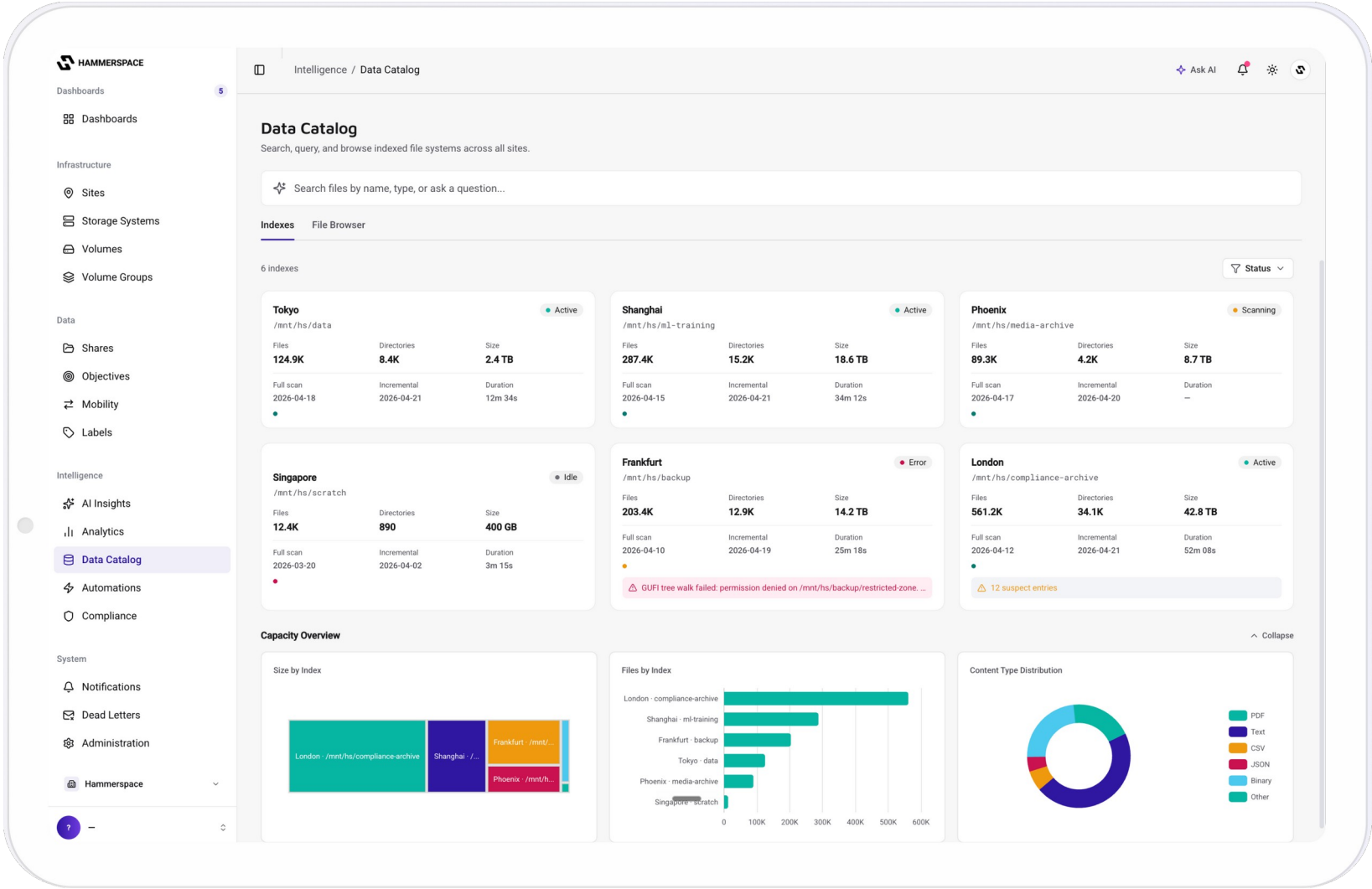
GUFI and AI



HAMMERSPACE

GUFI Indexing and Visual Builder

Query petabytes instantly.



Natural Language Queries — File Browser

AI translates intent into precise queries: *“Find all the data that is classified violation with a tag key of GDPR”*.
→ 6 matches

The screenshot displays the HAMMERSPACE Data Catalog interface. On the left is a sidebar with navigation options: Dashboards, Infrastructure (Sites, Storage Systems, Volumes, Volume Groups), Data (Shares, Objectives, Mobility, Labels), Intelligence (AI Insights, Analytics, Data Catalog, Automations, Compliance), and System (Notifications, Dead Letters, Administration, Hammerspace). The main panel is titled 'Intelligence / Data Catalog' and contains a search bar with the query 'Find all the data that is classified critical with a tag key of GDPR'. Below the search bar, a summary states 'Found 6 files for "Find all the data that is classified critical with a tag key of GDPR"'. A table lists the results, including file names, sizes, classifications, tags, and modification times.

File Name	Size	Classification	Tags	Modified
annual-report-2025.pdf /mnt/hs/data/reports/annual-report-2025.pdf	15.0 MB	Critical	PII, GDPR, SOX	3 days ago
access-log-20260421.txt /mnt/hs/user-homes/admin/logs/access-log-20260421.txt	84.9 MB	Critical	PII, GDPR, HIPAA, +1	12 minutes ago
site-inventory.xlsx /mnt/hs/data/admin/site-inventory.xlsx	554.6 KB	Critical	PII, GDPR	5 days ago
user-mapping.csv /mnt/hs/user-homes/admin/user-mapping.csv	22.9 KB	Critical	PII, GDPR, HIPAA	3 days ago
compliance-scan-results.json /mnt/hs/data/compliance/compliance-scan-results.json	446.1 KB	Critical	GDPR, SOX	9 hours ago
quarterly-revenue-summary-q1.pdf /mnt/hs/data/finance/quarterly-revenue-summary-q1.pdf	3.3 MB	Critical	GDPR, SOX, region=APAC	9 days ago



Natural Language Queries — File Browser

AI translates intent into precise queries: *“Find all .log files modified in the last 7 days”*.
→ 5 matches

The screenshot displays the 'Data Catalog' interface, specifically the 'File Browser' tab. The search bar contains the query: 'Find all .log files modified in the last 7 days'. Below the search bar, there are filters for 'All Types', 'Any Size', and 'All PII'. The search results are displayed in a table with columns: name, modified, and site. The results show 5 matches:

name	modified	site
access-log-20260421.txt	12 min ago	Shanghai
cluster-metrics-apr.csv	6 hours ago	Tokyo
raid-health-report.txt	5 hours ago	Frankfurt
compliance-scan-results.json	9 hours ago	Tokyo
hammerscript-migration.hs	21 hours ago	Tokyo

On the right side, the 'Data Catalog Assistant' panel shows the same query and a summary of the 5 most recently modified files, which matches the main results table. At the bottom right, there is a search bar labeled 'Ask anything...' and a note 'Powered by HAP'.



Additional Natural Language Queries

Finding and moving data where it is needed, when it is needed...

- Find all files within <share/directory> accessed with the last 7 days and move it to the nearest available NVMe volume to these GPUs <gpu-machine-name>
- Finde alle Dateien innerhalb von <share/directory>, auf die in den letzten 7 Tagen zugegriffen wurde, und verschiebe sie auf das nächstgelegene verfügbare NVMe-Volume zu diesen GPUs <gpu-machine-name>
- Trouve tous les fichiers dans <share/directory> qui ont été consultés au cours des 7 derniers jours et déplace-les vers le volume NVMe disponible le plus proche de ces GPU <gpu-machine-name>
- <share/directory> 안에서 지난 7 일 동안 액세스된 모든 파일을 찾아 , 이 GPU <gpu-machine-name> 에 가장 가까운 사용 가능한 NVMe 볼륨으로 이동하세요

Natural language is not limited to English. Work in the language in which you are comfortable



HIPAA – Sample AI Queries

Compliance officers ask risk questions, not “find me PHI.”

- Which PHI-containing shares are accessible to users outside the clinical workforce?
- Are there DICOM or HL7 files with patient identifiers on non-encrypted volumes?
- Show me PHI accessed by service accounts in the last 24 hours.
- What PHI exists in dev, test, or staging environments? (de-identification gap)
- Which shares hold ePHI but lack the required 6-year retention policy?



SOX – Sample AI Queries

Financial integrity, segregation of duties, change control.

- Which files in financial reporting paths were modified by users outside the finance role?
- Show me SOX-scoped data (GL, journal entries, revenue recognition) accessed by privileged or admin accounts.
- Are there SOX-classified files on shares without immutability or audit logging enabled?
- Which financial close artifacts were modified during the quarterly blackout window?
- Where do auditor working papers live, and who has read access?



GDPR — Sample AI Queries

Location, consent, retention, subject rights.

- Is there any GDPR-classified data currently stored outside the EU?
- Which shares contain EU personal data replicated to a non-adequacy country in the last 30 days?
- Show me EU PII files not accessed in 7+ years (retention violation).
- Which files reference data subject “Jane Doe” — across every site? (Article 15 DSAR)
- What personal data would be affected by an Article 17 erasure request for subject X?
- Which shares hold special-category data (health, biometric, religion) without encryption?



PII – Sample AI Queries

Exposure, location, and minimization across every site.

- Where is unmasked PII stored — across every site?
- Which PII files were created or modified in the last 24 hours?
- Show me PII residing in shares accessible to contractors or third-party accounts.
- Are there PII files in dev, test, or backup locations that should have been redacted?
- Which users or service accounts touched the most PII this week?
- Where do we have PII co-located with other regulated data (PHI, PAN, MNPI)?



Cross-Framework — Executive Dashboard

Classification × Location × Access × Time.

- What is our total exposure surface by framework and by site?
- Which shares hold data classified under multiple frameworks (e.g., PHI + PCI)?
- Show me the 10 highest-risk files in the environment right now.
- Where has classified data moved in the last 24 hours — by framework?
- Which users or service accounts touched the most regulated data this week?



Final Thought

“Classification isn't needed for AI training”

We are proud to present the fastest filesystem on the planet — and the only one that can tell you, before the training job starts, whether your corpus contains EU PHI, unmasked PANs, or material non-public information





HAMMERSPACE

Michael Kade, Sr Director – Solutions Engineering– mike.kade@hammerspace.com
Eve Wang, Solutions Engineer – UI/UX – Solutions Engineering– eve.wang@hammerspace.com