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

Concrete Application: DataMender targets *ride-sharing companies* processing 5-10GB daily trip datasets with inconsistent formatting, missing timestamps, and duplicate records. Current solutions like OpenRefine require manual rule creation, while Trifacta/Alteryx use predefined transformations without LLM integration.

The Gap: No existing tool combines **automated LLM-powered rule discovery** with **human validation workflows** for large-scale data quality management. Data scientists spend 60-80% of their time on data cleaning, yet lack tools that can suggest context-aware cleaning rules for multi-gigabyte datasets.

Recent Work (2024-2025): Recent studies show LLMs can generate data transformation rules, but existing implementations lack: (1) Large-scale file handling, (2) Human-in-the-loop validation, (3) Domain-specific rule templates, and (4) Hallucination mitigation strategies.

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 and Novelty

Technical Challenges: Profiling multi-gigabyte files without memory overflow, designing effective LLM prompts, creating intuitive validation UI, ensuring vectorized operations for large-scale transformations.

LLM Hallucination Mitigation: (1) **Confidence Scoring:** Each suggested rule includes confidence metrics based on data coverage and pattern consistency, (2) **Multi-Model Validation:** Cross-validate rules using GPT-4, Claude-3, and local models, (3) **Human Validation Required:** All rules must pass human review before application, (4) **Reversible Operations:** All transformations are logged for easy rollback, (5) **Test Data Validation:** Rules tested on sample data before full application.

Novel Contributions: (1) First LLM-powered rule discovery for large-scale CSV/Parquet files, (2) Human-in-the-loop validation workflow, (3) Domain-specific templates for ride-sharing data, (4) Hallucination-resistant rule generation pipeline.

4 Solution Approach and Competitors

DataMender Components: (1) **Data Profiler:** Fast scanning using Pandas/NumPy/Dask, (2) **LLM Rule Discovery Engine:** Multi-model prompts (GPT-4, Claude-3) with confidence scoring, (3) **Human Validation:** Interactive UI for reviewing/editing rules.

Key Competitors: OpenRefine (manual rules, no LLM), Trifacta/Alteryx (expensive, rule-based only), HoloClean (anomaly detection only, no rule generation).

DataMender's Competitive Advantages: (1) **First LLM-powered rule discovery** for large-scale data, (2) **Human-in-the-loop validation** preventing hallucinations, (3) **Domain-specific templates** for ride-sharing data, (4) **Multi-model validation** reducing false positives, (5) **Free and open-source** unlike enterprise tools.

5 Deliverables and Work Division

8-Week Timeline: Weeks 1-2: Dataset selection and profiler. Weeks 3-4: LLM prompts and UI tool. Weeks 5-6: Batch processing engine. Weeks 7-8: Testing and documentation.

Final Deliverables: (1) Working UI tool, (2) YAML configuration files, (3) Demo video, (4) Technical report with performance metrics vs. GPT-4, OpenRefine, Trifacta baselines.

Work Division: **Ahmad:** Project coordination, LLM integration, rule discovery engine development. **Daniel:** Data profiler implementation and optimization. **Iliya:** Validation UI design and interface development. **Nicholas:** Batch processing engine and data transformation. **Tamali:** Testing, metrics collection, documentation, and final report preparation.

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

Expected Impact: DataMender demonstrates LLM-powered rule discovery for large-scale data cleaning workflows, providing foundation for future AI-assisted data quality research.