DataMender: Smart Cleaning for Large CSV/Parquet Files

Project Proposal

AMS 560 / CSE 542 - Fall 2025

Team Members:
Ahmad Javadi Nezhad • Daniel Bazmandeh • Iliya Mirzaei • Nicholas Tardugno • Tamali Halder

September 2025

1 Background and Motivation

Data quality issues affect organizations working with large-scale datasets. Current solutions like OpenRefine require manual intervention, while Trifacta/Alteryx rely on rule-based transformations. However, **no existing tool uses Large Language Models (LLMs) for rule discovery** in large-scale data quality management.

Data scientists spend 60-80% of their time on data cleaning tasks, yet lack automated tools that can suggest cleaning rules for multi-gigabyte datasets. This project addresses this need by combining data profiling with LLM-powered rule discovery and human validation.

2 Problem Statement and Importance

Core Problem: Organizations struggle to efficiently clean large CSV/Parquet files due to lack of automated rule discovery systems that can handle multi-gigabyte datasets.

Why This Matters: Data cleaning consumes 60-80% of data science workflows, manual rule creation doesn't scale to large datasets, existing tools lack LLM integration, and poor data quality leads to unreliable analytics and ML models.

Target Use Case: Ride-sharing companies with 5-10GB CSV files containing millions of records with inconsistent formatting, missing values, and data quality issues requiring cleaning rules.

3 Challenges

Key challenges include: (1) Profiling multi-gigabyte files efficiently without memory overflow, (2) Designing effective LLM prompts for rule discovery across different data types, (3) Creating UI tool

for human validation, (4) Ensuring vectorized operations handle large-scale transformations, and (5) Balancing automated suggestions with human oversight.

4 Solution Approach

DataMender combines three key components:

(1) **Data Profiler:** Fast scanning using Pandas/NumPy/Dask to analyze row counts, column types, missing percentages, and histograms.

(2) **LLM Rule Discovery Engine:** Prompts that suggest range constraints, uniqueness checks, monotonicity validation, and sanity checks.

(3) **Human Validation:** UI tool for reviewing, accepting, editing, or rejecting suggested rules.

5 Deliverables and Timeline

8-Week Timeline: Weeks 1-2: Dataset selection and profiler implementation. Weeks 3-4: LLM prompt templates and UI tool. Weeks 5-6: Batch processing engine. Weeks 7-8: Polish, testing, and documentation.

Final Deliverables: (1) Working UI tool for data cleaning, (2) YAML configuration files with discovered cleaning rules, (3) Demonstration video showing complete workflow, (4) Technical report with performance metrics and lessons learned.

6 Division of Work Among Group Members

Ahmad Javadi Nezhad: Project coordination, LLM integration, and rule discovery engine development.

Daniel Bazmandeh: Data profiler implementation and optimization.

Iliya Mirzaei: Validation UI design and UI interface development.

Nicholas Tardugno: Batch processing engine and data transformation implementation.

Tamali Halder: Testing, metrics collection, documentation, and final report preparation.

All members will contribute to dataset selection, testing, and final presentation preparation.

Expected Impact: DataMender will demonstrate how LLM-powered rule discovery can improve large-scale data cleaning workflows, providing a foundation for future research in AI-assisted data quality management.