Diskover

CUSTOMER SUCCESS STORY

Half a petabyte found in hours. $542K reclaimed in days.

How a cancer research center indexed 27 million pathology files across three sites and gave the capacity its pipelines had swallowed back to researchers.



$542K	0.5 PB	27M	4.7 PB
of storage reclaimed within days of the first index	of temporary, transient, and test data identified within hours	pathology files indexed across three sites	indexed in place, with no agents and no data movement

CUSTOMER PROFILE

A world-renowned cancer center on the US East Coast that combines a specialty hospital, a research institute with more than 100 laboratories, and a pathology department holding over a century of records.

Its pathology and research teams run imaging scanners, image management platforms, and machine learning pipelines that generate petabytes of data a year, spread across several file systems and storage clusters in three locations.

A small infrastructure team keeps that storage available for thousands of researchers, clinicians, and administrators.

At a glance

The center's pathology pipelines run on scanners and an image management platform. Each one produces valuable output and, on the side, large volumes of temporary, transient, and test files that the applications never clean up. With departments budgeting for multi-petabyte storage expansions every year, adding capacity had stopped being an answer.

Diskover indexed 27 million files across 4.7 PB and three sites in hours, with the data staying in place. Smart Searches and tags separated critical research and clinical data from pipeline leftovers by fields such as path, file type, age, and owner. Within hours the team had a list of more than half a petabyte of reclaimable data, and within days the data owners had reviewed and approved the cleanup and archive requests.

The reclaimed capacity went back to researchers and clinicians, planned storage expansions were deferred, and the center now has one place to see which pipelines fail to clean up after themselves, so the next half petabyte is caught before it lands.

Facts

INDUSTRY
Cancer research and clinical pathology.

DATA
Pathology imaging, scanner output, image management platform files, and over a century of clinical and research records.

STORAGE INDEXED
4.7 PB across multiple file systems and storage clusters, indexed in place.

USERS
Infrastructure and storage administrators running the platform, with data owners in pathology and research approving cleanup and archive requests.

LOCATIONS
Three primary sites supporting global cancer research and clinical care.

DEPLOYMENT STAGE
Production. Reclaimable data identified within hours of the first index, cleanup approved within days.

WHAT'S NEXT
Tagging policies across pathology workflows and capacity controls on machine learning training pipelines, so temporary data is flagged at each scan.

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

Diskover

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

CUSTOMER SUCCESS STORY

> *Pathology pipelines are built to produce results, not to clean up after themselves. Seeing this team find half a petabyte of leftovers within hours of their first index, and get it approved for cleanup within days, is exactly why we built Diskover around the metadata instead of the files.*
>
> **Paul Honrud, CPO**
> Diskover

CHALLENGES

PETABYTE-SCALE GROWTH

Departments were budgeting for multi-petabyte storage expansions every year, with no view of how much of that growth was research data and how much was pipeline leftovers.

TRANSIENT DATA CLOGGED TIERS

Pathology scanners and the image management platform generated temporary, transient, and test files at petabyte scale and left them on primary storage.

NO VIEW ACROSS THE ESTATE

Building a picture of storage and data across several file systems, clusters, and three sites was slow, manual, and error-prone, so decisions ran on estimates.

CAPACITY HIT RESEARCHERS

Storage over-utilization disrupted researchers and clinicians at the moments their work mattered most, and capacity requests landed on a small infrastructure team.

SOLUTIONS

RAPID INDEX, FULL ESTATE

Diskover indexed 27 million files across 4.7 PB and three sites in hours, from metadata only, with no agents on storage and no data movement.

OWNER APPROVAL, THEN CLEANUP

Reclaim candidates went to the data owners in pathology and research as review lists. Approved sets were archived or removed, with a record of who approved what and when.

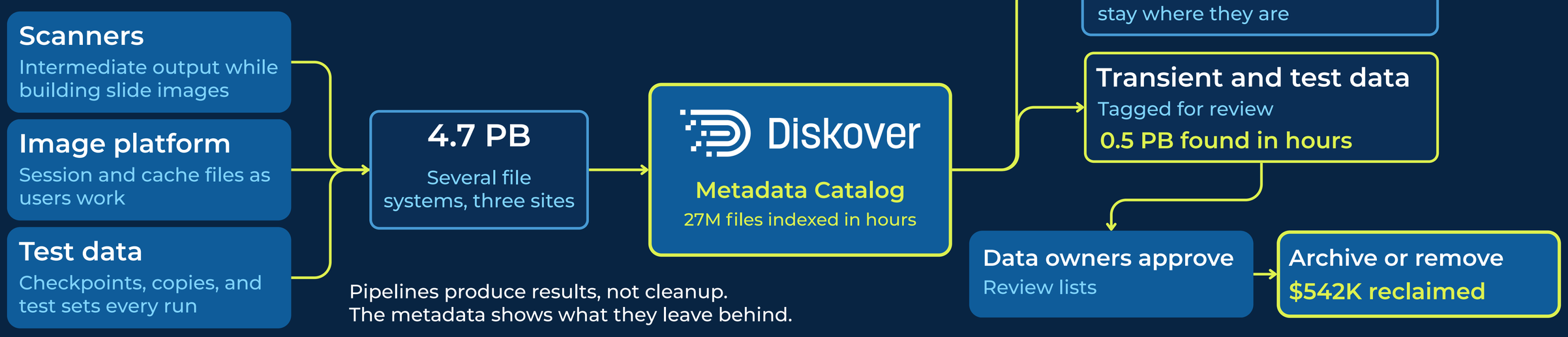
SMART SEARCHES AND TAGGING

Saved searches on fields such as path, file type, age, and owner separated research and clinical data from scanner output and platform temp files.

PIPELINE ACCOUNTABILITY

Reports by application and location show which pipelines fail to clean up, so lifecycle policies target the source of the growth instead of chasing the symptoms.

Where the half petabyte was hiding.



OUTCOMES

$542K RECLAIMED

More than half a petabyte of reclaimable data identified within hours and approved for cleanup within days, with no migration and no new hardware.

PETABYTE BUDGETS DEFERRED

Departments avoided budgeting for unnecessary multi-petabyte expansions because they could see what was actually consuming the estate.

HOURS BACK TO RESEARCH

Researchers and clinicians stopped chasing capacity and resumed their work, and the infrastructure team stopped fielding emergency storage requests.

ONE POINT OF CONTROL

A single integration point for lifecycle, retention, and governance policies replaced fragmented tools and simplified an unwieldy problem.

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.

Diskover

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