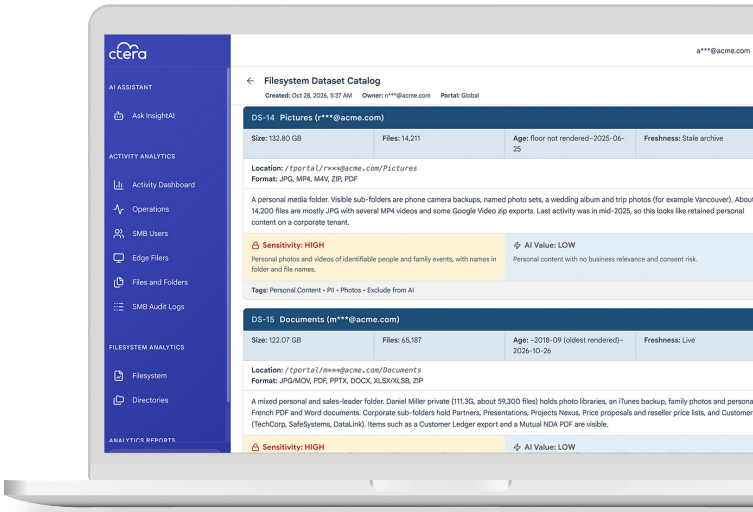
SOLUTION BRIEF

Unstructured Data Catalog

Know What Is in Your File Storage Before You Feed It to AI

Enterprises are investing in AI, but most cannot say what sits in their unstructured data estate: what exists, where it lives, what is worth using for AI initiatives, and which files contain risky sensitive data. **The CTERA Unstructured Data Catalog**, now available as part of CTERA InsightAI, maps your file storage in minutes. It shows which datasets carry value for AI, which carry risk, which are redundant, obsolete or trivial (ROT), and where to potentially start, without opening a single file.



Challenges with large unstructured data environments

- **No inventory.** Directories and files have built up over decades across file systems, often by people who have left the organization. Without insights from a dataset inventory, attempts to make data management decisions for AI, migration, archiving or compliance use cases are guesswork.
- **AI on unqualified data.** Pointing an LLM at a billion-file estate does not work. It cannot ingest at that scale, the cost of trying is prohibitive, and the answers that come back are inaccurate or outdated.
- **Risk in plain sight.** Regulated and personal content, often with the wrong permission set, sits in files or directories nobody has reviewed in years.
- **No ownership.** With the original owners potentially gone or moved on to new projects, nobody has standing to approve archiving or deletion. Your file storage only grows, and so do cost and exposure.

The CTERA Unstructured Data Catalog Solution maps your file storage in minutes. It shows which datasets have value for AI, which carry risk, and which are redundant, obsolete or trivial (ROT).

© 2026 CTERA Networks Ltd. All rights reserved.

What an Unstructured Data Catalog Should Deliver

Know what is in your file estate

CTERA InsightAI can provide a map of every major dataset across your file systems: what it is, where it lives, how big it is and how much looks redundant, obsolete or trivial. You get a data catalog report in minutes, not after a months-long survey.

Start AI on the right data, in the right order

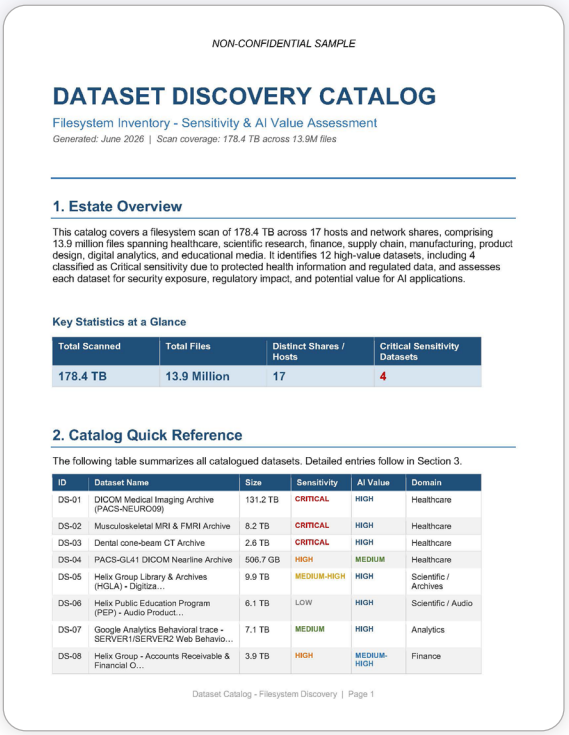
Know which datasets are worth feeding to AI, and which to leave out, so AI projects start on trustworthy data instead of paying to ingest everything and getting wrong answers back.

Find sensitive data before it causes a problem

See which datasets likely hold regulated or personal content, so security and privacy teams review those first, before an AI agent or an attacker reaches them, without opening and scanning every file.

Clean up stale data with evidence behind the decision

See which datasets have gone stale, with a recommended action for each, such as archive, reassign or dispose. Decision-makers get the evidence they need to approve cleanup, so storage cost and exposure stop growing by default.



NON-CONFIDENTIAL SAMPLE

DATASET DISCOVERY CATALOG

Filesystem Inventory - Sensitivity & AI Value Assessment

Generated: June 2026 | Scan coverage: 178.4 TB across 13.9M files

1. Estate Overview

This catalog covers a filesystem scan of 178.4 TB across 17 hosts and network shares, comprising 13.9 million files spanning healthcare, scientific research, finance, supply chain, manufacturing, product design, digital analytics, and educational media. It identifies 12 high-value datasets, including 4 classified as Critical sensitivity due to protected health information and regulated data, and assesses each dataset for security exposure, regulatory impact, and potential value for AI applications.

Key Statistics at a Glance

Total Scanned	Total Files	Distinct Shares / Hosts	Critical Sensitivity Datasets
178.4 TB	13.9 Million	17	4

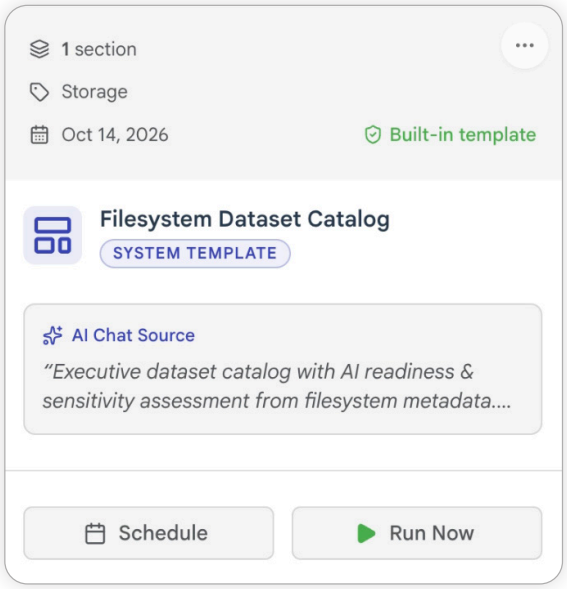
2. Catalog Quick Reference

The following table summarizes all catalogued datasets. Detailed entries follow in Section 3.

ID	Dataset Name	Size	Sensitivity	AI Value	Domain
DS-01	DICOM Medical Imaging Archive (PACS-NEURO09)	131.2 TB	CRITICAL	HIGH	Healthcare
DS-02	Musculoskeletal MRI & FMRI Archive	8.2 TB	CRITICAL	HIGH	Healthcare
DS-03	Dental cone-beam CT Archive	2.6 TB	CRITICAL	HIGH	Healthcare
DS-04	PACS-GL41 DICOM Nearline Archive	506.7 GB	HIGH	MEDIUM	Healthcare
DS-05	Helix Group Library & Archives (HGLA) - Digitiza…	9.9 TB	MEDIUM-HIGH	HIGH	Scientific / Archives
DS-06	Helix Public Education Program (PEP) - Audio Product…	6.1 TB	LOW	HIGH	Scientific / Audio
DS-07	Google Analytics Behavioral trace - SERVER1/SERVER2 Web Behavio…	7.1 TB	MEDIUM	HIGH	Analytics
DS-08	Helix Group - Accounts Receivable & Financial O…	3.9 TB	HIGH	MEDIUM-HIGH	Finance

Dataset Catalog - Filesystem Discovery | Page 1

The exported catalog summarizes the scanned estate and ranks each dataset by sensitivity and AI value.



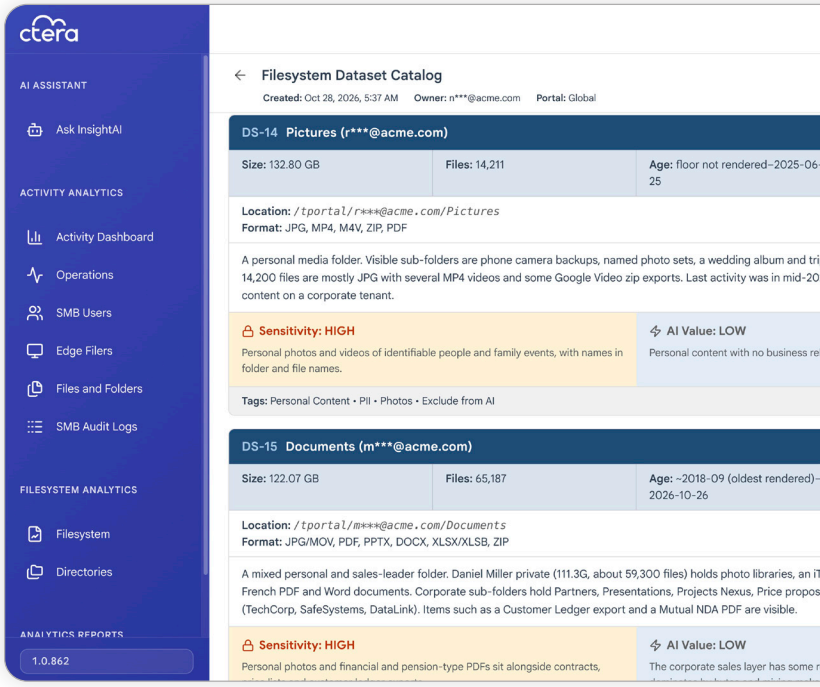
The Filesystem Dataset Catalog is a built-in template in CTERA InsightAI. Run it now or schedule it.

© 2026 CTERA Networks Ltd. All rights reserved.

How CTERA Delivers It

CTERA InsightAI builds the catalog from **metadata only**: file and folder names, folder structure, file types, sizes, dates and audit-log activity. It never opens a file. A token-efficient file system discovery algorithm (patent pending) gives CTERA InsightAI's LLM a compact representation of the estate, so the product can reason across billions of files without hitting the context-window limit.

- **Dataset grouping and description.**
  Files are grouped into datasets, and each is described in plain language with its location, size, file formats and freshness.
- **AI value and sensitivity ratings.**
  Each dataset is rated for its value to AI and its likely sensitivity, with a recommended order for ingestion.
- **ROT identification.**
  Redundant, obsolete and trivial data is flagged so it can be kept out of AI ingestion or considered for archiving.
- **Natural-language customization.**
  Start from the built-in report template, tailor it with prompts and ask follow-up questions such as "Show me high-value, low-sensitivity datasets ready for a RAG pilot."
- **Re-run at any time.**
  Because it runs on metadata, the catalog can be refreshed as the estate changes, including before data is migrated.



Each dataset entry shows size, file count, age, freshness, location and formats, a plain-language description, and sensitivity and AI value ratings.

Two ways to get started

Existing CTERA InsightAI customers

1. Open the Reports section in CTERA InsightAI and select the built-in Filesystem Dataset Catalog template.
2. Tailor it with natural-language prompts if needed, then run it. The cataloging pass completes in minutes.
3. Review the map of major datasets with their AI value and sensitivity ratings, ask follow-up questions and export the report.

Organizations not yet using CTERA InsightAI

1. Work with a CTERA Forward Deployed Engineer to run the CTERA Migrate Discovery tool on your current file shares, whether on CTERA or third-party storage.
2. CTERA engineers will privately analyze the output with CTERA InsightAI.
3. Review the catalog and recommendations then determine next steps for your data environment, IT initiative or AI project.

© 2026 CTERA Networks Ltd. All rights reserved.

From Estate-Wide Visibility to File-Level Detail

Metadata cannot show what is inside an individual file, so the catalog works at the dataset level, the granularity needed to set priorities. Once you know which datasets matter, CTERA Content Services, including CTERA Search and CTERA Classify, delivers full-text search, file-level classification and RAG on the selected subset to confirm what is in the files, and deliver accurate answers for your AI initiatives.

Beyond AI readiness. The same approach answers other questions about your estate. Describe your own rules, such as a compliance policy, and CTERA InsightAI reports on your file systems against them.

Sensitivity Tiers

CRITICAL	PHI / legally regulated medical data; unauthorised disclosure triggers mandatory breach notification
HIGH	Trade secrets, financial records, licensed IP, PII - significant legal and reputational risk
MEDIUM-HIGH	Scientific records with restricted subsets; mixed public/non-public content
MEDIUM	Pseudonymous behavioural data; GDPR/PEPA obligations likely; limited direct harm potential
LOW	Public domain or institutional content; no PII; standard copyright applies

AI Value Tiers

HIGH	Large-scale, labelled, domain-specific corpus with clear model training or fine-tuning application; scarce in open datasets
MEDIUM-HIGH	Valuable for specific applications; some curation or rights clearance required before use
MEDIUM	Useful for analytics and model evaluation; may lack scale or recency for frontier model training

AVAILABILITY

The Unstructured Data Catalog is now available in CTERA InsightAI, part of the CTERA Intelligent Data Platform. To see a catalog of your own file estate, contact your CTERA representative or email info@ctera.com.

REQUEST A DEMO TODAY

CONTACT US

CTERA.COM
info@ctera.com
USA: (917) 768-7193
Intl: +972-3-679-9000

ctera your files. your cloud.