SMART INDIA HACKATHON 2024

SMART INDIA HACKATHON 2024

- **Problem Statement ID – SIH1659**
- **Problem Statement Title- Data download Duplication Alert System (DDAS)**
- **Theme- MISCELLANEOUS**
- **PS Category- Software**
- **Team ID- 1811**
- **Team Name- SVVV_TECHNOBYTE**

SVVV_TECHNO BYTE

AI-Powered Data Download Duplication Alert System (DDAS)

SMART INDIA HACKATHON 2024

Proposed Solution:

- **AI-Powered Duplication Detection:** AI models like Multimodal Deep Learning Model (BERT, CNN, Simhash, FAISS) will intelligently identify duplicate files based on content similarity rather than simple hashing.
- **Real-Time Alerts:** Users will receive instant notifications of potential duplicates, saving time and resources.
- **Centralized Data Management:** The system provides secure, role-based access to existing files, reducing unnecessary downloads.

How It Solves the Problem:

- **Reduces Bandwidth:** AI optimizes data downloads, avoiding file duplication across users.
- **Optimizes Storage:** Only one copy of a dataset is saved, preventing redundant storage.

Innovation:

- **AI Content Matching**: Unlike traditional hash comparisons, AI can detect **partial duplicates** and **similar files**, even if names differ (Rodrigues and Pinto, 2023).
- **Machine Learning Predictions**: The system learns from user behavior, proactively suggesting files they may need before downloading.
- Incorporates Role-Based Access Control (RBAC) and Audit Logs to protect privacy (Ghazal et al., 2020).
- Allows users direct access to existing files via a secure LAN-based system.

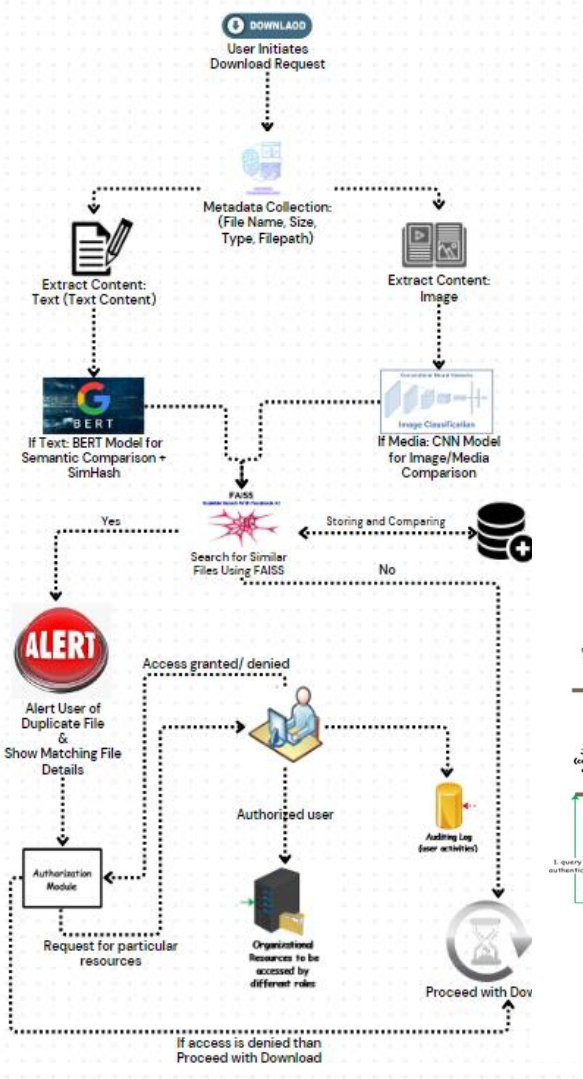
SVVV_TECHNO BYTE

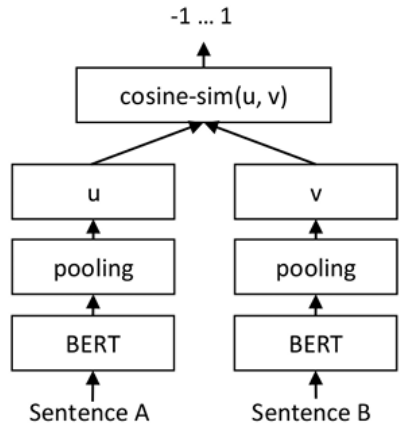
TECHNICAL APPROACH

SMART INDIA HACKATHON 2024

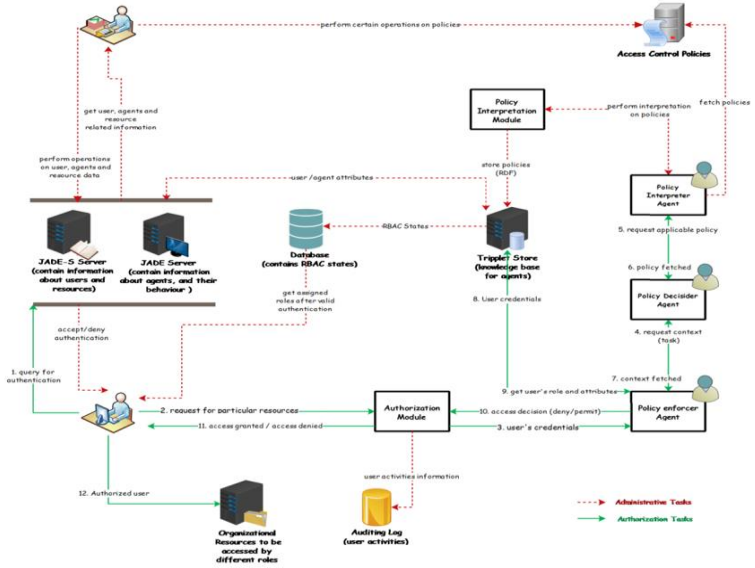
Technologies:

- **Programming Language**: Java
- **AI Models:**
  - **BERT** helps find text duplicates by analyzing the meaning of content.
  - **CNN** analyzes images for similarity.
  - **SimHash** quickly detects near-duplicate files.
  - **FAISS** is used for fast and efficient similarity search across large datasets, enabling quick retrieval of similar files.
- **AI/ML Frameworks**:
  - **Deeplearning4j** for deep learning (Kaluza, 2016).
  - **Weka** for traditional machine learning tasks (Aher and Lobo, 2011).
- **Database**: MySQL (stores metadata and AI-generated insights).
- **Real-Time Alerts**: WebSocket (for instant duplicate notifications).
- **Access Control**: Role-Based Access Control (RBAC) ensures file security (Ghazal et al., 2020).
- **Audit Logging**: Tracks user actions for transparency and security.





Source: Bonetti and Torroni, 2021



Source: Ghazal et al., 2020

SVVV_TECHNO BYTE

FEASIBILITY AND VIABILITY

SMART INDIA HACKATHON 2024

Feasibility:

- The system is designed to be easy to implement within existing organizational structures. Users won't need to change their behavior; they simply download files as usual, and the system works in the background to alert them.
- The AI models continuously learn and improve over time, allowing for adaptive detection of similar files based on usage patterns, making it increasingly accurate and beneficial over time.
- Role-Based Access Control (RBAC) ensures that only authorized users can access files, maintaining security standards without adding complexity to user workflows. (Javed et al., 2022).

Challenges:

- **AI Training**: Ensuring the models are trained on diverse file types for accurate duplication detection.
- **Real-Time Response**: Managing network latency for instant alerts in large-scale deployments.

Solutions:

- **AI Model Updates**: Regular updates and training to improve accuracy and file recognition.
- **Optimized WebSocket**: Configuring WebSocket for faster response times and parallel processing to handle high load.

IMPACT AND BENEFITS

Impact:

- **Optimized Resources:** Reduced bandwidth and storage usage, lowering the system's overall resource consumption.
- **Increased Efficiency:** Users save time by avoiding unnecessary downloads, improving productivity across teams.
- **Better Data Management:** Streamlined access to existing files, reducing confusion over duplicate file versions and enhancing collaboration..

Benefits:

- **Social**: DDAS improves collaboration by preventing confusion over multiple file versions, making it easier for teams to share and access accurate data.
- **Economic**: By reducing **bandwidth usage by up to 40%** and saving **25% of storage**, institutions can cut down on operational costs related to infrastructure and data management.
- **Environmental**: Efficient storage usage reduces the need for additional data centers, lowering energy consumption and contributing to a more sustainable IT environment.

RESEARCH AND REFERENCES

- Ghazal, R., Malik, A.K., Qadeer, N., Raza, B., Shahid, A.R. and Alquhayz, H., 2020. Intelligent role-based access control model and framework using semantic business roles in multi-domain environments. *IEEE Access*, *8*, pp.12253-12267.
- Javed, A.R., Ahmed, W., Alazab, M., Jalil, Z., Kifayat, K. and Gadekallu, T.R., 2022. A comprehensive survey on computer forensics: State-of-the-art, tools, techniques, challenges, and future directions. *IEEE Access*, *10*, pp.11065-11089.
- Kaluža, B., 2016. Machine Learning in Java. *UK: Packt Publishing Ltd*.
- Aher, S.B. and Lobo, L.M.R.J., 2011, March. Data mining in educational system using weka. In *International conference on emerging technology trends (ICETT)* (Vol. 3, pp. 20-25). Foundation of Computer Science.
- Rodrigues, M. and Pinto, F., 2023, July. Content Matching for City Improvement. In *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)* (pp. 1-6). IEEE.
- Bonetti, L. and Torroni, P., *Design and Implementation of a Realworld Search Engine Based on Okapi Bm25 and Sentencebert* (Doctoral dissertation, MS Thesis, Department of Computer Science and Engineering, Alma Mater Studiorum-Università di Bologna, Bologna, Italy, 2021, www. amslaurea. unibo. it/24774/1/Thesis_Bonetti. pdf (Accessed January 20, 2024)).