DataMender

Smart Cleaning for Large CSV/Parquet Files

AMS 560 / CSE 542 - Fall 2025

Group 8 September 2025

Team Members:

Ahmad Javadi Nezhad • Daniel Bazmandeh

Iliya Mirzaei • Nicholas Tardugno • Tamali Halder

Background & Problem Statement

What Has Been Done

- OpenRefine: Manual data cleaning workflows [1]
- Trifacta/Alteryx: Rule-based transformations [2]
- HoloClean: ML-based anomaly detection [3]

The Gap We Address

No LLM-powered rule discovery for large-scale data quality

Our Objectives

- Fast profiling with Polars for multi-GB files
- LLM-suggested cleaning rules
- Human-in-the-loop validation

Deliverables & Timeline

What We'll Deliver

1. Streamlit cleaning app
2. Polars-based profiler
3. LLM rule discovery engine
4. Reusable configurations
5. Demo with metrics

8-Week Plan

- **Weeks 1-2:** Dataset & profiler
- **Weeks 3-4:** LLM & validation UI
- **Weeks 5-6:** Batch processing
- **Weeks 7-8:** Demo & report

References I

1. OpenRefine. "A free, open source tool for working with messy data." `https://openrefine.org/`
2. Trifacta Inc. "Data Preparation Platform for Analytics & ML." Acquired by Alteryx, 2022.
3. Rekatsinas, T. et al. "HoloClean: Holistic Data Repairs with Probabilistic Inference." *VLDB Endowment*, 2017.