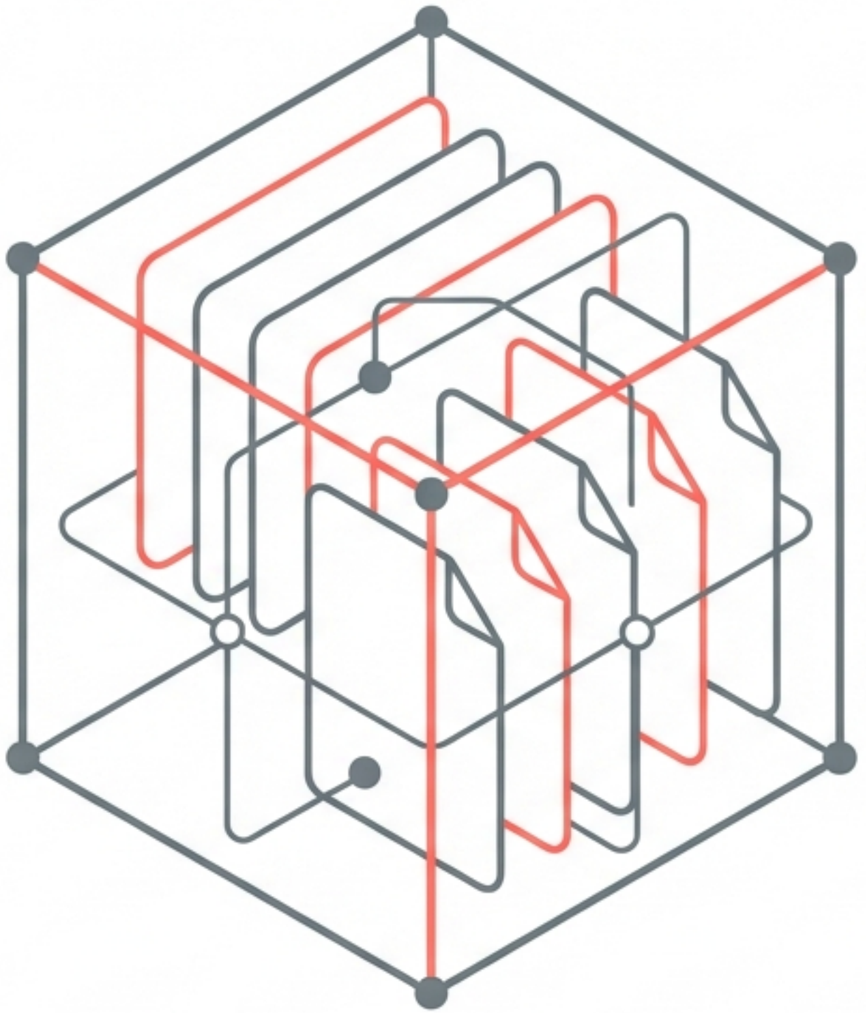
FileMind

An Intelligent Local File Management System



Project Review Zero
Department of Data Science and Business Systems

Kandi Karthik (RA2311027010119)
Akhandam Bhagavan Karthikeya (RA2311027010084)
Pritham Mukesh Krishna (RA2311027010084)

The Problem: Uncontrolled Digital Growth

The Chaos

File Redundancy

Digital collections grow without supervision. Identical copies accumulate with different filenames (e.g., 'report.pdf' vs 'report_final.pdf'), consuming massive amounts of disk space.

The Blind Spot

Inefficient Discovery

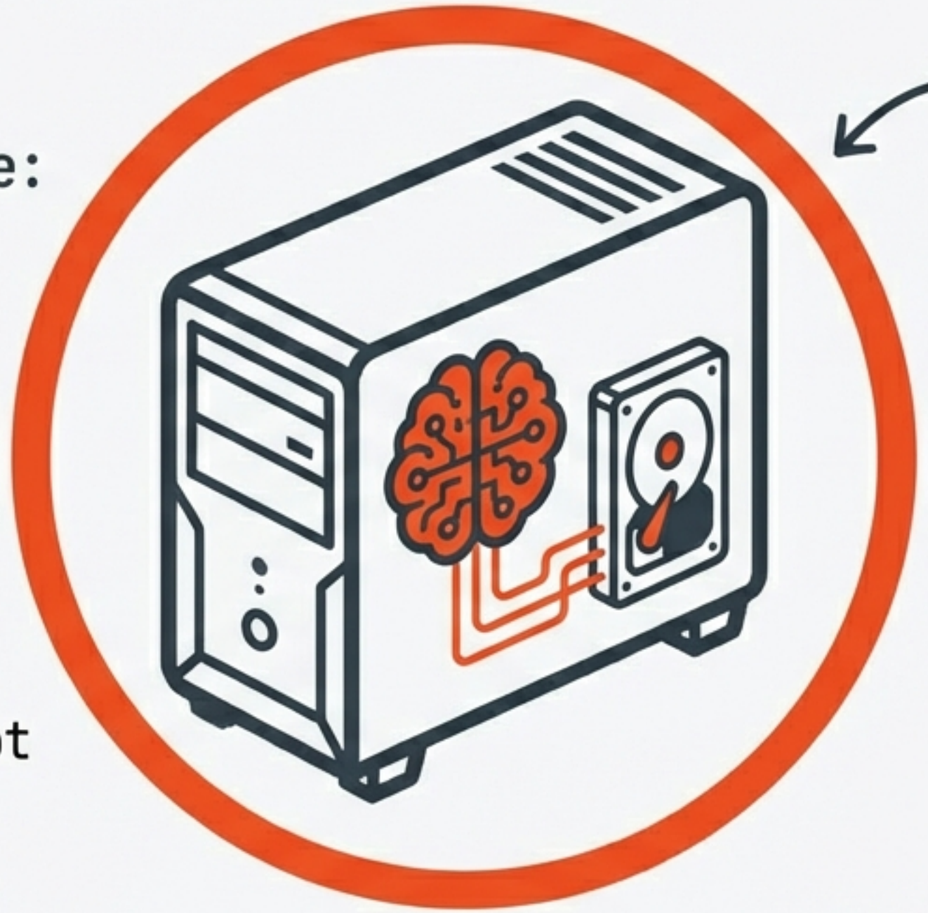
Standard OS tools rely on strict filename matching. Finding documents based on internal content or concepts is slow, ineffective, or impossible.

IMPACT: Wasted Storage & Lost Productivity.

Proposed Solution

The Local Intelligence Engine

Local-First Architecture:
Operates entirely on the user's machine.



PRIVACY BARRIER:
NO INTERNET CONNECTION

Deduplication:
Bit-for-bit exact copy detection.

Content-Aware:
Indexes meanings, not just filenames.

IMPACT: Reclaimed Storage & Enhanced Productivity.

Project Objectives

01	**Analyze Content:** Identify exact duplicates via hash analysis.
02	**Text Extraction:** Implement pipelines for PDF, DOCX, and TXT.
03	**Hybrid Search Engine:** Combine keyword matching with semantic vector analysis.
04	**User Interface:** Develop a robust Command-Line Interface (CLI).
05	**Privacy Assurance:** Zero-trust, fully offline operation.

Proposed Features & Workflow

filemind init

One-time environment setup. Downloads AI models and configures the local SQLite database.

filemind scan

Ingests a target directory. Extracts text, chunks content, and generates vector embeddings.

filemind search

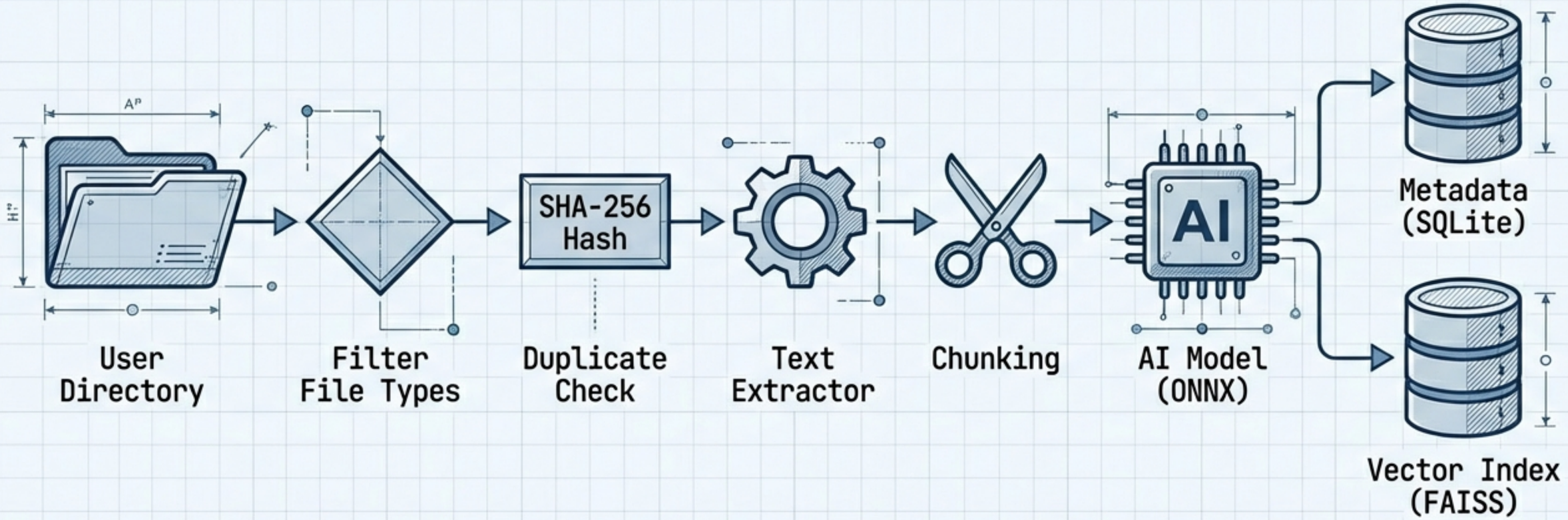
Hybrid retrieval. Matches natural language queries (concepts) and specific keywords.

filemind duplicates

Storage recovery. Identifies bit-for-bit identical files using SHA-256 hashing.

Proposed System Architecture

The Ingestion Pipeline



Technology Stack Overview

Language

Python 3.8+

CLI Framework

Typer

AI & Inference

BAAI/bge-small-en-v1.5

ONNX Runtime

Similarity Search

FAISS (IndexFlatIP)

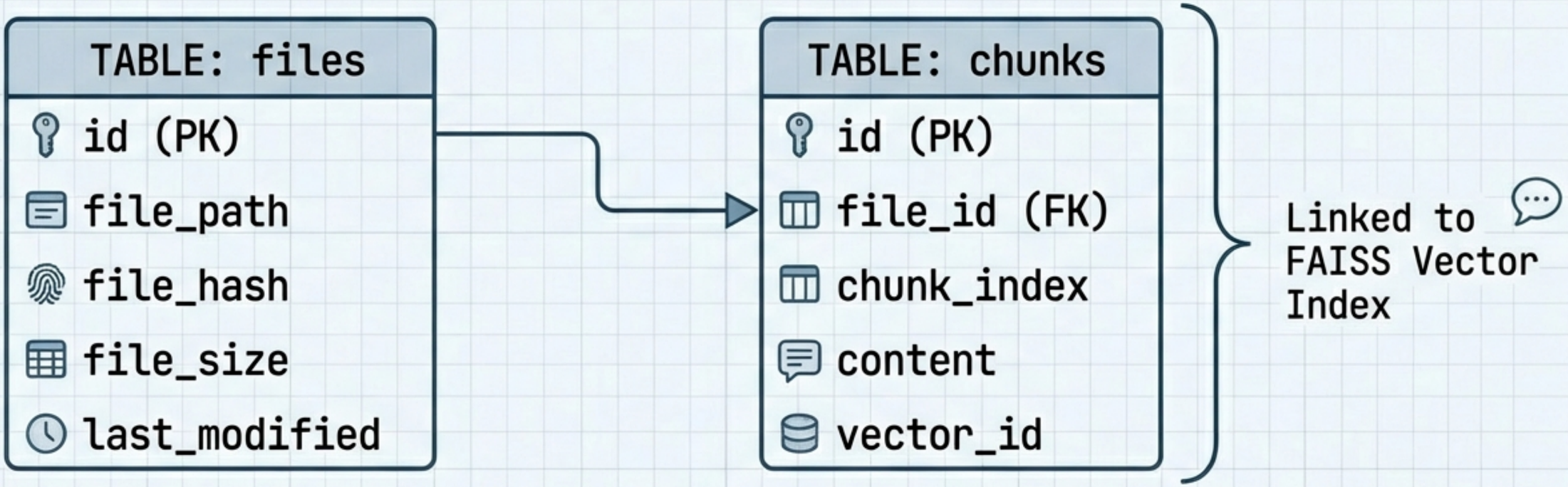
Metadata

SQLite + FTS5

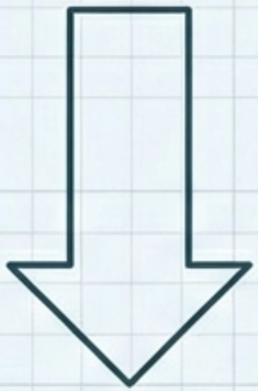
Dev Tools

Ruff (Linting) /
Black (Formatting)

Data Strategy & Schema



Expected Outcomes



Search Time

Drastic reduction in retrieval latency.

Disk Waste

Recovery of storage via deduplication.

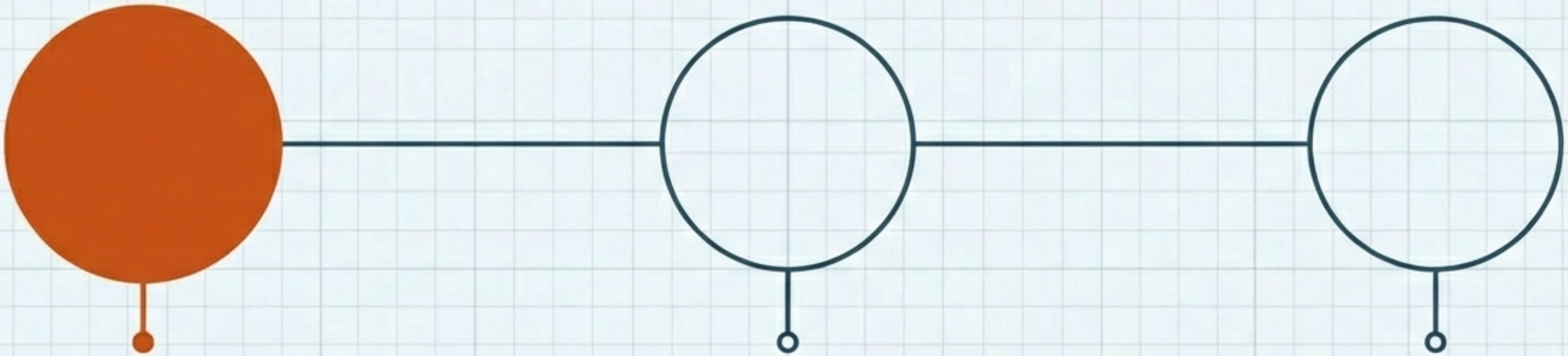
Discovery

Unlock files via conceptual matching.

Privacy

100% Data sovereignty.

Project Status & Roadmap



CURRENT STATE

- Idea & Planning Phase.
- Problem Definition.
- Architecture Design.

PHASE 1 (v0.0.1)

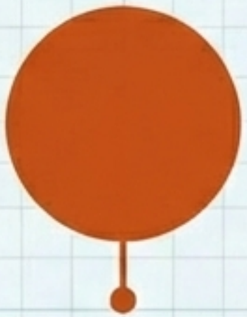
- Environment Setup (pyproject.toml). 'Init' Command Implementation.
- Model Integration.

PHASE 2

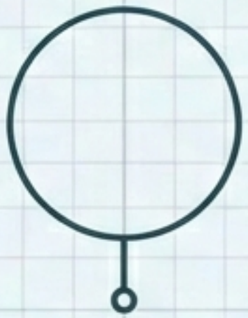
- Full Scanning Logic.
- Search Optimization.

CONCLUSION & FUTURE SCOPE

“FileMind: Your files, your data, intelligently managed.”



Graphical User Interface (GUI)



Real-time File Watching

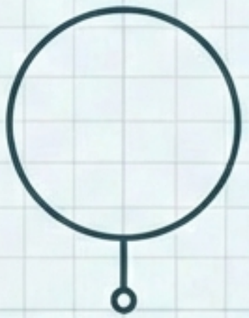


Image Similarity Search

Proposed by Team: Kandi Karthik, Akhandam Bhagavan Karthikeya, Pritham Mukesh Krishna.