Data Duplication and Anomaly Alert System (for company database)

Introduction

The **Data Duplication and Anomaly Alert System** is an advanced security framework designed to prevent unauthorized or excessive data downloads. It employs **cryptographic hashing, text similarity algorithms, and anomaly detection techniques** to ensure data integrity, prevent misuse, and enhance security.

Duplication Detection Mechanism

SHA-256 Hashing for File Integrity

To detect duplicate downloads, the system employs **SHA-256 hashing**, a cryptographic function that generates a unique fingerprint for each downloaded file. Even the slightest modification in content results in a completely different hash, making it an effective method for duplicate detection.

How It Works:

When a file is requested for download, the system generates its SHA-256 hash.

This hash is compared against previously downloaded files stored in the database.

If a match is found, the system flags it as a **duplicate download attempt** and can trigger an alert or block the action.

Text-Based Similarity Detection using TF-IDF & LSA

For text-based downloads such as documents and reports, hashing alone may not be effective, as minor changes in formatting or content can alter the hash. To handle this, the system integrates **TF-IDF (Term Frequency-Inverse Document Frequency)** and **LSA (Latent Semantic Analysis)** to detect content duplication.

TF-IDF Analysis:

Assigns importance weights to words based on their frequency in a document.

Helps measure similarity between text documents with high accuracy.

If similarity exceeds a predefined threshold (e.g., 85%), the document is flagged as a potential duplicate.

Latent Semantic Analysis (LSA):

Uncovers hidden patterns in textual data to detect conceptual similarities.

Enhances duplication detection by understanding context beyond simple word matches.

Useful for detecting paraphrased content that might bypass traditional similarity checks.

Anomaly Detection Mechanism

To prevent **unusual and unauthorized data access patterns**, the system incorporates advanced anomaly detection techniques. It continuously monitors user activity, flags suspicious behavior, and takes action when necessary.

User Activity Logging & Threshold Monitoring

The system tracks:

Download frequency: Number of files downloaded within a given time window.

Time of access: Identifies downloads occurring at unusual hours.

IP address monitoring: Detects suspicious location changes or multiple accounts using the same IP.

If a user exceeds a predefined threshold (e.g., downloading 10+ files within 5 minutes), an **immediate alert is triggered**, and the user may be temporarily restricted.

Machine Learning for Anomaly Detection

To improve the detection of abnormal download patterns, the system integrates **machine learning models** such as **Isolation Forest**:

How Isolation Forest Works:

Learns normal download behaviors based on historical data.

Flags unusual behavior as anomalies based on deviation from expected patterns.

Helps identify users attempting bulk downloads or automated data extraction.

Automated Actions Upon Detection:

Immediate notification to administrators.

Temporary suspension of suspicious accounts.

Enforcement of additional authentication steps for flagged users.

Conclusion

The **Data Download Duplication Alert System** provides a **comprehensive and efficient solution** for detecting and preventing unauthorized data downloads. By leveraging **SHA-256 hashing, TF-IDF, LSA, and Isolation Forest algorithms**, it ensures that duplicate downloads and anomalous activities are promptly detected and mitigated. This multi-layered security approach enhances data integrity, prevents misuse, and safeguards sensitive information against potential threats.