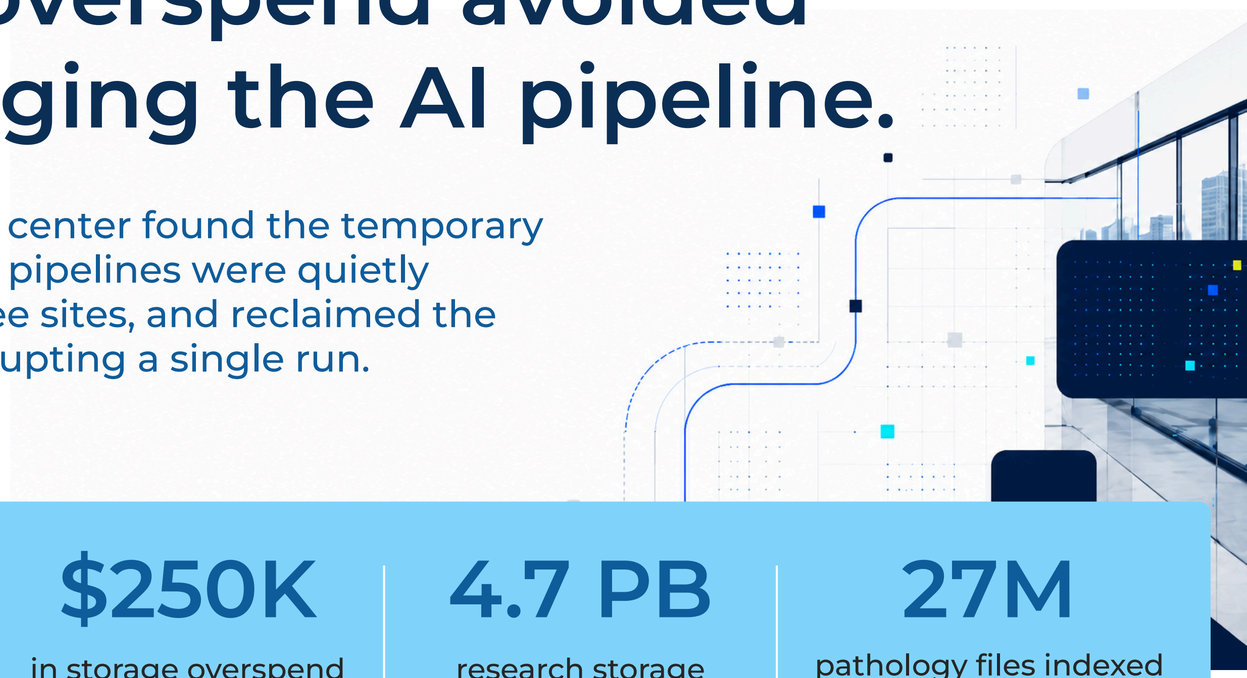
Diskover

CUSTOMER SUCCESS STORY

$250K in overspend avoided by unclogging the AI pipeline.

How a cancer research center found the temporary and training data its AI pipelines were quietly stacking up across three sites, and reclaimed the capacity without interrupting a single run.



500+ TB	$250K	4.7 PB	27M
of temporary and training data identified and cleared	in storage overspend avoided	research storage estate indexed in place	pathology files indexed into one catalog across three sites

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

Only 7 percent of respondents say their organization’s data is completely ready for AI, according to Cloudera and Harvard Business Review Analytic Services. This center faced the same challenge. Its bioinformatics and research IT teams ran data-intensive AI pipelines across a 4.7 PB research storage estate of 27 million files, spread over three geographic locations, and as those workloads scaled, storage costs climbed with them.

What the team could not see was why. There was no way to tell whether the pipelines were running efficiently or simply accumulating specialized data on the side. Diskover indexed the full estate into one searchable catalog of files, owners, and ages, then classified what was piling up and pinpointed the temporary, transient, and training data consuming capacity across all three sites.

The center cleared 0.5 PB of temporary, transient, and training data and avoided $250K in storage overspend, while keeping its AI pipelines running throughout. The same index also reclaimed $542K of the data estate within days. The team came away with something it did not have before: ongoing visibility into which pipelines produce results and which ones produce leftovers.

Facts

INDUSTRY
Cancer research and clinical pathology.

DATA
AI training data, temporary pipeline output, and specialized research datasets.

STORAGE INDEXED
4.7 PB and 27 million files across the center's research storage estate.

USERS
Bioinformatics, research IT, and lab operations teams.

LOCATIONS
Three geographic locations.

DEPLOYMENT STAGE
Production. Pipelines optimized and wasted capacity reclaimed.

WHAT’S NEXT
Curate AI-ready datasets and stream them into downstream research pipelines.

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

> *Every AI pipeline produces two things: the result somebody asked for, and the working files nobody planned for. At petabyte scale, that second pile quietly turns into your next storage purchase. Reading the metadata is how you tell the two apart before you buy more disk.*
>
> **Paul Honrud, CPO**
> Diskover

CHALLENGES

UNCONTROLLED STORAGE GROWTH
AI workloads generated data faster than the storage estate could absorb, driving costs up across three sites.

UNKNOWN DATA MIX
Specialized research data accumulated across the estate with little insight into what it was or whether it was still needed.

PIPELINE BLIND SPOTS
The team had no way to tell whether its AI data pipelines ran efficiently or simply piled up data on the side.

NO SINGLE OWNER
Pipelines crossed team boundaries, so no one group could say which data was still needed or authorize clearing it.

SOLUTIONS

INDEXED THE FULL ESTATE
Diskover crawled 4.7 PB and 27 million files across three locations into one searchable catalog of files, owners, and ages, with the data staying in place.

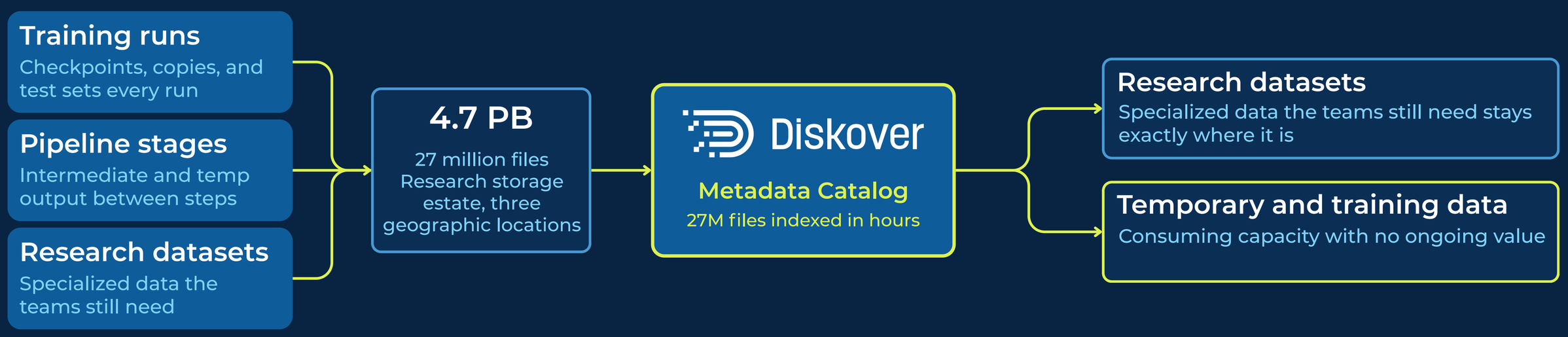
OPTIMIZED THE AI PIPELINES
Reports by application and location showed which pipelines left the most data behind, so the team could fix the biggest producers first.

PINPOINTED WASTED CAPACITY
Searches on fields such as path, age, and owner found temporary and training data unread since the run that created it.

CLASSIFIED THE DATA MIX
Tags on fields such as path, file type, and age separated critical research data from pipeline leftovers, so the team knew what to keep.

The pipeline produced results. It also produced half a petabyte.

AI pipelines are built to produce results, not to clean up after themselves.



0.5 PB cleared and $250K overspend avoided
via global visibility and data curation

Zero
interruption to a running pipeline

Ongoing
view of which pipelines leave data behind

OUTCOMES

$250K OVERSPEND AVOIDED
The center cleared 0.5 PB and avoided $250K in storage overspend, while keeping its AI pipelines running throughout.

CLEAR PIPELINE VISIBILITY
The team gained ongoing insight into pipeline efficiency and the data each run leaves behind.

HALF A PETABYTE RECLAIMED
Diskover identified and cleared over 500 TB of temporary and training data, freeing storage resources.

RESEARCH KEPT MOVING
Data management stayed efficient and uninterrupted throughout, supporting the center's research mission.

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