**Technical Documentation of Data-Quality Product**

**Contents**

1. Technology Stack
   1. Data Loading Requirements
   2. Data Quality Metrics requirements
   3. Detailed Report Generation
   4. Quality Summary Generation
2. Product Architecture
3. Upload Dataset
4. Dataset statistics
5. Quality Metric Section
6. Quality Score
7. Missing Value Analysis
8. Alerts
9. Column-Wise Visualizations
10. Quality Summary
11. Sample Dataset
12. Chatbot Assistant
13. Column Analytics
14. Correlation Matrix
15. **Technology Stack**

|  |  |
| --- | --- |
| **Component** | **Technology** |
| Operating System | Windows 11 & above |
| Visual Studio Code | Version 1.96 |
| LLM | Gemini 2.0 Flash-Lite |
| Programming Language | Python (3.12.7) , Flask 3.1.0 |
| Web Requests | requests (for downloading datasets from URLs) |
| Data Handling | Pandas (2.2.3), NumPy (2.0.2) |
| Data Visualization | Matplotlib (3.9.4), Seaborn (0.13.2) |
| Reporting | HTML (generated reports), CSS, JavaScript |

* 1. **Data Loading Requirements**

|  |  |
| --- | --- |
| Data Source | Kaggle |
| Data Loading | Pandas v2.2.3 |

* 1. **Data Quality Metrics requirements**

|  |  |
| --- | --- |
| Dataframe | Pandas v2.2.3 |
| Programming | Python v3.12.7 |
| Library | NumPy v2.0.2 |

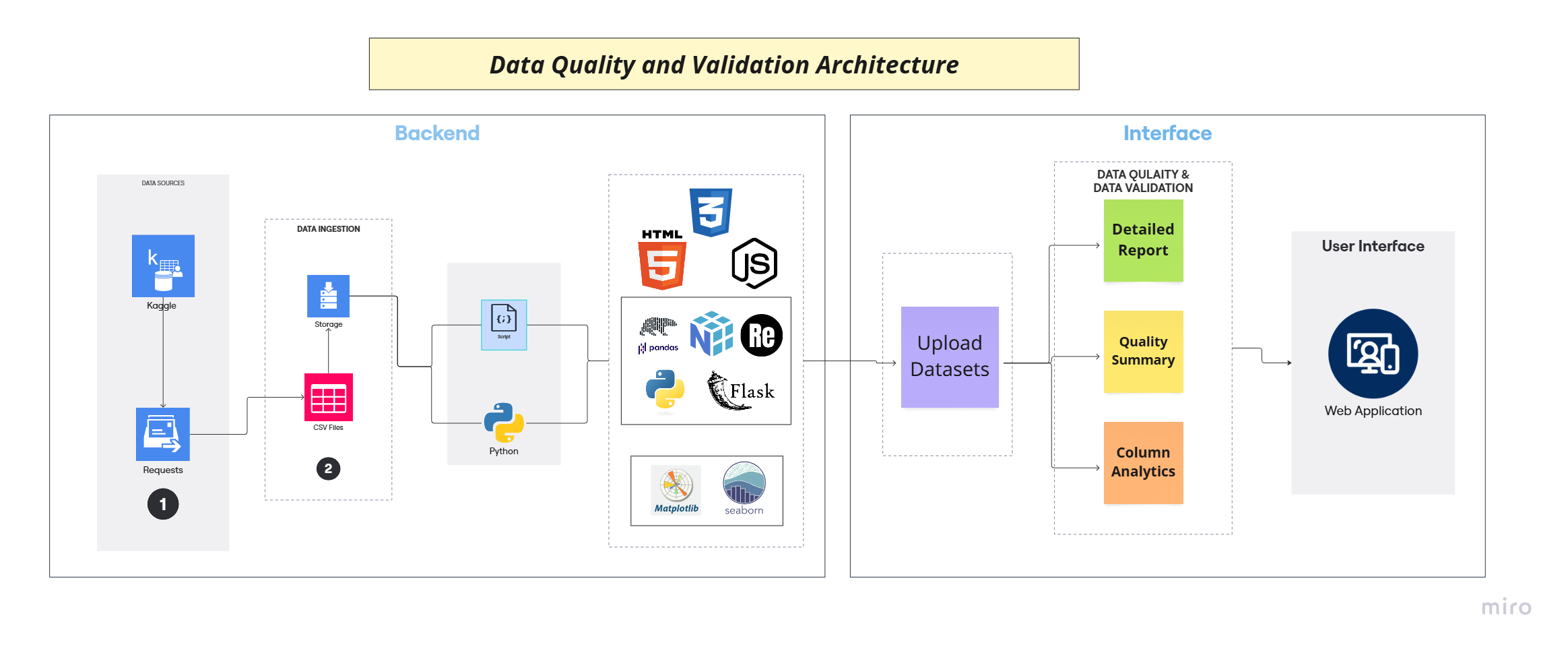
* 1. **Detailed Report Generation**

|  |  |
| --- | --- |
| Dataframe | Pandas v2.2.3 |
| Visualizations | Matplotlib v3.9.4, Seaborn v0.13.2 |
| Report | HTML, CSS, JavaScript |

* 1. **Quality Summary Generation**

|  |  |
| --- | --- |
| Quality Metrics | Python v3.12.7, NumPy v2.2.3 |
| Dataframe | Pandas v2.2.3 |

1. **Product Architecture**
2. This document provides technical documentation for the Data Quality Validation process.
3. It evaluates datasets to measure and ensure data integrity.
4. The system computes data quality scores based on predefined metrics.
5. It generates both detailed and summary reports for analysis.
6. It offers a clear assessment of dataset quality.
7. The process includes:
   * Loading datasets into the system.
   * Calculating data quality metrics such as completeness, accuracy, consistency, and uniqueness.
   * Generating reports to provide insights into data health.

****

1. **Upload Dataset**

The system requires users to upload two datasets in **CSV format** as input for the **Data Quality Validation** process.

A screenshot of a data quality validation system

AI-generated content may be incorrect.

1. **Dataset statistics**

After uploading, the system provides a **basic dataset overview**, including:

1. **Total Rows & Columns** – Dataset dimensions.
2. **Missing Values** – Count and percentage per column.
3. **Unique Values** – Number and percentage per column.
4. **Data Types** – Identifies column data types.

Screens screenshot of a computer

AI-generated content may be incorrect.

1. **Quality Metric Section**

The system dynamically generates the **Quality Metrics** section, allowing users to customize and display only the selected metrics.

1. **Customizable Metrics** – Users can choose specific quality metrics to display.
2. **Real-Time Calculation** – Only selected metrics are processed and shown.
3. **Percentage Representation** – Each metric is displayed with its computed percentage for better clarity.

A screenshot of a survey

AI-generated content may be incorrect.

**6. Quality Score**

* The system calculates an **Overall Data Quality Score** based on selected quality metrics.
* It is computed dynamically, reflecting real-time assessments of completeness, accuracy, consistency, and other chosen metrics.
* A higher score indicates **better data quality**, helping users quickly gauge the dataset's overall health and suitability for further processing.

A screenshot of a score

AI-generated content may be incorrect.

**7. Missing Value Analysis**

The system provides a **visual representation** of missing values using a **bar chart**, making it easier to identify data gaps across columns. Each column is analyzed for:

* + **Present Values** – The number of non-missing entries.
  + **Missing Values** – The count of missing entries.

A screenshot of a graph

AI-generated content may be incorrect.

**8. Alerts**

The **Alerts** section provides **real-time warnings** about potential data quality issues, helping users quickly identify and address anomalies. The system generates alerts for:

* **Missing Values** – Highlights columns with significant data gaps.
* **High Duplicate Count** – Warns about excessive duplicate values.
* **Low Distinct Values** – Flags columns with minimal variation.
* A screenshot of a computer

  AI-generated content may be incorrect.**High Memory Usage** – Identifies columns consuming excessive storage.

**9. Column-Wise Visualizations**

It provides clear insights into data quality with:

* Bar Chart – Shows all the metric scores for specific column.
* The heatmap highlights column-level data metric scores, using color codes

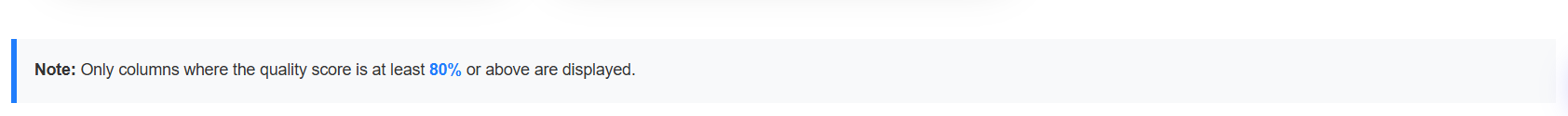
A close-up of a graph

AI-generated content may be incorrect.

**10. Quality Summary**

The **Quality Summary** section highlights only the columns that achieve an **80% or higher score** in the quality metrics. This ensures that users focus on **high-quality data** while filtering out lower-performing columns.

A screenshot of a computer

AI-generated content may be incorrect.

**11. Sample Dataset**

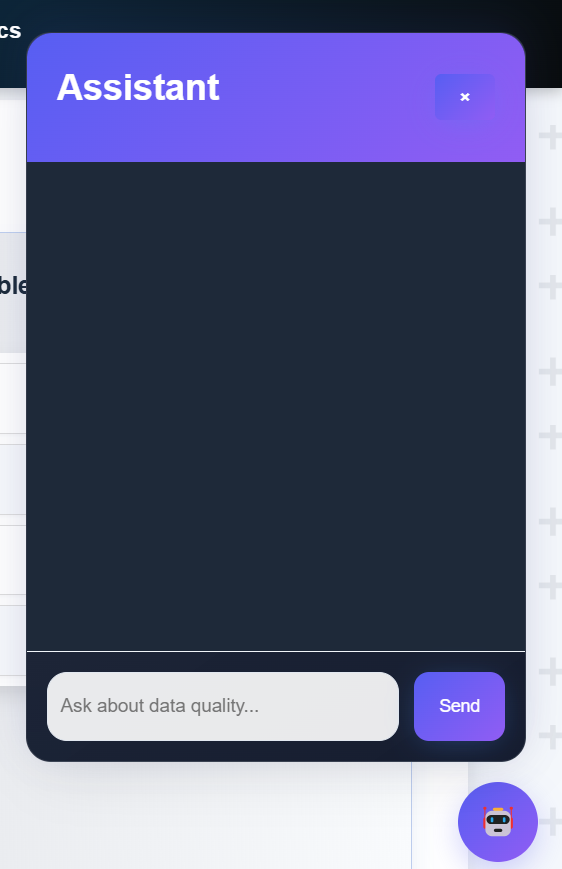
The **Sample Dataset** section displays a preview of the **first 10 and last 10 rows** of the uploaded dataset. This helps users quickly inspect the dataset's structure, data types, and quality from both the beginning and end of the data.

A screenshot of a calendar

AI-generated content may be incorrect.

**12. Chatbot Assistant**

The **Chatbot Assistant**, powered by **Gemini Free LLM**, provides real-time support for users navigating the data quality validation process. It explains quality metrics, and offers troubleshooting recommendations for common data issues.



**13. Column Analytics**

The **Column-Wise Analytics** section provides a **detailed breakdown** of each column's data characteristics, helping users assess data integrity and efficiency. For every column, the system analyzes:

* + **Missing Values** – Count and percentage of empty cells.
  + **Duplicate Values** – Number of repeated entries.
  + **Distinct Values** – Unique values present in the column.
  + **Memory Usage** – The amount of memory consumed by the column.

A screenshot of a computer

AI-generated content may be incorrect.

**14. Correlation Matrix**

The **Correlation Matrix** provides a **visual representation** of relationships between numerical columns in the dataset. It helps users identify **patterns, dependencies, and potential multicollinearity** by measuring how strong variables are related.

A screenshot of a graph

AI-generated content may be incorrect.