Project Overview

The project is a data cleaning automation agent designed to preprocess CSV files using both rule-based and LLM-powered quality checks. The agent runs in a workflow orchestrated by n8n and interacts with external APIs as needed.Data_Cleaning-agent-4.json

Tools and Technologies

- **n8n workflow**: The core automation and integration platform orchestrating the data flow and node execution.Data_Cleaning-agent-4.json
- **Custom JavaScript (n8n "code" node)**: Used for data profiling, preliminary cleaning, and summarization of CSV columns.Data_Cleaning-agent-4.json
- **Webhook (n8n "webhook" node)**: Receives CSV uploads via HTTP POST requests.Data_Cleaning-agent-4.json
- **File Extraction (n8n "extractFromFile" node)**: Reads and parses the uploaded CSV file.Data_Cleaning-agent-4.json
- **HTTP Request (n8n "httpRequest" node)**: Calls an external Hugging Face LLM endpoint to perform advanced data analysis and recommendations.Data_Cleaning-agent-4.json
- **Set Node ("set")**: Prepares JSON payloads for LLM requests.Data_Cleaning-agent-4.json

Workflow Structure

1. **Webhook - Receive CSV**
    - Receives HTTP POST requests at `/data-cleaner-1` endpoint, accepting CSV files as binary data in the `file` property.Data_Cleaning-agent-4.json
2. **Extract From File CSV**
    - Parses the incoming CSV file and outputs row-wise JSON data for downstream processing.Data_Cleaning-agent-4.json
3. **Profile Summary ("code" node)**
    - Runs JavaScript to:
        - Get all column names

        - Generate column-wise summaries (missing, unique values, inferred type)
        - Extract first five data rows as a sample
        - Return error if the file is emptyData_Cleaning-agent-4.json

4. **Prepare LLM Body ("set" node)**
    - Structures a JSON payload containing the profile summary and data sample for the LLM.Data_Cleaning-agent-4.json

5. **HTTP Request to LLM**
    - Sends a POST request to the Hugging Face endpoint with the sample and summary, securing the request with an API key.Data_Cleaning-agent-4.json
    - The LLM is instructed to return only a valid JSON object with the following keys: `issues`, `duplicates`, `typemismatches`, `summary`, `recommendations`.Data_Cleaning-agent-4.json

6. **Respond to Webhook**
    - Sends the response from the LLM back to the user as the final output.Data_Cleaning-agent-4.json

Detailed Flow Diagram

Step	Tool/Node	Purpose/Description
1	Webhook	HTTP POST endpoint to receive CSV files Data_Cleaning-agent-4.json
2	ExtractFromFile	Parses CSV file to JSON Data_Cleaning-agent-4.json
3	Profile Summary (Code)	Profiles data: missing, uniqueness, type, sample Data_Cleaning-agent-4.json
4	Set (Prepare LLM Body)	Assembles JSON payload for LLM analysis Data_Cleaning-agent-4.json
5	HTTP Request (LLM)	Sends data to Hugging Face LLM for advanced QC Data_Cleaning-agent-4.json
6	RespondToWebhook	Sends analysis results back to user Data_Cleaning-agent-4.json

API Endpoints and Integration

- **/data-cleaner-1 (POST)**
    - Accepts a CSV file upload, processes the file, orchestrates the workflow, and returns AI-driven cleaning suggestions and profile stats.Data_Cleaning-agent-4.json
- **External LLM (Hugging Face endpoint)**
    - Used for AI-enriched data cleaning insights.Data_Cleaning-agent-4.json

Value Proposition

- **Automation**: Reduces time and manual effort for data profiling and quality checks.Data_Cleaning-agent-4.json
- **Scalable and Reusable Workflow**: Extensible via n8n for more nodes and complex transformations.Data_Cleaning-agent-4.json
- **Quality Insights**: Advanced LLMs deliver nuanced recommendations and identify common data issues (missing values, duplicates, mismatches).Data_Cleaning-agent-4.json
- **Plug-and-play**: Easily integrated into data pipelines via a simple API endpoint.Data_Cleaning-agent-4.json
- **Immediate Feedback**: Returns actionable data cleaning suggestions in real time, supporting agile data teams.Data_Cleaning-agent-4.json