Empowering Every School Across the Czech Republic

Problem

- Educational data is scattered
- Systems don't speak the same language
- Hours wasted on manual cleaning & formatting
- No data engineers in most schools
- Insights delayed → decisions become guesswork

Our Solution

- Fully automated ETL pipeline (inbox → bronze → silver → gold)
- AI transcription, schema unification, validation & modelling
- Structured datasets generated automatically
- Instant dashboards + AI insights (Streamlit + LLM)
- Zero technical knowledge required

The Scaling Challenge

Current State

Until now, the foundation operated primarily in one region using manual processes: Excel spreadsheets, email chains, audio recordings, and one-off reports. This already creates fragmented, inconsistent data that's difficult to leverage.

The Scaling Reality

Now imagine expanding this approach to all 13 regions plus Prague over 3 years. Different data formats, inconsistent quality standards. With manual methods, **this simply doesn't scale**. The foundation risks drowning in data instead of transforming it into actionable insights.

Understanding the Data Landscape

The foundation works with remarkably diverse data sources that span structured and unstructured formats:

Structured Data

Training attendance tables, participant rosters, registration forms with varying completeness and format consistency

Feedback & Surveys

Post-training evaluations, satisfaction surveys, program feedback forms with mixed question types and response quality

Audio Interviews

Czech-language interviews with teachers and school principals discussing experiences, challenges, and outcomes

Qualitative Notes

Transcripts, field notes, follow-up evaluations conducted months after initial interventions to measure lasting impact

Some data arrives clean and standardized, while much of it contains typos, missing values, and inconsistent formats. The critical challenge: connecting information across time - linking what happened during a training with measurable changes in that school 6 months later.

Essential System Requirements

The foundation needs a **practical, production-ready system** that addresses real-world operational needs.

Here's what success looks like:

01

Multi-Source Data Ingestion

Seamlessly accept diverse data types from all regions.

02

Automated Quality Control

Leverage AI to clean, validate, and standardize incoming data.

03

Comprehensive Analytics

Enable meaningful analytics across schools, regions, and years to identify patterns, trends, and successful interventions

04

Privacy & Compliance

Maintain strict GDPR compliance with robust data protection measures, ensuring student and educator privacy at every step

05

Sustainable Operations

Design for deployment and maintenance by a small foundation team without dedicated IT infrastructure or technical specialists

AI as the Essential Assistant

Artificial intelligence isn't a buzzword in this context - it's the fundamental enabler that makes the entire system practical and scalable:

1

Czech Speech/image Recognition

Enhanced speech-to-text capabilities specifically optimized for Czech language interviews, capturing nuanced educator feedback accurately

2

Intelligent Data Validation

Automatic detection of errors, inconsistencies, and patterns in messy tables.

3

Data unification

Data unification transforms scattered, mismatched educational files into a consistent structure so every dataset speaks the same language.

4

Insight Generation

Automated dashboards and insights through natural language processing.

Our goal is making AI a **quiet, reliable assistant** working in the background—transforming raw data into trustworthy, actionable intelligence that foundation staff can confidently use to support schools.

Cost & Maintenance

One important element we cannot ignore is budget. The foundation does not have a large IT department, so any system we propose must be affordable to use.

The good news is that modern data tooling allows us to keep maintenance costs extremely low. Instead of expensive cloud databases or big servers, we rely on lightweight technologies like DuckDB and MinIO, which are open-source and almost cost-free to operate.

In practical terms:

€150-200

Monthly infrastructure costs in production

The Vision: Impact at Scale

When we successfully implement this system, the foundation will have transformed how it operates and creates value:

Seamless Regional Expansion

New regions can begin contributing data immediately without requiring system redesigns or lengthy onboarding processes

Evidence-Based Feedback

Schools and municipalities receive clear, data-driven insights about which interventions truly work and create lasting improvement

Demonstrable Impact

The foundation can present compelling, quantified evidence of real outcomes to donors, government authorities, and stakeholders

---

Most importantly: Teachers and schools receive better, more targeted support delivered faster. That's the vision behind this challenge—and that's precisely what our solution is designed to enable for every school across the Czech Republic.

The next steps

1. In the next phase, we integrate **Airflow** to automate and schedule all regional data pipelines as the system scales.

2. Later, **Kafka** can be added to handle real-time events from schools and regions, enabling instant processing and alerts.

3. Finally, GitHub Actions ensures fully automated testing, deployment, and nightly data-quality checks, keeping maintenance costs minimal.

Yee.ai