

Rubrik Annapurna

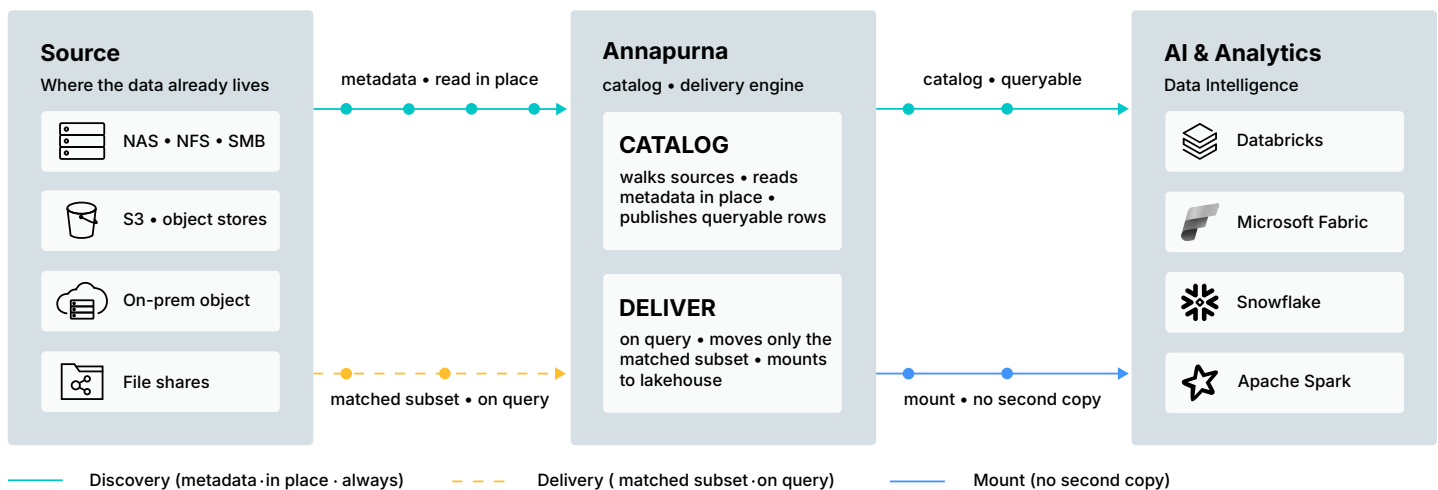
The AI-ready unstructured data layer for enterprise Data Intelligence platforms.

THE CHALLENGE

Enterprise AI runs on data. Roughly 90% of enterprise data is unstructured, and less than 10% of it ever reaches an AI pipeline.¹ The reason is plumbing. Before a data science team can train a model or ground a RAG application, raw files have to be copied out of NAS and object stores into a landing zone, access controls have to be rebuilt by hand, provenance is lost, and storage and ETL bills double.

Annapurna replaces that pattern. It scans unstructured data in place across NAS and object stores, publishes a queryable catalog of file metadata into a lakehouse, and stages only the subset a query actually pulls. Source permissions ride into the catalog. Lineage and versioning are recorded on every staged dataset. No bulk copy, no parallel storage estate, no rebuilt ACLs.

HOW IT WORKS



Crawl in place. Annapurna auto-discovers, scans, and indexes billions of files across NAS and object stores without copying source data in bulk.

Catalog for AI. Annapurna publishes a queryable catalog of unstructured file metadata natively into a lakehouse. Source access controls are captured and surfaced in the catalog.

Pull on demand. When a data team queries the catalog, Annapurna stages only the matched subset into a managed object store and mounts it to the lakehouse. Lineage and versioning are recorded on every staged dataset.

KEY BENEFITS

- Activate unstructured data for AI without standing up another ETL stack or duplicating petabytes of source storage.
- Cut time-to-models from weeks of pipeline work to a catalog query that delivers a governed dataset to the lakehouse.
- Preserve source permissions and produce auditor-defensible lineage from the source file to the AI-ready staged dataset.
- Align spend with consumption. Pay for what AI actually pulls, not for a parallel copy of the entire estate.
- Deploy alongside existing NAS, object storage, and applications.

KEY CAPABILITIES

- 1

Automated data discovery. Auto-discover, scan, and index unstructured files across NAS (Isilon, NetApp, NFS/SMB) and S3-compatible object stores.
- 2

Native lakehouse integration. Publish a queryable catalog of unstructured file metadata directly into a lakehouse. Data engineers query the index to pinpoint exact file targets, and Annapurna stages only that subset for downstream workflows.
- 3

Demand-driven pipeline economics. Stage only what a query needs into Annapurna’s managed object store, hosted in your S3 (or locally for smaller datasets). Mount that subset to the lakehouse for training, RAG, search, or BI.
- 4

Continuous governance. Capture source file system permissions, owners, and access controls at scan time and surface them in the catalog so downstream platforms enforce the right policy on the right user.
- 5

Verifiable chain of custody. Record lineage and versioning on every staged dataset, from the source file to the AI-ready dataset published to the catalog. Built for audit and reproducibility.
- 6

Built on Rubrik Security Cloud. Deploy as part of the same control plane as the rest of your Rubrik footprint, with the operational model and licensing your team already uses.

CAPABILITY TO OUTCOME

Capability	Operational outcome
In-place scan and index across NAS and S3	Inventory of every unstructured file across the estate without a bulk copy.
Queryable catalog of file metadata in a lakehouse	Data teams find and stage unstructured data from the same console they already use.
On-demand staging of matched subsets to a managed object store	AI consumes only the data it needs. No second storage estate to fund or secure.
Source ACLs captured and surfaced in the catalog	Existing permissions flow into governance decisions made downstream in the lakehouse.
Lineage and versioning on staged datasets	Auditable record of which file became which training input, and when.
Runs on Rubrik Security Cloud	One control plane, no new infrastructure.

SUPPORTED SOURCES AND INTEGRATIONS

Unstructured sources	NAS: Dell PowerScale (Isilon), NetApp ONTAP, NFS and SMB shares. Object: Amazon S3 and S3-compatible stores.
Lakehouse integration	Native lakehouse connector support in future releases.
Staging tier	Managed object store hosted in customer S3, or locally for smaller datasets. Mounted to the lakehouse on demand.
Control plane	Rubrik Security Cloud (RSC).
Compliance	Supports customer compliance programs across regulations.

Source
1 DalleMule, L. and Davenport, T. “What’s Your Data Strategy?” Harvard Business Review, May 2017.



Global HQ

3495 Deer Creek Road

Palo Alto, CA 94304

United States

1-844-4RUBRIK

inquiries@rubrik.com

www.rubrik.com

Rubrik (NYSE: RBRK), the Security and AI Operations Company, operates at the intersection of data protection, cyber resilience and enterprise AI acceleration. The Rubrik Security Cloud platform is designed to deliver robust cyber resilience and recovery including identity resilience to ensure continuous business operations, all on top of secure metadata and data lake. Rubrik’s offerings also include Predibase to help further secure and deploy GenAI while delivering exceptional accuracy and efficiency for agentic applications.

For more information please visit www.rubrik.com and follow [@rubrikinc](#) on X (formerly Twitter) and [Rubrik](#) on LinkedIn.

Rubrik is a registered trademark of Rubrik, Inc. All company names, product names, and other such names in this document are registered trademarks or trademarks of the relevant company.