APARAVI®

Case Study: How Visibility Rescued an AI Project Before It Began

[Challenge]

An enterprise software company scoped an AI training project expecting 1TB of data. But after deploying Aparavi Baseline to validate before ingestion, the truth emerged—they were sitting on 8TB+ of unstructured data. That included duplicate PDFs, obsolete documentation, sensitive HR files, and years of unused internal assets.

Without a clear view into what existed or what mattered, the project was poised for inflated cloud costs, hallucinated outputs, and extended delivery timelines.

[Breakthrough]

Baseline delivered instant visibility and deep insights across cloud and on-prem sources. With built-in OCR, classification, and contextual tagging, the team isolated what mattered—and discarded what didn't.

- 8TB scanned and indexed in under 24 hours
- 4TB flagged as ROT (redundant, obsolete, trivial)
- Sensitive and misaligned files tagged and isolated from training
- Authoritative content prioritized for ingestion
- Export-ready metadata and structured outputs generated

Baseline provided instant clarity, helping the team stay on budget by eliminating manual review, surfacing only relevant data, and filtering out sensitive or conflicting files—streamlining prep while strengthening governance

[Results at a Glance]

8TB of unstructured data indexed

50% reduction in storage and compute needs

70% less manual time spent on data prep

Model accuracy improved via curated inputs

Project timeline preserved from scope creep

[Benefits]

The Smarter Start to AI Projects

Aparavi Baseline helps organizations avoid costly missteps and scope creep by giving teams a clear, searchable understanding of what data they have before building any AI workflow.

Build Intelligence, Not Waste

Reveal hidden data volumes, remove the junk, and surface only what's relevant—so you can focus AI efforts where they'll make the most impact.

Reduce Cost & Scope Overrun

Cut unnecessary cloud spend and infrastructure requirements by identifying ROT data before it ever hits your AI pipeline.

Strengthen Data Trust

Classify sensitive files and surface compliance risks early to avoid exposing non-permissible data to LLMs or external APIs.