DataGenie

FROM DETECTION TO ACTION:
DATA QUALITY REMEDIATION ASSISTANT

Detect issues. Get AI powered fixes. Improve Data Quality at scale.

THE PROBLEM

- **Extremely difficult to detect early**
- **Damage happens before anyone notices**
- **Breaks ML pipelines and models**
- **Disrupts forecasts and operations**
- **Costs both time and money**

Poor data quality costs US organizations an average of

$3.1 Trillion
per year
-IBM

BUSINESS IMPACT

TECH/GAMING

ML PIPELINE CORRUPTED BY STATISTICAL OUTLIERS

UNITY SOFTWARE

$110M in losses

AVIATION

NULL ANOMALIES CRASHED CREW SCHEDULING SYSTEM

SOUTHWEST AIRLINES

$800M+ in losses

FINANCE

DATA QUALITY ERROR IN +11K REGULATORY REPORTS

CITIGROUP

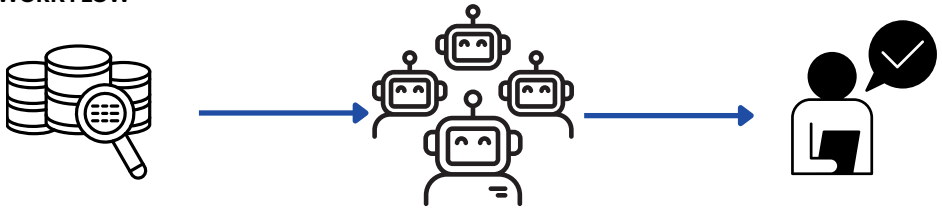
$536M+ in fines

Sources: Unity Q2 2022 Earnings Call, Southwest Airlines Q4 2022 Earnings Call, OCC Consent Order AA-EC-2020-64 (Oct 7, 2020)

OUR SOLUTION

A **MULTI-AGENT AI SYSTEM** THAT GOES FROM BAD DATA DETECTED TO FIX APPROVED IN A FEW MINUTES.
THEY TELL YOU THE PIPE IS LEAKING. WE HAND YOU THE WRENCH.

WORK FLOW



SPARK DETECTS → **MULTI-AGENT GENERATES FIX** → **HUMAN IN THE LOOP**

Pipeline

1 DETECT
Apache Spark scans millions of rows for null spikes, statistical outliers, and format inconsistencies

2 ROUTE
Router classifies each issue and assigns the right remediation strategy

3 DIAGNOSE
Agent 1 identifies root cause and business impact, not just flags the problem

4 REMEDIATE
Agent 2 generates runnable PySpark fix code, ready to paste into your pipeline

5 VALIDATE
Critic Agent verifies every output and retries if invalid

6 APPROVE
Human-in-the-loop by design, no fix is applied without analyst approval

7 LEARN
Every human decision is logged and fed back to improve future recommendations

Why DataGenie?

	DATA GENIE	MONTE CARLO/ANOMALO
ANOMALY DETECTION	Statistical & Heuristic	Automated Observability
ROOT CAUSE ANALYSIS	LLM-Driven: **why + so what**	Lineage Based: **where**
RUNNABLE FIX CODE	Auto-Generated PySpark	SQL only
CODE VALIDATION	Critic Auto-retry up to 2x if invalid	Manual Validation
FEEDBACK LOOP	Audit Trail & Feedback loop	No learning loop

Use Cases

Tech / AI
Training data outlier detection

Healthcare
Patient record completeness checks

Aviation / Ops
Real-time null & format anomaly

Retail / E-commerce
Inventory & transaction anomaly alerts

Finance
Schema drift & compliance validation

Manufacturing
Sensor data drift & predictive maintenance

Apache Spark · databricks · GPT-4o mini · python · Streamlit · plotly

Team Six


Sean Cabaniss
caban025@umn.edu


Chingfen Hung
hung0133@umn.edu


Omkar Thombare
thomb017@umn.edu


Yung Hsuan Hsieh
hsieh203@umn.edu


Yonghui Kim
kim02452@umn.edu