SMART INDIA HACKATHON 2024

TITLE PAGE

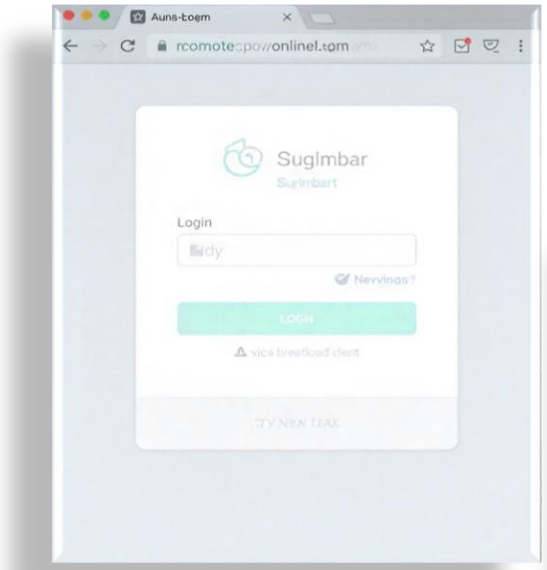
- **Problem Statement ID – 1659**
- **Problem Statement Title- Data download Duplication Alert System (DDAS)**
- **Theme- Miscellaneous**
- **PS Category- Software**
- **Team ID- 46183**
- **Team Name- D Tech Eternals**

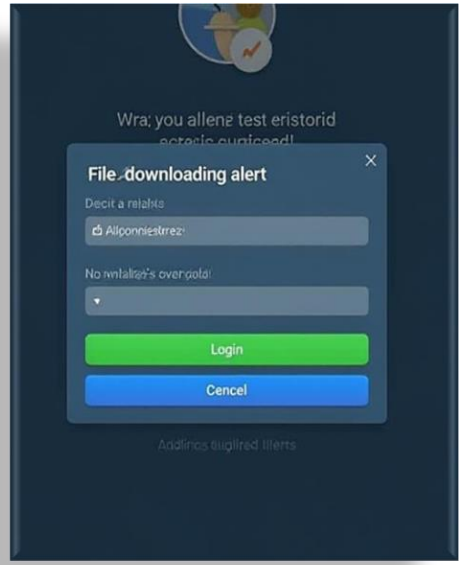
D Tech Eternals

Data download Duplication Alert System (DDAS)

SMART INDIA HACKATHON 2024

- **Overview :**
  - A DDAS addresses the issue of multiple users inadvertently downloading duplicate copies of the same datasets across various fields. The DDAS operates by maintaining a repository or database that records metadata of all downloaded datasets.
  - When download request occurs, we will compare the file's information with existing entries and if duplicate is found the alert message will be displayed.





- **Addressing the problem :**
  - Prevent multiple users to download same data.
  - Saves lots of bandwidth.
  - Saves storage resources.

- **Innovation in the project :**
  - Will use unique identifiers (hash values) to detect duplicates.
  - By this system will be more user-friendly and will prevents delays.

D Tech Eternals

TECHNICAL APPROACH

SMART INDIA HACKATHON 2024

• Technologies: -

Frontend: React.js or Angular.js for building the user interface.

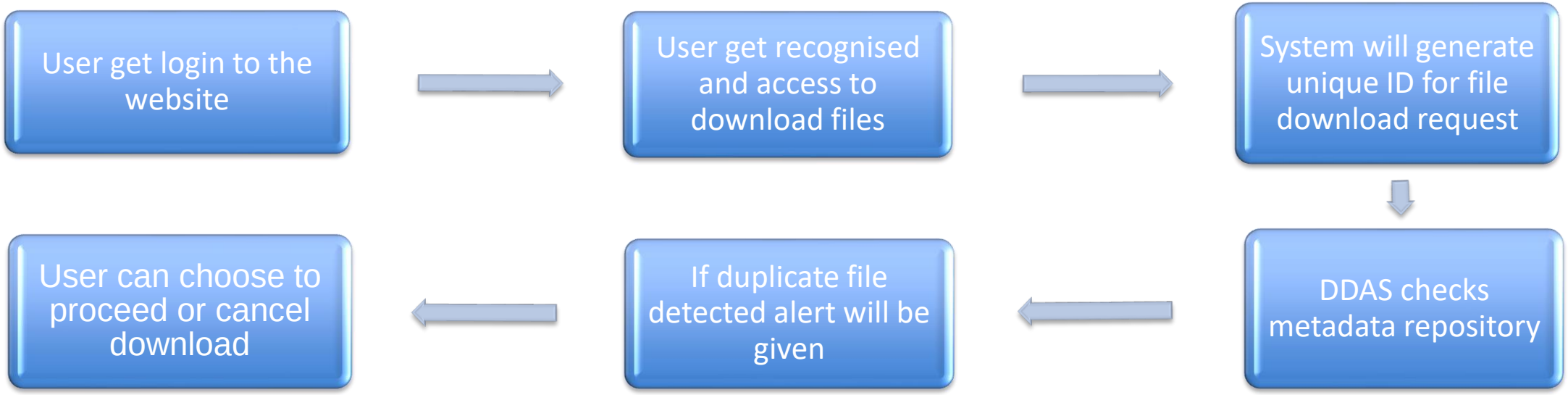
Backend: Node.js with Express.js for handling API request and managing metadata.

Database: MongoDB/PostgreSQL for storing metadata, Elasticsearch for search functionality.

• Duplicate detection: Uses MD5/SHA-256 hashing algorithms.

Storage integration: Works with AWS S3, Google Cloud, etc.

- **Process for implementation:**



D Tech Eternals

FEASIBILITY AND VIABILITY

SMART INDIA HACKATHON 2024

- **Feasibility:**
  - Assess the existing infrastructure to determine if it can support the system.
  - Evaluate the volume of data downloads to ensure the system can handle peak loads without performance issues.
- **Potential challenges:**
  - Potential delays in detecting duplicates due to large dataset volumes.
  - Handling variations in file storage structures.
- **Strategies for overcoming these challenges:**
  - Use hashing techniques to quickly identify duplicate data.
  - Maintain a log of downloads with timestamps and user identifiers to track patterns effectively.
  - Create a user-friendly alert system that notifies users of potential duplication, providing options to confirm or reject the alert.

D Tech Eternals

IMPACT AND BENEFITS

SMART INDIA HACKATHON 2024

- **Impact:-**

1. Reduces unnecessary bandwidth and storage consumption.
2. Streamline data management processes.

- **Benefits:-**

1. Saves time by avoiding redundant downloads.
2. Minimize storage cost.
3. Improve overall organisational efficiency and collaboration.

D Tech Eternals

RESEARCH AND REFERENCES

SMART INDIA HACKATHON 2024

1.Data Deduplication Techniques:

1. H. Wang, H. Huang, and J. Liu, "A Survey on Data Deduplication Techniques," *Journal of Computer Science and Technology*, 2020.
2. S. B. H. Chowdhury et al., "Efficient Data Deduplication in Cloud Storage," *IEEE Transactions on Cloud Computing*, 2021.

2.Duplicate Detection Algorithms:

1. A. K. Jain, "An Overview of Duplicate Detection Techniques," *IEEE Transactions on Knowledge and Data Engineering*, 2019.
2. J. A. R. DeHaan et al., "Duplicate Detection in Large Data Sets," *Data Mining and Knowledge Discovery*, 2021.

3.Data Quality and Governance:

1. "Data Quality: The Accuracy Dimension" by Jack E. Olson. This book covers principles of data quality, including deduplication.