Diskover

SOLUTION BRIEF | LIFE SCIENCE

From sample to archive, every file accounted for.

One enriched metadata catalog for sequencing, imaging, and analysis data across HPC, NAS, object, cloud, and archive.

4.7 PB indexed and 0.5 PB reclaimed at one research institute

$542K reclaimed from temporary, transient, and test data within days of the first index

27M pathology files indexed across three sites at the same institute

At a glance

KEY CHALLENGES

Sequencing pipelines spawn BAM, SAM, and sorted copies that look distinct on disk but came from one run. Imaging and analysis systems each keep their own metadata, so teams cannot trace which sample, run, or instrument produced a given file. Slide scanners and proteomics tools leave logs, swap files, and database backups behind. Training pipelines quietly consume petabytes of intermediate data, and a manual cleanup risks removing something regulated.

THE SOLUTION

Diskover indexes every file across your entire data estate into one catalog, then enriches it. Plugins parse the folder path, read directory lineage, and pull attributes from the storage system, and read provenance where a format carries it, such as a BAM alignment header. Researchers trace results to their source, policy clears ephemeral output from premium tiers, retention carries an audit trail, and curated datasets reach AI pipelines ready to use.

WHY IT MATTERS

Results trace back to the command and the reference that produced them, so an AI output or a published finding can be defended. Premium capacity comes back from ephemeral and duplicate files without anyone writing scripts. Retention and audit questions are answered from one current inventory. Curated imaging and sequencing datasets reach analysis and AI pipelines in days rather than weeks.

Why Diskover

ONE CATALOG FROM INSTRUMENT TO ARCHIVE

Index every file across HPC scratch, lab NAS, object, cloud, and archive into one searchable view, by fields such as project, pipeline, format, and age, so one search covers sequencing output, whole slide images, proteomics results, and archive, with data staying in place.

CONTEXT FROM THE ESTATE, NOT JUST THE FILE

Plugins add what a file name leaves out: tokens parsed from the folder path, lineage from the directory tree above it, classification tags, and attributes read from the storage system itself. Imaging, proteomics, and study records land in the catalog with the same richness as sequencing output.

SEQUENCING PROVENANCE FROM THE FILE ITSELF

The BAM plugin reads the alignment header, so fields such as the aligner, its version, its command line, and the annotation reference become searchable. It also stores a checksum of that command line, which is how files that look distinct on disk are identified as one run.

RETENTION YOU CAN EVIDENCE

Tag the files a study, a protocol, or a regulation requires you to keep, protect them from policy, and run removal with a preview and a full log of what went, when, and by whom, ready for an audit or an inspection.

EPHEMERAL OUTPUT CLEARED BY POLICY

Slide scanners, sequencers, and analysis tools leave logs, swap files, temporary output, and database backups behind. Diskover identifies them by metadata such as extension, path, and age, then policy clears or tiers them on schedule, so premium capacity comes back without scripts or a manual cleanup.

ASK IN PLAIN LANGUAGE

Ask the catalog, "Which whole slide images from closed studies are still sitting on premium storage?" and get the matching files back with a recommended action, such as tagging, tiering, removing, or packaging for a collaborator. AI-assisted, human-approved.

About Diskover

Diskover helps enterprises structure unstructured data across cloud, on-prem, and edge environments, making it easier to search, govern, optimize, and deliver file data for compliance, analytics, and AI-ready workflows.

Contact

diskoverdata.com
sales@diskoverdata.com
(800) 560-5853
San Jose, CA

Diskover

SOLUTION BRIEF

How it works

1 CONNECT
STORAGE OR IT TEAM

Point Diskover at HPC file systems, lab NAS, object, cloud, and archive. Indexing runs in place, so images, sequencing output, and analysis files stay where instruments and pipelines expect them.

Example: a sequencing share on HPC scratch is cataloged where it sits.

2 CATALOG
AUTOMATION

Every file becomes a searchable record with rich metadata, such as name, path, and owner, enriched with path tokens, directory lineage, classification tags, and storage attributes.

Example: a BAM file now shows its aligner, version, and reference.

3 CURATE
ANY AUTHORIZED TEAM

Tag by fields such as study, protocol, project, and status. Build saved searches for what teams ask weekly, such as ephemeral scanner output still on premium storage, or every file tied to a closed study.

Example: a scientist or lab manager saves "closed studies" as a search.

4 ACTIVATE
BY POLICY

Run tiering, removal, and packaging by policy with an audit trail, and deliver curated imaging, sequencing, and analysis metadata to analytics and AI pipelines through the API.

Result: one research institute avoided $250K in storage overspend while its AI pipelines kept running.

WORKS ACROSS Cloud • On-prem • Hybrid • NAS • SAN • Object • Tape • Archive • Data lakes • Lakehouses • AI/BI/ML

Every sample, slide, and run in one view.

Sequencing
BAM and SAM alignments, sorted copies, pipeline output
HPC

Imaging
.svs, .ndpi, and .mrxs slide images, scanner logs, swap files
NAS

Analysis and AI
proteomics output, checkpoints, abandoned training sets
HPC

Archive and audit
study records, protocols, legacy tape
TAPE

Diskover

ENRICHED METADATA CATALOG
Examples: project • study • lineage • pipeline • format • age | from files, paths, plugins, and tags
Every stage of the study and every team, from research to compliance, read from the same inventory.

Clear the ephemeral output
Found by fields such as extension, path, and age, tiered off premium storage.

Retention with evidence
Study and protocol records tagged, removed, and logged for the auditor.

Curate high-value datasets for AI
Clean imaging, sequencing, and analysis datasets reach pipelines in days.

Key outcomes

TRACEABLE RESULTS

Researchers trace any image, read, or result back to the run that produced it in one search, so review, reproduction, and publication questions stop pulling people off study work.

STORAGE THAT STAYS CLEAN

Scanner logs, swap files, duplicate reads, and abandoned training output move off premium tiers by policy, so capacity comes back on its own and the next storage purchase is smaller or later.

AUDIT AND AI READY

Retention and inspection reviews run from a single current inventory with the evidence attached, and curated imaging and sequencing datasets reach analysis and AI pipelines in days instead of weeks.

GET STARTED

Ready to see it on your estate?

Start with a 30-day proof of concept on your own data, and recover your investment in 3 to 6 months.

About Diskover

Diskover helps enterprises structure unstructured data across cloud, on-prem, and edge environments, making it easier to search, govern, optimize, and deliver file data for compliance, analytics, and AI-ready workflows.

Contact

diskoverdata.com
sales@diskoverdata.com
(800) 560-5853
San Jose, CA

Revised November 2026