Laboratoire ERI C

Varunya ATTASENA
Nouria HARBI
Jérôme DARMONT

Sharing-based Privacy and Availability of Cloud Data Warehouses
Business intelligence (BI) and data analytics have been an ever-growing trend in

- **Business**
  - Finance
  - Telecoms
  - Insurance
  - Logistics

- **Non-business**
  - Agriculture
  - Medicine
  - Health
  - Environment

**DATA WAREHOUSE**
Introduction

Characteristic of Cloud Computing

- On-Demand Self-Service
- Broad Network Service
- Resource pooling
- Rapid Elasticity
- Measured Service
Introduction

Move DW to Cloud Computing

DATA WAREHOUSE

CLOUD COMPUTING

High security

High performance of data analysis

High costs of implements, maintenances...

EDA 2013- Sharing-based Privacy and Availability of Cloud Data Warehouses
Introduction

CLOUD DATA WAREHOUSE

- Low security
- Performance of data analysis?
- Low costs of implements, maintenances...
Introduction
- Data warehouse
- Cloud computing

Problems
- Cloud data warehouse

Scheme I
- A new \((m, n, t)\) multi secret sharing

Scheme II
- Sharing a data warehouse in the cloud

Security analysis & performance evaluation

Conclusions
- Conclusions
- Future researches
Problems: Cloud security issues
Problems

1. Data Security
   - Data Privacy
   - Data Availability
   - Data Integrity

2. Data Analysis
   - Data Analysis

EDA 2013 - Sharing-based Privacy and Availability of Cloud Data Warehouses
Related work

- Data Availability
- Data Integrity
- Data Privacy
- Data analysis

Data replication
- Synchronous
- Asynchronous
- Semi-synchronous

Data anonymization
- k-anonymization
- l-diversity

Data verification
- Inner code verifying
- Outer code verifying

Data security
- (By encryption)

Homomorphic encryption (HE)
- A partially HE
- A fully HE

Incrystal encryption

Secret sharing (SS)

Multi-secret sharing (MSS)

Verification secret sharing (VSS)
**Scheme-I:** A new \((m, n, t)\) multi secret sharing scheme

- **Data Availability**
- **Data Integrity**
- **Data Privacy**
- **Data analysis**

### Related Works
- **Data replication**
  - Synchronous
  - Asynchronous
  - Semi-synchronous
- **Data anonymization**
- **Data security** (By encryption)
  - Homomorphic encryption (HE)
    - A partially HE
    - A fully HE
  - Incremental encryption
  - Secret sharing (SS)
  - Multi-secret sharing (MSS)
  - Verification secret sharing (VSS)
- **Data Verification**
  - Inner code verifying
  - Outer code verifying

- **k-anonymization**
- **l-diversity**

---

EDA 2013 - Sharing-based Privacy and Availability of Cloud Data Warehouses
**Scheme-I: A new \((m, n, t)\) multi secret sharing scheme**

**Principle:** share data over several cloud service providers
(Each provider will only store part of the data which will also not be exploitable neither by the provider nor any intruder.)

\(m = \) a number of Data.
\(n = \) a number of cloud service providers (CSPs).
\(t = \) a sufficient number of CSPs for reconstructing data.

---

**Data Privacy**  **Data Availability**  **Data Integrity**  **Data analysis**

EDA 2013- Sharing-based Privacy and Availability of Cloud Data Warehouses
How to encrypt data?

1. Data are organized into blocks. Each block is encrypted and decrypted all at once.
2. All data in the block are encrypted by mapping them and their signature to coefficients of a polynomial equation of degree $t - 1$.

How to verify the correctness of data?

1. Create a signature $s_{\text{in}}$ by a hash function (verify the trusted of providers)
2. Create a signature $s_{\text{out}}$ by a hash function (reduce a cost of data transfer in reconstructing process because no error encrypted data is transferred)
**Sharing Process**

1. Data are organized into blocks.
2. Create a signature in each block.
3. Encrypt data and a signature in each block by Polynomial equation.
4. Create a signature of each encrypted data.

**Scheme-I: A new \((m, n, t)\) multi secret sharing scheme**

1. **Store in provider 1**
   - \(d_1\)
   - \(d_2\)
   - \(\ldots\)
   - \(d_{t-1}\)
   - \(S_{\text{in}_1}\)

2. **Store in provider 2**
   - \(d_t\)
   - \(d_{t+1}\)
   - \(\ldots\)
   - \(d_{2t-2}\)
   - \(S_{\text{in}_2}\)

3. **Store in provider 3**
   - \(d_{2t-1}\)
   - \(\ldots\)
   - \(d_{m}\)
   - \(\ldots\)
   - \(S_{\text{in}_n}\)

4. Encrypt data and signatures:
   - \(e_{1,1}\) \(e_{1,2}\) \(\ldots\) \(e_{1,n}\)
   - \(e_{2,1}\) \(e_{2,2}\) \(\ldots\) \(e_{2,n}\)
   - \(\ldots\)
   - \(e_{o,1}\) \(e_{o,2}\) \(\ldots\) \(e_{o,n}\)

The diagrams illustrate the process of sharing data and creating signatures in different providers, ensuring privacy and availability of cloud data warehouses.
**Reconstructing Process**

1. Select $t$ CSPs from $n$ CSPs
2. Verify a correctness of encrypted data in each CSP.
3. Transfer encrypted data to user.
4. Compute original data and a signature.
5. Verify the correctness of data.

---

**Scheme-I: A new $(m, n, t)$ multi secret sharing scheme**

- **Step 1:** Select $t$ from $n$ shares.
- **Step 2:** Store encrypted data in different providers.
- **Step 3:** Reconstruct the process.
- **Step 4:** Decrypt the data.
- **Step 5:** Verify the correctness of data.

EDA 2013 - Sharing-based Privacy and Availability of Cloud Data Warehouses
### Scheme-II: Sharing a data warehouse in the cloud

<table>
<thead>
<tr>
<th>Original data</th>
<th>Encrypted data at CSP(_1)</th>
<th>Encrypted data at CSP(_2)</th>
<th>Encrypted data at CSP(_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>id</strong></td>
<td><strong>name</strong></td>
<td><strong>salary</strong></td>
<td><strong>sex</strong></td>
</tr>
<tr>
<td>124</td>
<td>Bob</td>
<td>75€</td>
<td>M</td>
</tr>
<tr>
<td>125</td>
<td>Anna</td>
<td>80€</td>
<td>F</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Encrypted data at CSP(_2)</strong></th>
<th><strong>Encrypted data at CSP(_3)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>id</strong></td>
<td><strong>name</strong></td>
</tr>
<tr>
<td>124</td>
<td>(6,6),(6,6),(2,2)</td>
</tr>
<tr>
<td>125</td>
<td>(2,2),(5,5),(5,5),(0,0)</td>
</tr>
</tbody>
</table>
Data Analysis over shares
Can analyze data (search and aggregation operations) over shares while not decrypting all data first.

<table>
<thead>
<tr>
<th>Original data</th>
<th>Encrypted data at CSP_1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>id</td>
</tr>
<tr>
<td>id</td>
<td>124</td>
</tr>
<tr>
<td></td>
<td>125</td>
</tr>
</tbody>
</table>

Select name from customer where sex='M'.
At CSP_1: Select name from customer where sex='9'.

Select avg(salary) from customer.
At CSP_1: Select avg(salary) from customer.
Security analysis and performance evaluation

Security analysis

Reliability analysis

Cost analysis
- Time complexity
- Stored data volume

Reliability analysis

Security analysis
Security analysis

- Neither the CSP nor any intruder can decode the original data from only one share.
- It is very difficult to retrieve shares from all CSPs’ by attacking them simultaneously.
- In the case that an intruder can steal shares from $x$ CSPs such that $x \leq t$, the probability of discovering $b_j$ is very low.

**Scheme I:** \( \frac{1}{p^{2t-x-1}} \)

**Scheme II:** \( \frac{1}{p^{2t-x-1}} \)

$x$ is a number of pilfered shares.

- Probability of discovering an original data block from some or all shares
Reliability analysis

- **Data availability**: Our schemes guarantee the user can reconstruct $D$ if $t$ or more CSPs are honest and their shares are accessible.
- **Data integrity**: Our schemes can verify both the honesty of CSPs and the correctness of CSPs’ shares.
- **Data recovery**: If some shares are erroneous (lost, damaged, alternative...), they are reconstructed from $t$ other shares.

![Probability of incorrect data not being detected (false negative)](image)
Cost analysis: Time complexity

The time complexity in both schemes

- The time complexity of the data sharing process is $O(otn)$
- The time complexity of the data reconstruction process is $O(ot^2)$

The execution time of Scheme-II: in the data reconstruction process, the execution time is about 3:04 seconds, and throughput is 336 MB per second when $n = 4$ and $t = 3$.

Probability of incorrect data not being detected (false negative)
Cost analysis: Stored data volume

The Stored data volume
- The Stored data volume in Scheme I is indeed lower than $on || P ||$
- The Stored data volume in Scheme II is indeed lower than $on || p ||$

For example, with Scheme-II: 32 bits unsigned integers
(It are shared among 6 CSPs and 5 CSPs are sufficient for reconstruct them. Let $||p|| = 9$ bits.)
- The volume of all shares is lower than $1 \times 6 \times 9 = 54$ bits.
- The volume of each share is lower than $1 \times 9 = 9$ bits.

By implementation of Scheme-II:
- The volume of all shares is greater than the volume of D but less than $D \times 2$.
- The volume of each share is lower than the volume of D.

Volume of shares with Scheme-II
Conclusion

Our schemes

- Data Availability
- Data Integrity
- Data Privacy
- Data Analysis

\[ \text{COSTS} \]
- Time complexity
- Stored data volume
- Transferred data volume

\[ = \]

- Scheme-I
  - A new \((m,n,t)\) multi secret sharing scheme
- Scheme-II
  - Sharing a data warehouse in the cloud

Future researches

- Data Availability
- Data Integrity
- Data Privacy
- Data Analysis

\[ \text{Advanced Data Analysis} \]
- Sophisticated aggregations
- Complex predicates
  (conjunction, negation, disjunction...)

\[ + \]

- COST
  - Losing data volume of all shares
Thank you