A Privacy Leakage Upper-Bound Constraint Based Approach for Cost-Effective Privacy Preserving of Intermediate Datasets In cloud

**ABSTRACT**

Cloud computing provides massive computation power and storage capacity which enable users to deploy computation and data intensive applications without infrastructure investment. Along the processing of such applications, a large volume of intermediate datasets will be generated, and often stored to save the cost of re-computing them. However, preserving the privacy of intermediate datasets becomes a challenging problem because adversaries may recover privacy-sensitive information by analyzing multiple intermediate datasets. Encrypting ALL datasets in cloud is widely adopted in existing approaches to address this challenge. But we argue that encrypting all intermediate datasets are neither efficient nor cost-effective because it is very time consuming and costly for data-intensive applications to en/decrypt datasets frequently while performing any operation on them. In this paper, we propose a novel upper-bound privacy leakage constraint based approach to identify which intermediate datasets need to be encrypted and which do not, so that privacy-preserving cost can be saved while the privacy requirements of data holders can still be satisfied. Evaluation results demonstrate that the privacy-preserving cost of intermediate datasets can be significantly reduced with our approach over existing ones where all datasets are encrypted.

**EXISTING SYSTEM:**

Existing technical approaches for preserving the privacy of datasets stored in cloud mainly include encryption and anonymization. On one hand, encrypting all datasets, a straightforward and effective approach, is widely adopted in current research. However, processing on encrypted datasets efficiently is quite a challenging task, because most existing applications only run on unencrypted datasets. However, preserving the privacy of intermediate datasets becomes a challenging problem because adversaries may recover privacy-sensitive information by analyzing multiple intermediate datasets. Encrypting ALL datasets in cloud is widely adopted in existing approaches to address this challenge. But we argue that encrypting all intermediate datasets are neither efficient nor cost-effective because it is very time consuming and costly for data-intensive applications to en/decrypt datasets frequently while performing any operation on them.

**PROPOSED SYSTEM:**

In this paper, we propose a novel approach to identify which intermediate datasets need to be encrypted while others do not, in order to satisfy privacy requirements given by data holders. A tree structure is modeled from generation relationships of intermediate datasets to analyze privacy propagation of datasets. As quantifying joint privacy leakage of multiple datasets efficiently is challenging, we exploit an upper-bound constraint to confine privacy disclosure. Based on such a constraint, we model the problem of saving privacy-preserving cost as a con-strained optimization problem. This problem is then divided into a series of sub-problems by decomposing privacy leakage constraints. Finally, we design a practical heuristic algorithm accordingly to identify the datasets that need to be encrypted. Experimental results on real-world and extensive datasets demonstrate that privacy-preserving cost of intermediate datasets can be significantly reduced with our approach over existing ones where all datasets are encrypted.

**MODULE DESCRIPTION:**

# **Number of Modules**

After careful analysis the system has been identified to have the following modules:

1. **Data Storage Privacy Module.**
2. **Privacy Preserving Module.**
3. **Intermediate Dataset Module.**
4. **Privacy UpperBound Module.**

**1.Data Storage Privacy Module:**

The privacy concerns caused by retaining intermediate datasets in cloud are important but they are paid little attention. A motivating scenario is illustrated where an on-line health service provider, e.g., Microsoft Health Vault has moved data storage into cloud for economical benefits. Original datasets are encrypted for confidentiali-ty. Data users like governments or research centres access or process part of original datasets after anonymization. Intermediate datasets generated during data access or process are retained for data reuse and cost saving. We proposed an approach that combines encryption and data fragmen-tation to achieve privacy protection for distributed data storage with encrypting only part of datasets.

**2. Privacy Preserving Module:**

Privacy-preserving techniques like generalization can with-stand most privacy attacks on one single dataset, while preserving privacy for multiple datasets is still a challeng- ing problem. Thus, for preserving privacy of multiple datasets, it is promising to anonymize all datasets first and then encrypt them before storing or sharing them in cloud. Privacy-preserving cost of intermediate datasets stems from frequent en/decryption with charged cloud services.

**3. Intermediate Dataset Module:**

An intermediate dataset is assumed to have been ano-nymized to satisfy certain privacy requirements. However, putting multiple datasets together may still invoke a high risk of revealing privacy-sensitive information, resulting in violating the privacy requirements. Data provenance is employed to manage intermediate datasets in our research. Provenance is com-monly defined as the origin, source or history of deriva-tion of some objects and data, which can be reckoned as the information upon how data was generated. Re-producibility of data provenance can help to regenerate a dataset from its nearest existing predecessor datasets rather than from scratch

**4. Privacy UpperBound Module:**

Privacy quantification of a single data-set is stated. We point out the challenge of privacy quantification of multiple datasets and then derive a privacy leakage upper-bound con-straint correspondingly. We propose an upper-bound constraint based approach to select the necessary subset of intermediate datasets that needs to be encrypted for minimizing privacy-preserving cost. The privacy leakage upper-bound constraint is decomposed layer by layer.

**Algorithm used heuristic algorithm:**

**Algorithm 1. Privacy-Preserving Cost Reducing Heuristic**

we design a heuristic algorithm to reduce privacy-preserving cost. In the state-search space for an SIT, a state node SNi in the layer Li herein refers to a vector of partial local solutions, i.e., SNi corresponds to h\_1j1 ; . . . ; \_iji I where \_kjk 2 \_k, 1 \_ k \_ i. Note that the state-search tree generated according to an SIT is different from the SIT itself, but the height is the same. Appropriate heuristic information is quite vital to guide the search path to the goal state. The goal state in our algorithm is to find a near-optimal solution in a limited search space.

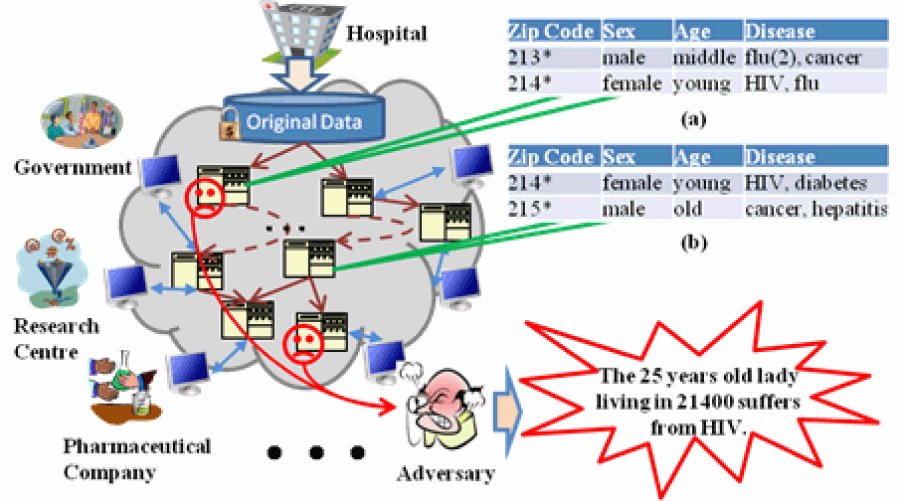
****

****

**Expected output:**

Evaluation results need to demonstrate that the privacy-preserving cost of intermediate data sets can be significantly reduced with our approach over existing ones where all data sets are encrypted.

**PROCESS FLOW**:



**SOFTWARE REQUIREMENTS**:

Operating System : Windows

Technology : Java and J2EE

Web Technologies : Html, JavaScript, CSS

IDE : Eclipse

Web Server : Tomcat

Database : My SQL

Java Version : J2SDK1.5

**HARDWARE REQUIREMENTS**:

Hardware : Pentium

Speed : 1.1 GHz

RAM : 1GB

Hard Disk : 20 GB

Floppy Drive : 1.44 MB

Key Board : Standard Windows Keyboard

Mouse : Two or Three Button Mouse

Monitor : SVGA