Data Modernization

Global Education Provider

How We Unified Millions of Student Data Points in 9 Weeks to Predict Learner Risk in Real Time

Our client is a global education provider with tens of thousands of employees and a presence across multiple continents. They support millions of learners preparing for high-stakes professional and academic exams.

Their business model is built on student success. Many of their courses include score guarantees. If students do not achieve the required results, refunds may apply. This makes early visibility into student performance critical not just for learning outcomes but for protecting revenue.

However, their data told an incomplete story. Student information was spread across multiple systems. Each platform tracked different parts of the learner journey. There was no single view of performance, engagement, or risk.

The organization needed a way to connect everything. They needed to spot struggling students early. And they needed to act before it was too late.

Utilized Technology Stack:

Cloud & Database: MongoDB Atlas (v7.0/8.0), PostgreSQL 16, MySQL 8.

Message Broker: Apache Kafka (Confluent Platform 7.5)

Change Data Capture (CDC): Debezium (PostgreSQL & MySQL connectors), Kafka Connect, MongoDB Kafka Connector

Stream Processing: Apache Flink 1.18 (PyFlink — Python API)

API Layer: Python FastAPI

Data Modeling: YAML-driven mapping engine, Pydantic data models

Observability: structlog (structured logging), Kafka UI

Infrastructure: Docker, Docker Compose

Testing: pytest (124 automated tests: 99 mapping + 25 performance)

Review of the Challenges

The client did not lack data; they lacked visibility which caused a rising risk to them.

Information was fragmented across multiple platforms. Each system used different formats. Learners had different identifiers depending on where they appeared. Even basic questions were difficult to answer.

This created a series of critical challenges:

- No unified view of each learner
- No consistent way to track engagement or progress
- No early warning system for disengagement
- No ability to intervene before failure or refund requests

At the same time, the business had a clear target: keep refund rates below 2.5%. Without better insight, this goal was increasingly difficult to maintain.

Our Solution

We built a unified platform that brings all learner data together and makes it usable in real time.

Instead of relying on delayed reports, the system now reacts instantly as new data arrives.

It connects multiple data sources, aligns learner identities, and builds a complete picture of each individual. From this, it calculates a live risk score for every learner.

This score is based on four simple but powerful signals:

- Progress against study plans
- Assessment performance
- Recent activity levels
- Areas of topic difficulty

Each learner is then classified as low, medium, or high risk. The system also suggests clear next steps, such as targeted support, outreach, or encouragement.

This turns complex data into a clear and practical action.

Our Approach: Structured, Collaborative, and Delivered at Pace

We worked at speed but with structure and collaboratively to ensure clarity was not lost.

In the first two weeks, we ran a series of deep discovery sessions with stakeholders across engineering, product, and leadership. This ensured alignment from day one.

We then designed a single, unified data model that could work across all systems. This avoided the risk of creating separate solutions for each platform.

From there, we built the solution in stages:

- Absorbing data from multiple systems
- Processing it in real time
- Structuring it into a unified format
- Making it accessible through simple APIs

Testing was built in from the start. Over 120 automated tests ensured the platform was accurate, reliable, and ready to scale.

All of this was delivered within a 9-week engagement.

The Outcome:

The results were immediate; they had real-time insights, faster intervention, and measurable impact.

For the first time, the organization has a single, unified view of each learner. Data that was once fragmented is now connected and accessible.

The platform now:

- Processes over 1,000 risk calculations per second
- Delivers end-to-end data updates in under 60 seconds
- Provides API responses in under 5 seconds
- Automatically adjusts as new data arrives, with no manual intervention

Most importantly, teams can now act early. They can identify at risk learners as soon as patterns emerge and provide support before outcomes decline. This shift from reactive to proactive support helps improve learner success and reduces the likelihood of refunds. The platform also creates a solid foundation for the future. With real-time, unified data in place, the organization is now positioned to introduce AI-driven personalization and predictive learning models.

Disconnected data became real-time intelligence, turning insight into immediate action.

We turned millions of disconnected student data points into real-time risk intelligence teams could act on immediately

Visit our Insights page for more articles about emerging technology trends, the Education Industry, interviews, and more!