Requirement Document: Universal File Loader

1. Problem Statement

In our ecosystem, data is scattered across multiple unstructured file formats such as:

- PDF
- Word Documents (DOC/DOCX)
- PowerPoint (PPTX)
- Excel (XLS/XLSX)
- CSV
- Text Files (TXT)
- JSON, XML, HTML

While modern vector databases are useful for working with unstructured data semantically, our existing legacy systems are built on structured databases like:

- MySQL
- PostgreSQL
- SQL Server
- MongoDB

These systems rely on rigid tabular schemas, and current workflows involve manual extraction, cleaning, and transformation, which is time-consuming and repetitive for analysts.

2. Objective

To address this challenge, we propose the development of a feature called:

Universal File Loader

This feature will enable automatic ingestion, extraction, prompt-based transformation, and structuring of data from various file formats.

3. Key Features

3.1 File Ingestion

Requirement Document: Universal File Loader

- Upload support for multiple file types: PDF, DOCX, PPTX, XLS/XLSX, CSV, TXT, JSON, XML, HTML

- Extract readable content including text, tables, lists, headers/footers, metadata

3.2 AI-Powered Data Extraction

- Automatically extract and convert relevant information into a structured format (like CSV or JSON)

- Differentiate and preserve structured data, semi-structured data, and contextual paragraphs

3.3 Prompt-Based Data Interaction

- Users can enter natural language prompts to:

  - Modify data (e.g., "Change all values in column 'Status' from 'Pending' to 'Open'")

  - Perform arithmetic operations (e.g., "Add 10% to all prices", "Calculate total revenue as price * quantity")

  - Filter rows, merge/split columns, rename fields, remove duplicates, format values

3.4 Structured Output & Integration

- Output transformed data in any structured format: CSV, Excel, JSON

- Allow field mapping and custom schema definition

3.5 Learning & Automation

- Learn from user prompts and transformations

- Suggest actions and offer auto-corrections or smart defaults

6. Enhancements to Maximize Analyst Productivity

6.1 Data Type Detection & Validation

- Detect data types per column and flag anomalies

- Suggest or auto-apply fixes to ensure data quality

6.2 Prebuilt Transformation Templates

- Provide a library of reusable prompt templates for common tasks

- Auto-suggest based on uploaded file patterns

6.6 Data Deduplication & Integrity Checks

- Detect duplicates, null values, and inconsistencies
- Apply user-defined rules for cleaning and merging data

7. Technical Considerations

- Use AI/LLM for prompt understanding and data transformation
- Modular file parsers for various formats
- Backend integration with structured DBs (MySQL, PostgreSQL)
- Optional integration with vector DBs
- Maintain logs, versioning, and prompt histories

8. Outcome

- Unified tool to ingest, analyze, and structure data from any source
- Significant time-saving for data analysts
- Reduced manual cleanup and repetitive operations
- Better integration with existing and future systems