# Product Requirements Document (Generated)

\*\*📌 Source: `reddit\_datacleaning\_hot\_500.json`\*\*

\*\*📊 Clusters analyzed:\*\* 1, 3, 7 (based on 59 posts)

\*\*📂 Pain Points Selected:\*\*

- Data professionals struggle with efficient data cleaning and manipulation techniques, particularly when dealing with inconsistencies like overlapping date ranges and outliers. They need robust automated solutions for validating data, identifying problematic observations, and performing granular time-based data extraction and splitting for analysis and visualization. Furthermore, there's a need for streamlined methods to clean data directly within databases, avoiding cumbersome import/export processes.

- Many users struggle with cleaning and standardizing messy, unstructured data from various sources, including free text logs, CSV files, scanned documents, and XML. This data often contains inconsistencies, typos, incomplete information, and varying formats, hindering analysis and data integration. The need for efficient automated solutions to extract, clean, and standardize this data is consistently highlighted.

- The pervasive challenge across various industries is ensuring data quality, encompassing issues like inconsistent formatting (e.g., phone numbers), incomplete or uncleaned datasets (e.g., hospital prices), and biased or unrepresentative data leading to skewed insights (e.g., tourism data). This poor data quality results in significant financial losses for businesses and hinders accurate analysis and decision-making. The lack of readily available resources and clear understanding of data quality itself further compounds these problems.

💡 \*To review raw discussions behind these pain points, open `cluster\_visualization.html` in your browser.\*

This file is located in the same folder as this PRD.

---

\*\*PRD Draft: DataCleanse AI\*\*

\*\*Problem Summary:\*\* Data professionals spend excessive time cleaning and preparing data from diverse sources, hindering analysis and decision-making. Current methods are inefficient for handling inconsistencies, unstructured data, and ensuring data quality across various formats and industries. This leads to delays, inaccurate insights, and ultimately, lost business value.

\*\*Why This Problem Matters:\*\* For data professionals, inefficient data cleaning translates directly into lost productivity and missed deadlines. Poor data quality impacts the accuracy of analyses, leading to flawed business decisions and potentially significant financial losses. The inability to efficiently process unstructured data limits the potential insights that can be derived, hindering innovation and competitive advantage.

\*\*Potential Solution Overview:\*\* DataCleanse AI is an automated data cleaning and preparation platform designed to streamline the entire data lifecycle for data professionals. It will offer intelligent automation for data validation, outlier detection, time-series data manipulation, and the cleaning and standardization of unstructured data from various sources. The solution will integrate directly with databases, minimizing the need for cumbersome import/export processes.

\*\*Suggested MVP Features:\*\*

\* \*\*Automated Data Quality Assessment:\*\* Identifies and flags inconsistencies, outliers, and potential biases within datasets, providing a comprehensive quality report.

\* \*\*Intelligent Data Cleaning Engine:\*\* Automates data cleaning tasks such as handling missing values, standardizing formats (e.g., dates, phone numbers), and resolving inconsistencies in overlapping date ranges. Supports various data formats (CSV, XML, free text).

\* \*\*Unstructured Data Extraction and Cleaning:\*\* Extracts key information from scanned documents, free text logs, and other unstructured sources using NLP techniques, then cleans and standardizes this extracted data for analysis.

\* \*\*Database Integration:\*\* Allows direct data cleaning and manipulation within databases (e.g., SQL, NoSQL), eliminating the need for data import/export.

\* \*\*Granular Time-Based Data Splitting:\*\* Enables users to easily split time-series data into meaningful segments for analysis and visualization.

\*\*Next Steps:\*\* Conduct user interviews to validate proposed features and gather additional requirements. Develop a functional prototype demonstrating key MVP features. Plan a sprint to build and test the core data quality assessment and intelligent data cleaning engine features.