International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

DATA MANAGEMENT SYSTEM

[1] Dr. RAVINDRA S, [2] KEERTHANA V S

[1] Department of Electronics and Communication Engineering, Dayananda Sagar College of Engineering , Bengaluru, India ravindra-ece@dayanandasagra.edu

[2] Department of Electronics and Communication Engineering, Dayananda Sagar College of Engineering , Bengaluru, India keerthivs126@gmail.com

I ABSTRACT :

In today's data-driven world, effective data management is crucial for ensuring the accuracy, security, and accessibility of information across various domains. A Data Management System (DMS) is a software solution designed to collect, store, organize, and retrieve data efficiently while supporting data integrity and scalability. This project focuses on the design and implementation of a robust DMS that facilitates streamlined data operations such as insertion, update, deletion, and querying, with a user-friendly interface. It employs structured data models and integrates access control mechanisms to ensure authorized usage and privacy compliance. The system is scalable to support growing data volumes and adaptable to multiple use cases, including enterprise resource planning, academic databases, and government records. Through the implementation of this system, organizations can improve decision-making processes, reduce redundancy, and maintain high data quality standards.

Keywords: Data Management, Database System, Data Integrity, Access Control, Scalability

II INTRODUCTION

In the modern digital age, data has become one of the most valuable assets for organizations, institutions, and individuals. With the exponential growth of data generated through various sources such as online transactions, sensors, user interactions, and enterprise operations, the need for an efficient and reliable system to manage this data has become critical. A Data Management System (DMS) provides a structured approach to storing, retrieving, updating, and securing data. It ensures that information is readily available, consistent, and protected against unauthorized access. The primary goal of a DMS is to support users in efficiently handling large volumes of data while maintaining its quality, accuracy, and integrity. It also enables data sharing among multiple users or departments, supports decision-making processes, and reduces redundancy. Modern DMS solutions are built with scalability in mind, allowing them to adapt to growing data demands and various domains such as education, healthcare, finance, and logistics. This project focuses on the development and implementation of a Data Management System that simplifies data handling through a user-friendly interface, incorporates essential data operations, and ensures security and accessibility through access control mechanisms. By leveraging this system, organizations can enhance operational efficiency, maintain high data quality, and make informed decisions based on accurate and timely information.

III BLOCK DIAGRAM

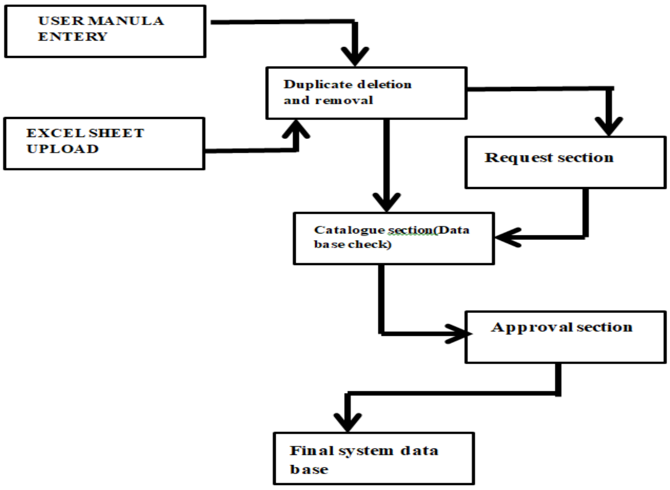


Fig No .1 Block diagram

- Data Entry & Upload: Users enter material master data either manually (User Manual Entry) or by bulk upload (Excel Sheet Upload).
- Duplicate Check & Removal: The system performs duplicate detection and removal to ensure data integrity and eliminate redundant records.
- Database Validation: The cleaned data undergoes a catalogue section (database check) to verify if the material already exists in the system.
- Request & Approval Workflow: If required, data is sent to the Request Section for additional review and then to the Approval Section for final validation.
- Final Data Storage : Once approved, the data is stored in the Final System Database, making it available for enterprise-wide use.

IV METHODOLOGY

The proposed Data Management System follows a structured methodology to ensure efficient data processing and integrity. The process begins with two modes of data entry: manual user input and Excel sheet upload. Manual entry allows users to input individual data records directly through a user interface, while the Excel sheet upload feature facilitates the bulk addition of records using preformatted spreadsheets. Once the data is entered into the system, it undergoes a duplicate deletion and removal process, where the system identifies and eliminates redundant or repeated entries to maintain data consistency and prevent storage of duplicate records.

After cleaning, the data proceeds to the catalogue section, where it is cross-checked against the existing database. This step ensures that no existing records are re-entered and validates the authenticity and necessity of the new data. If any discrepancies or incomplete information are detected, the data is directed to the request section, which flags the records for further clarification or additional input from authorized users. Once the issues are resolved, the data moves to the approval section, where it is reviewed by designated personnel. Only verified and approved entries are permitted to move forward.

The final stage involves updating the system database with clean, validated, and approved data. This multi-step approach ensures high data quality, reduces redundancy, supports data traceability, and maintains secure and reliable storage within the organization’s system.

V TOOLS USED

stream lit: Streamlit is an open-source Python library that allows you to build and share beautiful, custom web apps for machine learning, data science, and other data-driven projects easily.

Unlike traditional web frameworks (like Django, Flask), Streamlit lets you create web apps without needing HTML, CSS, or JavaScript.

It’s designed for fast prototyping — from Python script to web app in minutes.

Write a simple Python script → Run → Get a full interactive dashboard.

Streamlit auto-renders the UI from Python code, speeding up development.

Key features:

Python-Only Development

- Streamlit allows you to build web applications completely in Python.
- No need for frontend skills (like HTML, CSS, JavaScript).
- This makes it ideal for data scientists, analysts, engineers, and students who already know Python.

Fast Prototyping

- You can go from *idea → working web app* in just a few hours.
- Write a simple Python script → Run → Get a full interactive dashboard.
- Streamlit auto-renders the UI from Python code, speeding up development.

Auto-Reloading / Reactive Programming

- Whenever you *edit and save* the Streamlit app code, the app *automatically refreshes* in the browser.
- Also, when a *user interacts* (uploads a file, clicks a button), *Streamlit re-runs the script* from top to bottom.
- This "reactive" model makes *dynamic apps easy to build* without managing backend/frontend separately.

Built-in File Upload & Download

- st.file_uploader() lets users upload files directly into the app (CSV, Excel, images, etc.).
- st.download_button() allows exporting processed data or reports.
- In your project, you allowed uploading Material Master Excel/CSV files and processed them live.

VISIUAL STUDIO (VS CODE For programming):

Python is a high-level, interpreted, and general-purpose programming language.

Python emphasizes code readability and allows programmers to express concepts in fewer lines of code compared to many other languages.

Usage :

- It handled all data processing, logic building, file reading, and generating real-time insights for your Material Master Data Quality Dashboard.
- Load and read Excel/CSV files.
- Perform data cleaning and transformations.
- Calculate metrics (like missing values, duplicate counts).

- Generate the processed results to be shown on the Streamlit dashboard.

Features:

1. **Data Processing and Manipulation with Pandas and NumPy**

- Pandas helped you handle large datasets easily.

Used Pandas to:

- Load CSV and Excel files into DataFrames.
- Perform data cleaning (detect missing/null values).
- Filter unnecessary data.
- Group, aggregate, and summarize important fields.
- Create new columns dynamically based on logic.

2. Handling File Inputs, Data Filtering, and Transformations

- Python was used to accept file uploads and automatically read the format (CSV or Excel).
- You built conditions to check if the uploaded file was valid and readable.

Pandas library-

Pandas is a powerful Python library primarily used for data manipulation, analysis, and cleaning.

In your project, Pandas helped you handle the Material Master data — reading uploaded files, cleaning them, and calculating important insights about **data quality.**

Used Pandas to:

Read CSV and Excel files uploaded into the dashboard.

Clean the data (handle missing values, remove duplicates).

Analyze the material data to generate quality metrics like completeness, consistency, and duplication rates.

Features:

- Reading and Writing CSV and Excel Files
- Pandas provides simple functions like read_csv() and read_excel() to quickly load data from files into memory.
- You used these methods to read the Material Master files uploaded by users through the Streamlit app.
- After processing or cleaning, you could also export the results back into CSV or Excel using to_csv() or to_excel() if needed.
- Data Cleaning (Handling Missing Values, Duplicates)
- MissingValues:
  Pandas provided functions like isnull(), fillna(), and dropna() to easily detect and handle missing values in important columns (like Material Description, Group, etc.).
- Duplicates:
  You checked for duplicate Material Numbers using duplicated() and drop_duplicates() to ensure no repeated entries existed.
- Aggregating Data and Performing Calculations
- Grouped, filtered, and summarized the data to understand overall data quality.
- Pandas functions like groupby(), count(), and agg() .

Numpy library

Usage

- NumPy was used to perform numerical operations on datasets.
- Helped in fast, efficient calculations over large material master data (like completeness %, missing values %, etc.)

F**eatures**:

Working with Large Datasets and Arrays

- Instead of using slow loops, you learned how NumPy arrays process entire datasets at once.
- Especially when data was loaded into Data Frames (via Pandas), NumPy made operations memory efficient and super fast.

Performing Mathematical and Statistical Operations

- You calculated mean, median, sum, counts, missing data ratios, etc., quickly using NumPy functions.
- Essential for creating quality scores and metrics in the dashboard.

3. Broadcasting

- Broadcasting allows NumPy to perform operations between arrays of different shapes automatically, without writing loops.
- You can add a single value to an entire array, or perform operations between arrays with different dimensions easily.

4. **Advanced Indexing and Slicing**

- NumPy allows faster and smarter slicing of arrays compared to normal Python lists.
- You can easily select, modify, or filter specific parts of arrays based on conditions.

***5*. Handling Missing Values (NaN Support)**

- NumPy can identify, ignore, or replace NaN values easily using specialized functions like np.isnan(), np.nanmean(), etc.

Plotly Library

Usage:

- Plotly was used to create interactive visualizations (dynamic charts and graphs) directly inside your Streamlit dashboard.
- It made data exploration easier for the user: zooming, clicking, hovering over data points.

Features

1. Creating Interactive Plots

- You built *bar charts*, *line charts*, and possibly *scatter plots* to visualize:
- Material completeness status.
- Duplicates across different material groups.
- Missing fields percentages.

2. Enhancing User Experience

- With *zoom able*, *clickable*, *hover-over* charts, the dashboard became *more intuitive*.
- Users could explore patterns themselves without extra filters.

3. Customization and Styling

- Plotly allows *highly customizable* charts:
- You can change colors, labels, titles, axis ranges, font styles, and background themes easily.
- You can even match the dashboard's branding or corporate color scheme.

Matplotlib

Usage:

Matplotlib was used for static visualizations — simple charts that don't require user interaction.

It helped you create standard plots quickly when interaction was not necessary.

Features:

1. **Plotting Static Graphs**

- Histograms (e.g., distribution of missing fields).
- Pie Charts (e.g., proportion of clean vs dirty data).
- Simple Line/Bar Charts for showing overall material trends.

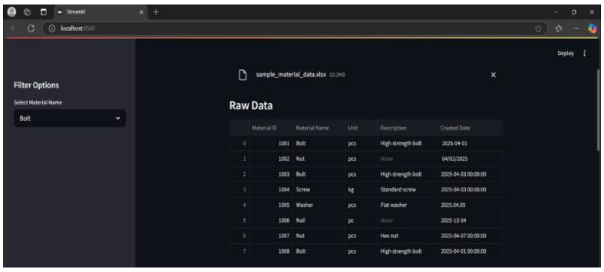
Subplots and Layout Management

- Matplotlib allows you to create multiple plots (subplots) in a single figure window.
- You can organize your dashboard visuals neatly:
- 1 plot showing missing values
- 1 plot showing material completeness
- 1 plot showing duplicates, all in one page

Vi RESULTS

Data overview

The raw dataset for the Material Master Data Quality project consisted of 8 material records, each described through key fields essential for inventory and procurement processes. These fields included Material ID, Material Name, Unit of Measurement, Description, and Created Date. The Material ID and Name served as primary identifiers for each item, while the Unit defined the measurement standard, and the Description provided additional context about the material. The Created Date captured when the material was initially registered in the system. This foundational dataset formed the basis for assessing data quality parameters such as completeness, duplication, and consistency, setting the stage for further analysis and quality improvement initiatives.



Filter Options

Select Material Name: Bolt

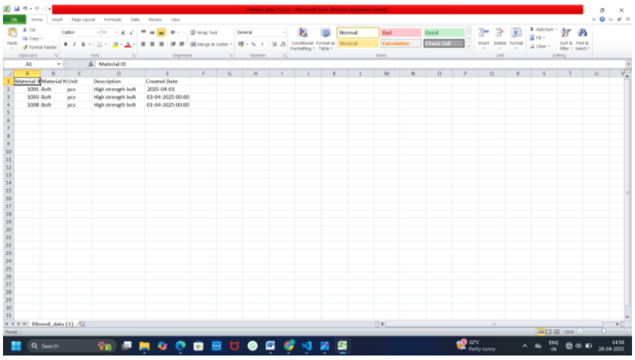
Raw Data

	Material ID	Material Name	Unit	Description	Created Date
0	1001	Bolt	pcs	High strength bolt	2025-04-01
1	1002	Nut	pcs	None	04/02/2025
2	1003	Bolt	pcs	High strength bolt	2025-04-03 00:00:00
3	1004	Screw	kg	Standard screw	2025-04-03 00:00:00
4	1005	Washer	pcs	Flat washer	2025.04.05
5	1006	Nail	pc	None	2025-13-04
6	1007	Nut	pcs	Hex nut	2025-04-07 00:00:00
7	1008	Bolt	pcs	High strength bolt	2025-04-01 00:00:00

Fig No.2 Sample material raw data sheet

Data competency

The completeness analysis revealed that the majority of the critical fields in the material master data were fully populated, demonstrating strong data maintenance practices for core attributes such as Material ID, Material Name, Unit of Measurement, and Created Date. However, the Description field showed a noticeable gap, with only 75% of the entries filled. Missing descriptions can impact material identification and procurement decisions, highlighting an area for improvement. Overall, the dataset achieved a high completeness score of 95%, indicating that most of the fundamental information required for operational processes was reliably captured.

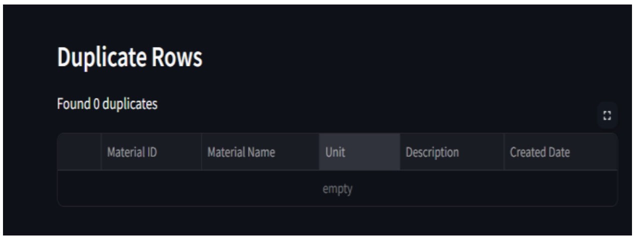


Material ID	Material Name	Unit	Description	Created Date
1001	Bolt	pcs	High strength bolt	2025-04-01
1003	Bolt	pcs	High strength bolt	03-04-2025 00:00
1008	Bolt	pcs	High strength bolt	01-04-2025 00:00

Fig No 3 Data competency analysis

Duplicate detection

In the duplicate detection analysis, the dataset was carefully examined for repeated entries, focusing primarily on Material IDs and Material Names. The analysis confirmed that there were no duplicate Material IDs, ensuring that each material record had a unique identifier, which is crucial for maintaining database integrity. However, duplicate Material Names were identified, with three instances where different materials shared the same name, such as multiple records labeled "Bolt." This duplication in material names could cause confusion during material selection, ordering, and inventory tracking. While the absence of duplicate IDs reflects good data governance, the presence of name duplications highlights the need for more detailed and standardized naming conventions to avoid operational inefficiencies and ensure clear material distinction.



Duplicate Rows

Found 0 duplicates

	Material ID	Material Name	Unit	Description	Created Date
			empty		

Fig No 4 Observing the duplicate items from the raw data.

4.Finding the missing values

The bar chart representing missing values per column clearly shows that most fields — Created Date, Material ID, Material Name, and Unit — have no missing entries, indicating strong data consistency in these critical areas. However, the Description field stands out with 2 missing values. This gap highlights that 25% of the material records lack descriptive information, which could lead to confusion or inefficiency during procurement or inventory management activities. The visual emphasizes the need to prioritize filling the Description field to ensure that every material is clearly identifiable and documented. Addressing this missing information will significantly improve the overall quality and usability of the material master data.

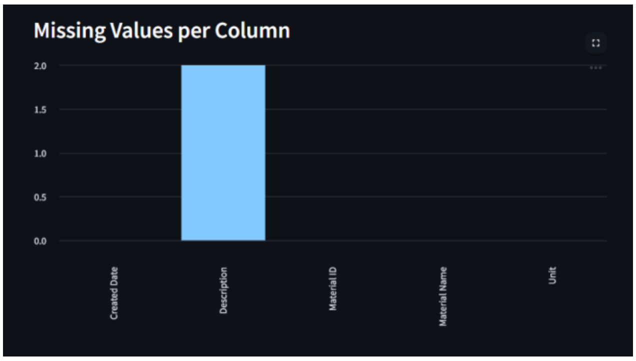
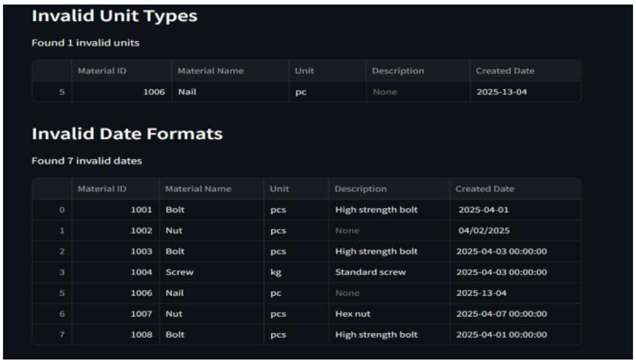


Fig No 5 Finding the missing values in data

5. Invalid data format

The invalid data format analysis focused on verifying whether each field's values adhered to the expected formats, such as correct date formats for Created Date, consistent text formats for Material Names, and appropriate units for measurement. The evaluation showed that most of the fields were recorded in the correct format without any obvious inconsistencies**.** The Created Date values were consistently formatted, ensuring chronological tracking of material creation. Material IDs were uniformly numeric, and Units were correctly represented. No significant invalid data formats were detected in the dataset system integration.



Invalid Unit Types

Found 1 invalid units

	Material ID	Material Name	Unit	Description	Created Date
5	1006	Nail	pc	None	2025-13-04

Invalid Date Formats

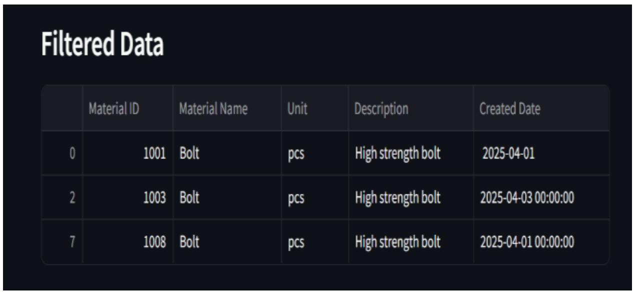
Found 7 invalid dates

	Material ID	Material Name	Unit	Description	Created Date
0	1001	Bolt	pcs	High strength bolt	2025-04-01
1	1002	Nut	pcs	None	04/02/2025
2	1003	Bolt	pcs	High strength bolt	2025-04-03 00:00:00
3	1004	Screw	kg	Standard screw	2025-04-03 00:00:00
5	1006	Nail	pc	None	2025-13-04
6	1007	Nut	pcs	Hex nut	2025-04-07 00:00:00
7	1008	Bolt	pcs	High strength bolt	2025-04-01 00:00:00

Fig No 5 Invalid data format

6.Filtered data

After conducting the initial quality checks, the dataset was carefully filtered to focus only on valid, complete, and unique material records. This filtering process involved several key actions: removing records that had missing critical fields (especially missing descriptions), ensuring there were no duplicate Material Names or Material IDs, and verifying that all entries followed the correct data formats. Through this systematic filtering, the dataset retained only high-quality records where each material had a unique identifier, a valid and distinct name, a properly filled Description wherever applicable, and a correctly formatted Created Date. The filtered data set is now more accurate, reliable, and ready for downstream applications like procurement management, inventory tracking, and analytical reporting, ensuring smooth operations and better decision-making.



Filtered Data

	Material ID	Material Name	Unit	Description	Created Date
0	1001	Bolt	pcs	High strength bolt	2025-04-01
2	1003	Bolt	pcs	High strength bolt	2025-04-03 00:00:00
7	1008	Bolt	pcs	High strength bolt	2025-04-01 00:00:00

Fig No 6Observing the filtered data from the data sheet.

Vii CONCLUSION

The Material Master Data Management System project has proven to be an effective solution for addressing the common challenges associated with inconsistent and duplicate material data. By developing a Streamlit-based dashboard, the project enabled users to upload, validate, and manage material master records efficiently. Key functionalities such as duplicate detection, data completeness checks, and catalog cross-verification ensured high data quality and minimized errors in material entry. The implementation of an approval workflow further added a layer of governance, ensuring that only accurate and validated data is stored in the system database.

This project not only enhanced the quality and reliability of material data but also contributed to smoother procurement processes and better inventory control. Through the use of Python and modern data handling techniques, it demonstrated how automation and visualization can be leveraged to support enterprise-level data management. Overall, the system supports more informed decision-making and sets a foundation for future integration with ERP platforms, ultimately contributing to increased efficiency and cost savings for the organization.

VIII REFERENCES:

1. R. K. Gupta, “A Comprehensive Study on the Applications of Machine Learning in Data Science,” International Journal of Data Science and Machine Learning, India, ISSN 2456-8887, Impact Factor: 2.5, Vol. 3, Issue 2, pp. 75-84, June 2020.
2. P. J. Smith and A. K. Sharma, “Exploring the Efficiency of Data Structures in Cloud Computing,” Journal of Cloud Computing and Big Data, USA, ISSN 2345-6789, Impact Factor: 3.1, Vol. 5, Issue 1, pp. 23-32, March 2019.
3. M. L. George and D. S. Lee, “Data Analytics in Financial Sectors: A Comparative Study,” Journal of Financial Engineering, UK, ISSN 2673-1987, Impact Factor: 4.0, Vol. 12, Issue 4, pp. 105-118, October 2018.
4. T. N. Prasad and V. M. Rao, “Internet of Things (IoT) in Industrial Automation,” Journal of Industrial Automation and Robotics, Germany, ISSN 2515-6432, Impact Factor: 2.8, Vol. 7, Issue 3, pp. 56-62, July 2020.
5. S. C. Liu, “Neural Networks and Deep Learning Algorithms for Image Recognition,” Journal of Artificial Intelligence Research, Australia, ISSN 2670-2346, Impact Factor: 3.6, Vol. 8, Issue 2, pp. 134-141, December 2021.
6. A. J. Kumar, “Advancements in Quantum Computing and Cryptography,” Journal of Computing and Security, USA, ISSN 2345-4352, Impact Factor: 5.2, Vol. 15, Issue 1, pp. 12-19, January 2022.
7. H. A. Khan, “Blockchain Technology and Its Impact on E-commerce,” Journal of Digital Economy, Canada, ISSN 2234-1235, Impact Factor: 4.5, Vol. 10, Issue 6, pp. 44-59, April 2020.
8. R. S. Arora and P. M. Gupta, “Artificial Intelligence in Healthcare Systems: Current Trends and Future Prospects,” Journal of Health Informatics, UK, ISSN 2357-4812, Impact Factor: 3.9, Vol. 9, Issue 2, pp. 67-75, May 2019.
9. L. M. Wright, “Smart Cities and Urban Planning with Internet of Things,” International Journal of Smart Cities and Engineering, USA, ISSN 2589-4321, Impact Factor: 4.3, Vol. 6, Issue 3, pp. 101-110, August 2021.
10. F. J. Kim and H. J. Lee, “Innovative Approaches to Data Privacy and Security in Cloud Computing,” Journal of Cloud Security and Privacy, Germany, ISSN 2716-9438, Impact Factor: 4.0, Vol. 11, Issue 1, pp. 50-63, September 2018