Secure Data Sharing and Distribution Platform for Integrated Big Data Utilization

- Handling all data with encryption -

Group Members

Waseda University  Hayato YAMANA
Institute of Information Security  Atsuhiro GOTO
Ochanomizu University  Masato OGUCHI
Kogakuin University  Saneyasu YAMAGUCHI
The University of Electro-Communications  Takahiko SHINTANI
Meiji Pharmaceutical University  Tamotsu NOGUCHI

Oct. 2015 - Mar. 2021 funded by Japan Science and Technology Agency
Brief Introduction of our Project

1. Research Background
2. Objective
3. Research Goal
4. Research Strategy
5. Experiment
6. Schedule
7. Progress in 2015FY
1. Research Background

At least 40% of it requires some level of security, from privacy protection to full-encryption ‘lockdown.’ …

Also unfortunately, the amount needing protection will grow …


---

**Examples:**
- Corporate financial data
- Personally identifiable information (PII)
- Medical records
- User account information

**Portion Protected:** 48%

**Portion Not Protected:** 52%

---

**Digital Universe**

- **Portion of DU Not Needing Protection:** 57%
- **Portion of DU Needing Protection:** 43%

---

**Information Security:** Much of the Data that Needs to Be Protected Is Not Yet Protected

---

*e.g. How should we manage private genome data?*
1. Research Background

- Anonymization
  - Attribute Linkage Model
  - k-anonymity, l-diversity, t-closeness
  - Probabilistic Model
  - differential privacy

Limitation of Anonymization

Link Attack
William was governor of Massachusetts and his medical records were in the GIC data. Governor Weld lived in Cambridge. According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code.
2. OBJECTIVE

<table>
<thead>
<tr>
<th>OUR APPROACH IS NOT ANONYMIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>OUR APPROACH IS HANDLING ALL DATA WITH ENCRYPTION 🗝️ THROUGHOUT DATA LIFE CYCLE</td>
</tr>
<tr>
<td>YOU CAN ADOPT ANONYMIZATION, BESIDES.</td>
</tr>
</tbody>
</table>
3. Research Goal

HANDLING ALL DATA WITH ENCRYPTION THROUGHOUT DATA LIFE CYCLE

Protecting Data from Leaking
3. Research Goal

HANDLING ALL DATA WITH ENCRYPTION THROUGHOUT DATA LIFE CYCLE

1. Confidentiality guarantee for kinds of contents
   - Using Fully Homomorphic Encryption (FHE)

2. Assurance of content source and provenance
   - Using Proof of storage

3. Flexible and assured access control
   - Using Attribute-based encryption
1. Confidentiality Guarantee for kinds of Contents

① Create PK & SK

Key generation (FHE/ideal lattices): 2.5 s/key
Key size: 17MB ~ 2GB
Encryption: 0.2 s/data
(7,500 hour / GB)

Not fit to CPU's cache memory

10^3 ~ 10^10 slower than normal operations

② Encrypt

Raw data

PK: Public Key
SK: Secret Key

③ Execute

No usable library to execute data mining and machine learning.
~10^10 slower than w/o encryption

④ Bootstrap

No noise part invades indispensable part

Bootstrapping: 6 s/data
(2,200 hour / MB)

30 times slower than encryption
2. Assurance of content source and provenance

Using proof of storage

- Signature size: twice the original data (2TB for 1TB original data)
- Requires a large storage space
- Re-tag creation is required depending on a kind of calculation
3. Flexible and assured access control

Using Attribute-based encryption, $10^2 \sim 10^3$ speedup is indispensable.
- handling “numeric number” as it is, not as character.
3. Research Goal

**BASELINE**
- CURRENT FHE,
- PROOF OF STORAGE,
- ATTRIBUTE-BASED ENCRYPTION

**GOAL**

1,000 TIMES FASTER THAN CURRENT ENCRYPTION METHODS

TO SHOW THE EFFECTIVENESS OF OUR PLATFORM WITH EXPERIMENTAL DEMONSTRATION
4. Research Strategy

- **Parallelization**
  1. For FHE, adopt “Ideal Lattice” whose basic operation is “matrix calculations,” to parallelize.

- **Escape Bootstrapping as possible as we can**
  2. If SWHE is applicable at some execution, use it.

<table>
<thead>
<tr>
<th>Types</th>
<th>Crypto operation</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE</td>
<td>Only Addition</td>
<td>High Speed</td>
</tr>
<tr>
<td></td>
<td>Only Multiplication</td>
<td>High Speed</td>
</tr>
<tr>
<td>SWHE (Somewhat HE)</td>
<td>Multiplication for Once + Addition</td>
<td>Medium Speed</td>
</tr>
<tr>
<td></td>
<td>Multiplication for a few times + Addition</td>
<td>Medium Speed</td>
</tr>
<tr>
<td>FHE (Fully HE)</td>
<td>Addition + Multiplication for N times</td>
<td>Very Low Speed</td>
</tr>
</tbody>
</table>

- **Ideal Lattice**
  - (i) Ideal Lattices-based
  - (ii) Integer-based
  - (iii) LWE-based

- **Closest Vector Problem**

- **It is difficult to find the closest point, as the lattice shape narrows!**
4. Research Strategy

- Off-load Engine/Stream Processing/Migration
  - Parallelization & adopt FPGA
  - Stream-processing called Queue Linker platform
  - Inter-cloud migration

- I/O tuning / optimization

- Cache unfriendly tuning of workload
  - Effective use of “memory hierarchy”

<table>
<thead>
<tr>
<th>Latency(clock)</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registers</td>
<td>1</td>
</tr>
<tr>
<td>L1 cache</td>
<td>4+</td>
</tr>
<tr>
<td>L2 cache</td>
<td>11+</td>
</tr>
<tr>
<td>L3 cache</td>
<td>24+</td>
</tr>
<tr>
<td>DRAM</td>
<td>200–400</td>
</tr>
<tr>
<td>SSD</td>
<td>350,000</td>
</tr>
<tr>
<td>HDD</td>
<td>35,000,000+</td>
</tr>
</tbody>
</table>

- Data Mining Library based on FHE

NEW CHALLENGE

SD² Platform for Integrated Big Data Utilization
5. Experiment

Experimental demonstration

- show the effectiveness of our platform

- Life Log Analysis (sensor data)
  - Gathering hundreds of thousands users data (raw 1TB data)
  - Proof of Storage
  - verifiable delegation of computation

- Drug Adverse Analysis (text data)
  - Gathering over 2 million users’ drug
  - Proof of Storage
  - Secure multiparty computation with fully homomorphic encryption
  - verifiable delegation of computation
  - attribute based encryption
### 6. Schedule

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outlook</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mid-term evaluation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Final Evaluation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Legal Study</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Legal Coordination</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guideline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Encryption</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attribute-based encryption</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>fully homomorphic encryption</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Proof of storage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computer Architecture</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friendly algorithm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improvement based on LS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over 30 times faster</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over 30 times faster</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hierarchy of Storage</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-fetch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I/O optimization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FPGA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Platform Const. at Waseda&amp;Ochanomizu</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Improvement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over 30 times faster</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over 30 times faster</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Experimental</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Life log</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over 1,000 logs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>drug adverse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Several drugstores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of platform</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over 10,000 logs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>drug adverse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Several drugstores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of platform</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over 10,000 logs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>drug adverse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Several drugstores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of platform</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over 20 times</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Several drugstores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Parallelizing by using “Ideal lattice” base encryption**
- **Over 10^3 times** faster (practical use level)

**Use the gap between Sequential and Random Access speed**
7. PROGRESS IN 2015FY

- **Legal Study**
  - Studied possible data transfer and analysis under the provision of 2015 Japanese amendment of Act on the protection of personal Information.

- **Encryption Algorithm**
  - Proposed a theory of FHE for real numbers called FHE4FX.
  - It enables Homomorphic Greater-Than-bit computation.

- **Implementation**
  - Implemented “Apriori algorithm,” 10 times faster than the state-of-the-art method by adopting packing with HElib.

- **Platform**
  - Analyzed I/O performance where data are on outer/inner zone of platter with large scale data access.
  - Prepared our Cloud Platform between Waseda Univ. and Ochanomizu Univ.
THANK YOU