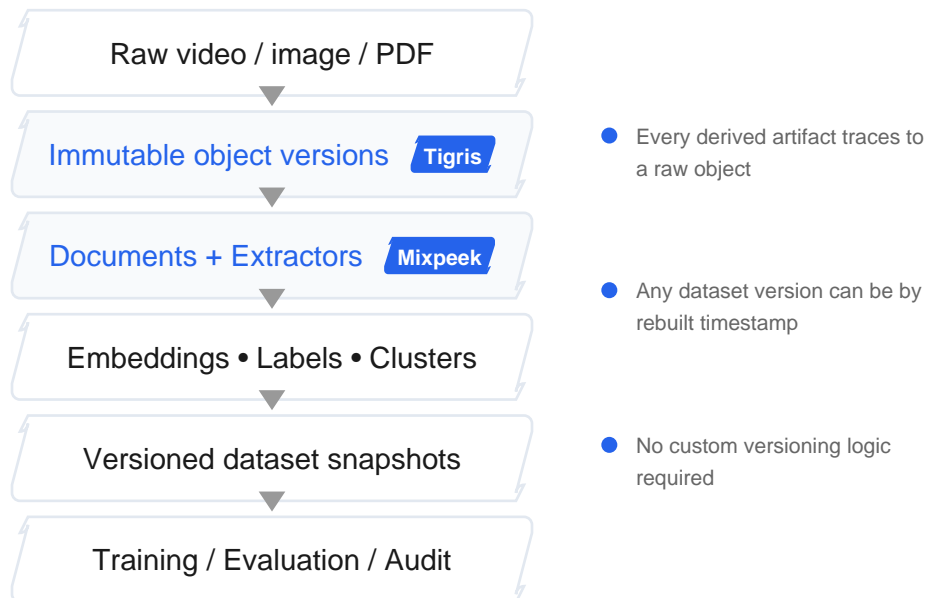# Reproducible Multimodal Datasets by Default

Versioned object storage + multimodal enrichment for ML teams retraining at scale

DATASET LINEAGE FLOW

Raw video / image / PDF

▼

Immutable object versions   **Tigris**

▼

Documents + Extractors   **Mixpeek**

▼

Embeddings • Labels • Clusters

▼

Versioned dataset snapshots

▼

Training / Evaluation / Audit

- Every derived artifact traces to a raw object
- Any dataset version can be by rebuilt timestamp
- No custom versioning logic required

WHY ENTERPRISES CARE

Reproducibility

Rebuild any training or evaluation dataset exactly as it existed.

Auditability

Trace predictions and models back to raw multimodal inputs.

Operational Simplicity

Object storage becomes the system of record—everything else is derived.

USE CASES

- Model retraining & regression analysis
- Regulated ML (finance, healthcare, safety)
- Long-lived content datasets (video, docs, learning content)

Works with existing S3-compatible pipelines. No migration required.   **Mixpeek + Tigris**