# System Design: Automated Data Quality Monitoring in Cloud Warehouses

## 1. Architecture Overview

+--------------------+  
| User Interface |  
| (Gradio Dashboard)|  
+--------------------+  
 |  
 v  
+---------------------+  
| CSV File Upload |  
| (.csv from user) |  
+---------------------+  
 |  
 v  
+------------------------+  
| Data Processing Layer |  
| - pandas (read CSV) |  
| - basic preprocessing |  
+------------------------+  
 |  
 v  
+-----------------------------+  
| Data Validation Engine |  
| (Great Expectations) |  
| - Null checks |  
| - Uniqueness validation |  
| - Row count checks |  
| - Duplicate detection |  
+-----------------------------+  
 |  
 v  
+------------------------------+  
| Visualization & Reporting |  
| - Matplotlib & Seaborn |  
| - Summary Stats (JSON) |  
| - Validation Log (Textbox) |  
| - Missing Values Chart (PNG) |  
+------------------------------+  
 |  
 v  
+--------------------------+  
| Cloud Deployment (Future)|  
| - AWS / Azure / GCP |  
| - Containerization (Docker)|  
| - Integration with ETL |  
+--------------------------+

## 2. Component Breakdown

Frontend (Gradio Interface)  
- CSV upload and results display  
  
Data Handling (pandas)  
- Reads and processes the CSV  
  
Validation Engine (Great Expectations)  
- Null, uniqueness, row count, and duplicate checks  
  
Visualization (matplotlib & seaborn)  
- Missing values chart  
  
Port Manager (socket)  
- Auto port assignment  
  
Logger  
- Logs all major events

## 3. Data Flow Diagram

[User Uploads CSV]  
 |  
 v  
 [CSV Read (pandas)]  
 |  
 v  
[Validate (Great Expectations)]  
 |  
 v  
[Summarize + Visualize]  
 |  
 v  
[Display on Gradio Dashboard]

## 4. Optional Cloud Integration (Future Scope)

- Cloud Storage Support: Read directly from S3, Azure Blob, GCS  
- CI/CD Integration: Trigger validation on data pipeline updates  
- Dockerization: Package app for scalable deployment  
- Authentication Layer: Secure access via user logins or tokens

## 5. Output Components

Summary JSON: {"Total Rows": 500, "Missing Values": 45, ...}  
Validation Log: - expect\_column\_values\_to\_not\_be\_null: ✅ Passed  
Image (Chart): Bar chart of missing values

## 6. Security Considerations

- Local file processing only (no upload to cloud)  
- Can be sandboxed using Docker  
- Extendable with HTTPS and login for production use

## 7. Suggested Enhancements

- Support Excel, Parquet, and JSON formats  
- Custom rule definitions via UI  
- Schedule-based auto-analysis for new datasets  
- Store historical analysis results