VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

VIT-SanDisk Hackathon 2026 - Project Idea Submission

1. Project Overview

Track Name: Track 1: AI / ML

Theme Name: Intelligent Storage Orchestration & Predictive Maintenance

Project Name: GuardianDrive

Tagline: AI-Powered Predictive Health, Risk-Aware Tiering, and Multi-Cloud Orchestration

Live Project URL: https://divyanshupatel.com/GuardianDrive-sandisk/

2. Problem Statement

Modern storage environments in both consumer and enterprise sectors face four critical, interconnected challenges that result in data loss, inflated costs, and inefficiency.

A. Silent Drive Failures & Reactive Maintenance

- The Issue: Storage drives (HDDs/SSDs) often degrade silently. Traditional S.M.A.R.T. metrics are cryptic and ignored by 87% of users until catastrophic failure occurs.

- The Impact: Data loss is reactive; backups are often triggered *after* corruption has started or are not comprehensive enough for the specific data risks.

B. The Cloud Cost Spiral

- The Issue: Enterprises and power users waste 40-60% of cloud storage budgets on improperly tiered data.

  - Hot data (frequently accessed) is often trapped in slow, cold archives.

  - Cold data (rarely accessed logs/backups) sits in expensive premium SSD tiers.

- The Impact: Massive, unnecessary monthly bills due to a lack of granular, file-level visibility.

C. Fragmented Storage Landscape

- The Issue: Infrastructure is siloed between Local NVMe/HDDs, External Arrays, and multiple Cloud Providers (AWS, Azure, GCP).

- The Impact: There is no "single pane of glass" to orchestrate data movement based on health, cost, and performance across these disparate layers.

D. "Data Gravity" & Compression Guesswork

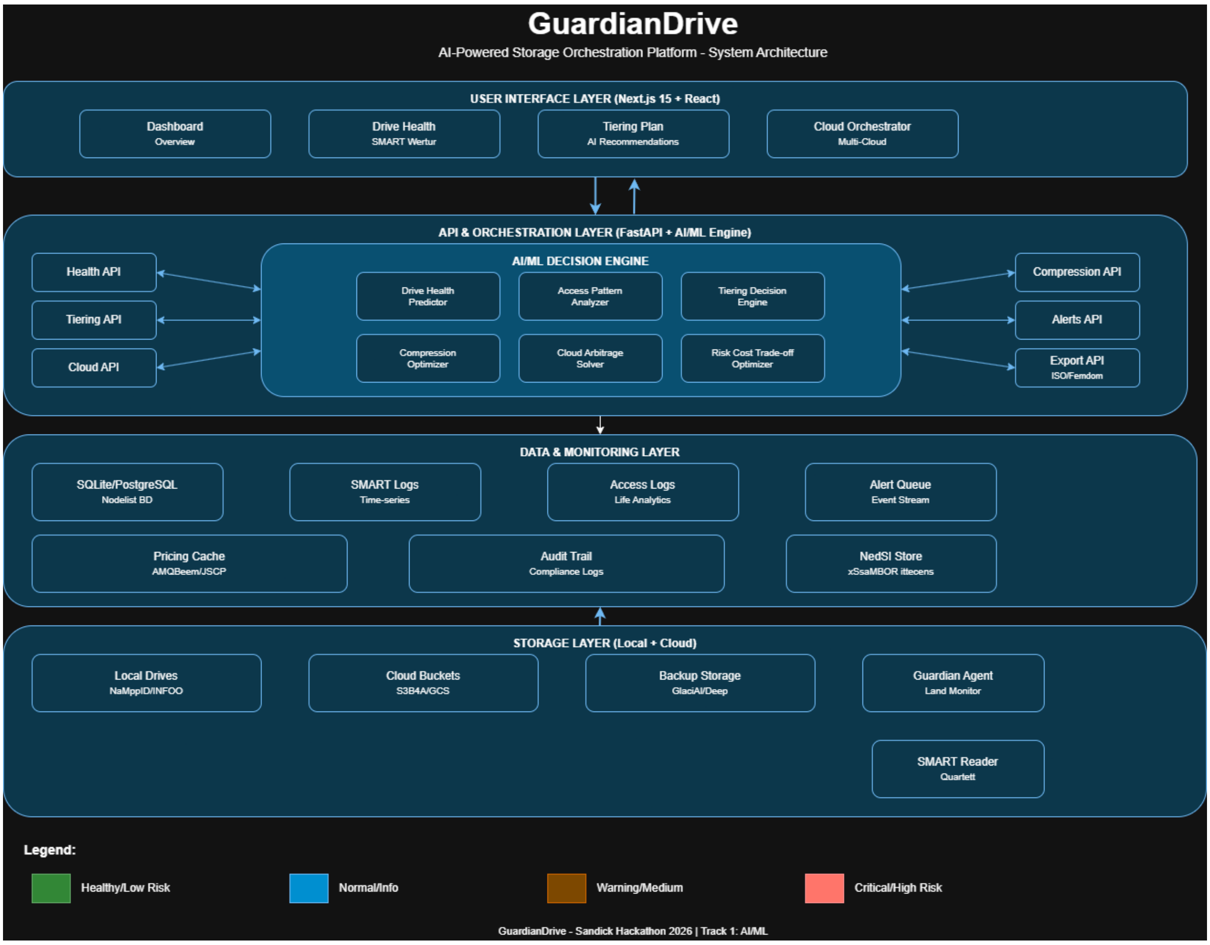
- The Issue: Moving large datasets is expensive (egress fees) and slow. Users blindly compress files without knowing if it will yield significant space savings.

- The Impact: Wasted CPU cycles compressing incompressible media (JPEGs, MP4s) and missed opportunities to compress text-heavy logs/databases.

GuardianDrive Solution:

An AI-native orchestration platform that unifies predictive maintenance, intelligent tiering, and multi-cloud arbitrage into a single, automated workflow.

3. Detailed Architecture / Block Diagram



System Architecture Overview:

Layer 1: USER INTERFACE LAYER

- React 19 + Vite + Tailwind Obsidian Dashboard

- Components: Overview, Health Monitor, Smart Tiering, Compression, Cloud Ops

Layer 2: API & ORCHESTRATION LAYER

- FastAPI + Python 3.11 Backend

- Modules: Health API, Tiering API, Cloud API

- AI/ML Decision Engine (The "Brain"):

  1. Drive Health Model (XGBoost Classifier)

  2. Access Pattern Model (K-Means Clustering)

  3. Cost Solver (Weighted Scalarization)

Layer 3: DATA LAYER

- SQLite/PostgreSQL Metadata

- SMART Telemetry DB

- File Access Logs

- Cloud Pricing Cache

Layer 4: STORAGE LAYER

- Local NVMe/HDD

- External Arrays

- NAS / DAS

Layer 5: EXTERNAL SERVICES

- AWS S3 API

- Azure Blob API

- GCP Storage API

4. Detailed Methodology

Our approach combines supervised learning for health prediction, unsupervised learning for data classification, and mathematical optimization for cost placement.

Phase 1: Predictive Drive Health (The "Guardian")

- Data Source: Trained on the Backblaze Hard Drive Dataset (100k+ drives, 5 years of history).

- Feature Engineering: Uses raw SMART attributes (Reallocated Sectors, Seek Error Rate, Power-On Hours, Temperature) and time-series features (Rolling averages) to detect degradation trends.

- Model: XGBoost Classifier.

- Objective: Predict failure within a 14-day horizon.

- Output: A granular Health Score (0-100) and a Risk Class (Low/Medium/Critical).

- Innovation: Unlike simple "Pass/Fail" SMART flags, our model provides a probabilistic "days to failure" countdown.

Phase 2: Intelligent Data Classification (The "Auditor")

- Problem: Files are not equal. A 10GB log file is different from a 10GB 4K video.

- Method: K-Means Clustering.

- Features: Recency (days since last access), Frequency (access count), and Size.

- Clusters: The model automatically segments files into 4 tiers:

1. HOT: Frequently accessed, recent. (Keep on NVMe)

2. WARM: Occasional access. (Move to HDD/SATA)

3. COLD: Rare access. (Move to Cloud Cool tier)

4. ARCHIVE: Never accessed, compliance data. (Move to Glacier/Deep Archive)

Phase 3: Risk-Cost Optimizer (The "Broker")

- Problem: Minimizing cost often increases risk. We need a balance.

- Method: Weighted Scalarization Optimization.

- Formula: Score = w1 * Norm(Cost) + w2 * Norm(Risk) + w3 * Norm(Latency)

- Execution: The system generates 3 distinct strategies for the user:

1. Conservative: Max redundancy, High Performance (High Cost)

2. Balanced: (Recommended) Optimal trade-off (Medium Cost)

3. Aggressive: Lowest possible cost, Cloud Archive focus (Low Cost)

Phase 4: Compression ROI Analysis

- Logic: Before compressing, we calculate the Return on Investment (ROI).

- Calculation: If (Predicted Savings > Compute Cost), compression is recommended.

- Tech: Uses Zstandard (zstd) for variable compression levels based on the file type.

Phase 5: Multi-Cloud Arbitrage

- Method: Real-time query of AWS/Azure/GCP pricing APIs.

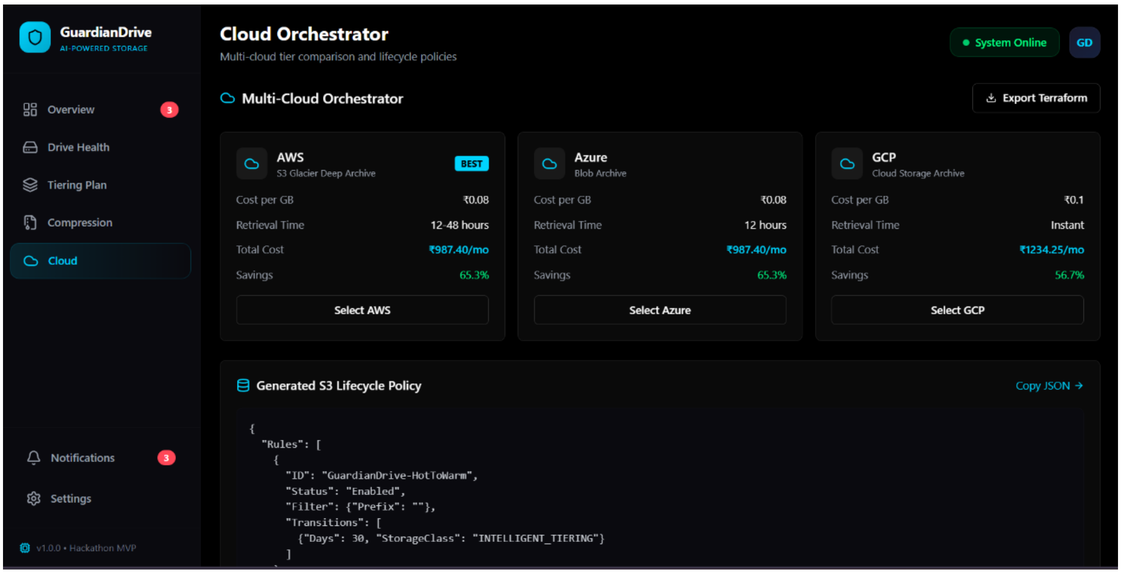
- Action: Compares current storage costs vs. competitors. Auto-generates Terraform scripts to migrate buckets if a cheaper provider is found. Generates S3 Lifecycle Policies to automate transitions.

5. Deliverables Expected & Achieved

A. Functional Web Dashboard (MVP)

- A fully responsive, dark-mode dashboard built with React & Vite.

- Includes real-time storage health monitoring and visual drive life expectancy gauges.



Cloud Orchestrator

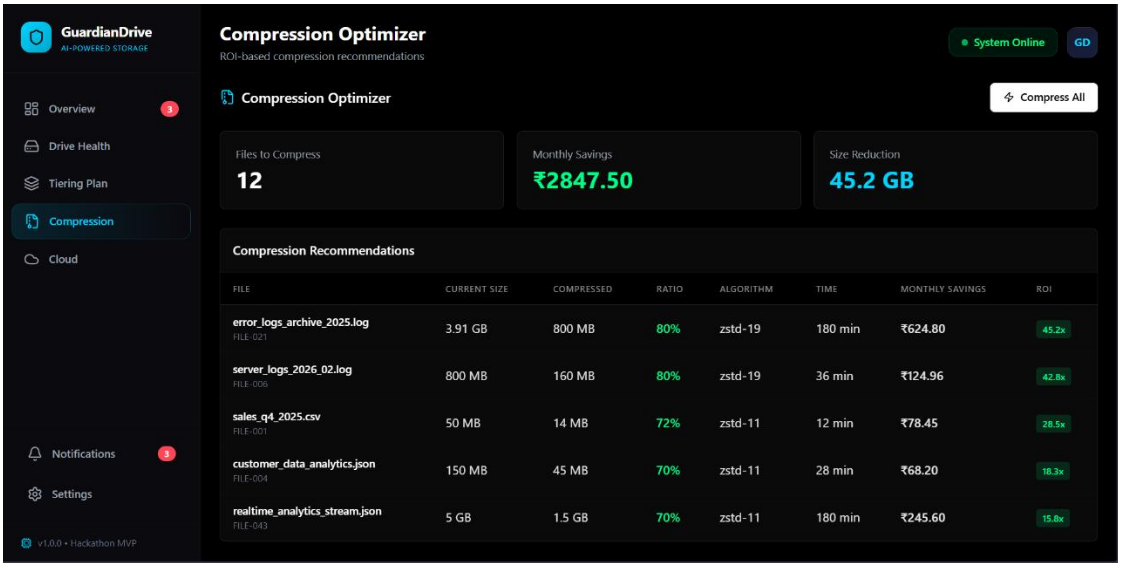
Multi-cloud tier comparison and lifecycle policies

Multi-Cloud Orchestrator

	AWS S3 Glacier Deep Archive	Azure Blob Archive	GCP Cloud Storage Archive
Cost per GB	₹0.08	₹0.08	₹0.1
Retrieval Time	12-48 hours	12 hours	Instant
Total Cost	₹987.40/mo	₹987.40/mo	₹1234.25/mo
Savings	65.3%	65.3%	56.7%

Generated S3 Lifecycle Policy

```
{
  "Rules": [
    {
      "ID": "GuardianDrive-HotToWarm",
      "Status": "Enabled",
      "Filter": {"Prefix": ""},
      "Transitions": [
        {"Days": 30, "StorageClass": "INTELLIGENT_TIERING"}
      ]
```



Compression Optimizer

ROI-based compression recommendations

Compression Optimizer

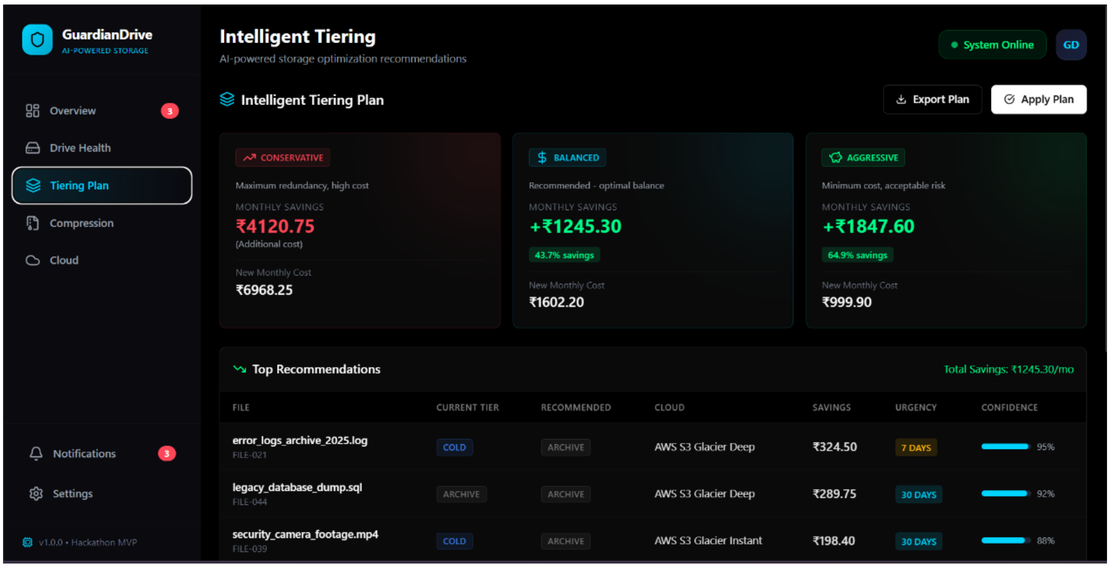
Files to Compress
12

Monthly Savings
₹2847.50

Size Reduction
45.2 GB

Compression Recommendations

FILE	CURRENT SIZE	COMPRESSED	RATIO	ALGORITHM	TIME	MONTHLY SAVINGS	ROI
error_logs_archive_2025.log FILE-021	3.91 GB	800 MB	80%	zstd-19	180 min	₹624.80	45.2x
server_logs_2026_02.log FILE-006	800 MB	160 MB	80%	zstd-19	36 min	₹124.96	42.8x
sales_q4_2025.csv FILE-001	50 MB	14 MB	72%	zstd-11	12 min	₹78.45	28.5x
customer_data_analytics.json FILE-004	150 MB	45 MB	70%	zstd-11	28 min	₹68.20	18.3x
realtime_analytics_stream.json FILE-043	5 GB	1.5 GB	70%	zstd-11	180 min	₹245.60	15.8x



Intelligent Tiering

AI-powered storage optimization recommendations

Intelligent Tiering Plan

CONSERVATIVE

Maximum redundancy, high cost

MONTHLY SAVINGS

₹4120.75

(Additional cost)

New Monthly Cost

₹6968.25

BALANCED

Recommended - optimal balance

MONTHLY SAVINGS

+₹1245.30

43.7% savings

New Monthly Cost

₹1602.20

AGGRESSIVE

Minimum cost, acceptable risk

MONTHLY SAVINGS

+₹1847.60

64.9% savings

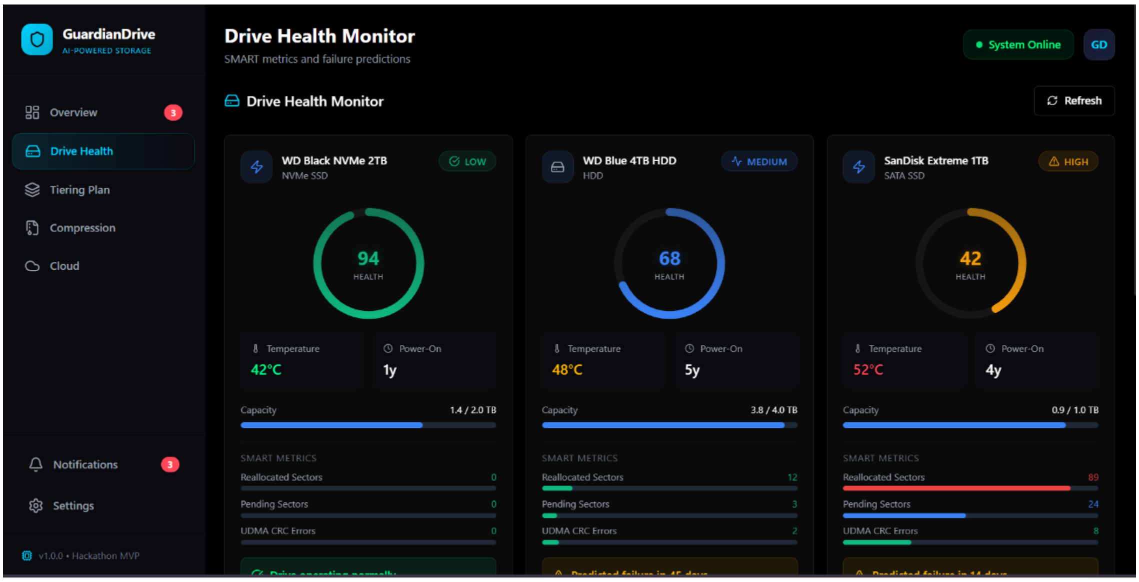
New Monthly Cost

₹999.90

Top Recommendations

Total Savings: ₹1245.30/mo

FILE	CURRENT TIER	RECOMMENDED	CLOUD	SAVINGS	URGENCY	CONFIDENCE
error_logs_archive_2025.log FILE-021	COLD	ARCHIVE	AWS S3 Glacier Deep	₹324.50	7 DAYS	95%
legacy_database_dump.sql FILE-044	ARCHIVE	ARCHIVE	AWS S3 Glacier Deep	₹289.75	30 DAYS	92%
security_camera_footage.mp4 FILE-039	COLD	ARCHIVE	AWS S3 Glacier Instant	₹198.40	30 DAYS	88%



Drive Health Monitor

SMART metrics and failure predictions

Drive Health Monitor

WD Black NVMe 2TB
NVMe SSD
LOW

94
HEALTH

Temperature 42°C
Power-On 1y

Capacity 1.4 / 2.0 TB

SMART METRICS

Reallocated Sectors 0
Pending Sectors 0
UDMA CRC Errors 0

Drive operating normally

WD Blue 4TB HDD
HDD
MEDIUM

68
HEALTH

Temperature 48°C
Power-On 5y

Capacity 3.8 / 4.0 TB

SMART METRICS

Reallocated Sectors 12
Pending Sectors 3
UDMA CRC Errors 2

Predicted failure in 45 days

SanDisk Extreme 1TB
SATA SSD
HIGH

42
HEALTH

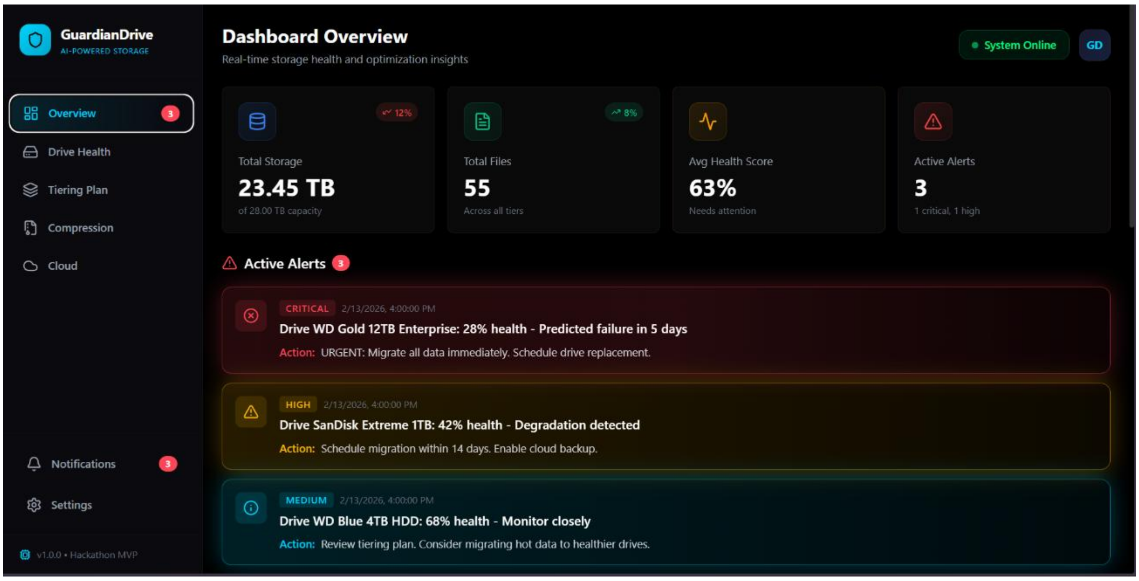
Temperature 52°C
Power-On 4y

Capacity 0.9 / 1.0 TB

SMART METRICS

Reallocated Sectors 89
Pending Sectors 24
UDMA CRC Errors 8

Predicted failure in 14 days



Dashboard Overview

Real-time storage health and optimization insights

12%
Total Storage
23.45 TB
of 28.00 TB capacity

8%
Total Files
55
Across all tiers

Avg Health Score
63%
Needs attention

Active Alerts
3
1 critical, 1 high

Active Alerts 3

CRITICAL 2/13/2026, 4:00:00 PM
Drive WD Gold 12TB Enterprise: 28% health - Predicted failure in 5 days
Action: URGENT: Migrate all data immediately. Schedule drive replacement.

HIGH 2/13/2026, 4:00:00 PM
Drive SanDisk Extreme 1TB: 42% health - Degradation detected
Action: Schedule migration within 14 days. Enable cloud backup.

MEDIUM 2/13/2026, 4:00:00 PM
Drive WD Blue 4TB HDD: 68% health - Monitor closely
Action: Review tiering plan. Consider migrating hot data to healthier drives.

B. AI/ML Engine

- A Python-based backend service (FastAPI) that runs the XGBoost inference for drive health and K-means clustering on file metadata.

C. Optimization Logic

- Automated logic to classify Hot/Cold data.

- Cost-benefit calculator for compression and cloud migration.

D. Documentation

- Architecture Diagram: System design and data flow.

- API Documentation: Interactive Swagger/OpenAPI docs.

- Setup Guide: Comprehensive README.md for local deployment.

6. Technology Stack

- Frontend: React 19, TypeScript, Tailwind CSS, Recharts, Lucide Icons.

- Backend: Python 3.11, FastAPI, Uvicorn, Pydantic.

- ML Libraries: XGBoost, Scikit-learn, Pandas, NumPy.

- Data: SQLite (Metadata), Backblaze Metrics (Training Data), Cloud Pricing JSONs.

- DevOps: GitHub Actions (CI/CD), GitHub Pages (Hosting).

7. Future Roadmap

1. Real-Time Hardware Integration: direct interfacing with smartctl / OS kernel for live telemetry.

2. Federated Learning: Allowing the model to learn from user-specific drive degradation patterns without uploading sensitive data.

3. Ransomware Detection: Analyzing "write velocity" anomalies to detect encryption attacks in real-time.

4. Desktop Agent: An Electron-based background service for continuous local monitoring.

8. Team Details

- Team Lead: Divyanshu Patel (divyanshu.patel2023@vitstudent.ac.in)
- Member 2: Varshith Pilli
- Member 3: Ashutosh Gunjal
- Member 4: Waqar Azim
- Member 5: Soumil Gandhi