Privacy-Preserving Data Mining

*Parallel Homomorphic Encryption*

Mandana Bozorgi  
Computer Science Department  
Las Vegas, NV, USA  
Mandana.Bozorgi@unlv.edu

Yoohwan Kim, Ph.D., CISSP, CISA, CEH, CPT  
Computer Science Department  
Las Vegas, NV, USA  
Yoohwan.Kim@unlv.edu
Agenda

- Hospitality data risk
- Big Data
- Map Reduce
- Homomorphic Encryption
- Parallel Homomorphic Encryption
- Map Reduce Parallel Homomorphic Encryption
Hotel Data Risks

- Between May 2014 and June 2015, hackers took credit card information from various branches of Trump hotels in US and Canada.
- Between Apr 2015 and Jul 2015, Hilton Hotels got hacked in four different branches.
- In Feb 2014, Sands corporation that owns several hotel properties like Venetian and Palazzo got hacked.
Data in Hospitality

- In Hotel industry, vast amount of data is collected and kept daily, e.g., customer’s reservation, food purchasing, video surveillance data, gaming activity records, credit card information, and employee SSN, etc.

- Working with this massive data brings some challenges.
Solving the Volume Problem

- Traditional data storage and processing
  - Relational database (SQL)
- Now
  - The issue for storing the massive and unstructured data is that the traditional way is not efficient and sometimes not possible.
- Why?
  - We cannot manage the unstructured data with SQL.
  - Working with the Petabyte size of data with SQL is not efficient.
- Solution
  - Hadoop is designed to work with massive and unstructured data.
Hadoop Data Storage

Hadoop File System (HDFS)

Big Data
(e.g., one or more txt files)

compute node 0
compute node 1
compute node 2
compute node 3
compute node 4
compute node 5

HDFS

big data piece
big data piece
big data piece
big data piece
big data piece
big data piece

## SQL vs. Hadoop

<table>
<thead>
<tr>
<th>Description</th>
<th>RDBMS (SQL)</th>
<th>Hadoop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Traditional row-column database for transactional system, reporting, and archiving</td>
<td>Distributed file system that stores large amount of data on a cloud. In addition, Hadoop provides an API for processing all stored data.</td>
</tr>
<tr>
<td>Data Size</td>
<td>Gigabytes (terabytes)</td>
<td>Petabytes (Exabytes)</td>
</tr>
<tr>
<td>Structure</td>
<td>Work with structured data only</td>
<td>Work with structured, semi structured, and unstructured data</td>
</tr>
<tr>
<td>Update</td>
<td>1000s queries per second</td>
<td>Millions of queries per second</td>
</tr>
</tbody>
</table>
MapReduce

- A programming model
  - Two functions: Map and Reduce.

- Designed for large-scale parallel data processing, like process 20 PB of data daily.
  - Pioneered by Google
  - Many companies use MapReduce like Google, Yahoo!, Facebook, etc.
Hadoop & MapReduce

- Hadoop is a platform for MapReduce programs.
- **Mapper**: Takes raw data file as input and maps input key/value pairs to a set of intermediate key/value pairs in the form of file and it is stored in the HDFS.
- **Reducer**: After processing, it produces a new set of output, which will be stored in the HDFS.
MapReduce Example

The overall MapReduce word count process

Input

Splitting

Mapping

Shuffling

Reducing

Final result

Deer Bear River

Car Car River

Deer Car Bear

Deer, 1
Bear, 1
River, 1

Car, 1
Car, 1
River, 1

Deer, 1
Deer, 1

River, 1

Bear, 1
Bear, 1

Car, 1
Car, 1
Car, 1

Deer, 1

Deer, 2

River, 1

River, 2

Bear, 2

Car, 3

Deer, 2

River, 2

Bear, 2

Car, 3

Deer, 2

River, 2

Bear, 2

Car, 3

Deer, 2

River, 2
Tying them together
Solving the Security Problem

- While using public cloud, my private data is exposed!

- Wouldn’t be nice if we are able to:
  - Encrypt our data before storing them in the cloud.
  - Store data in encrypted form in the cloud.
  - Send our queries to the cloud while they're encrypted and allow the cloud to process the encrypted queries.
  - Get encrypted answers and decrypt the result.
Challenges

- Once data is encrypted, it becomes meaningless until decrypted
  - Storing is ok, but cannot be processed.

- Solution: Homomorphic Encryption
  - Allows complex mathematical operation to be performed on encrypted data without decryption.
Understanding Homomorphic Encryption

**SEARCH**

- Alice
- ALICE
- BOB
- EVE
- IbAZ6
- Tz73b
- 4zLE9

**ADDITION**

- $B^2 \times B^3 = B^5$
- Encrypt: Raise B to this power
- Decrypt: Take log base B

- $2 + 3 = 5$

**MULTIPLICATION**

- $2A \times 3B = 6AB$
- Encrypt: Multiply by A
- Encrypt: Multiply by B
- Decrypt: Divide by AB

- $2 \times 3 = 6$
Some history

- Cryptographers and researchers were aware of homomorphic encryption since 1978.
  - 1978, Rivest, Adleman and Dertouzos: “On Data Banks and Privacy Homomorphism”
  - 1999, Pascal Paillier: Encrypted multiplication system
  - 2009, Craig Gentry (Stanford student then): first fully homomorphic algorithm, but million times slower
  - 2011, Raluca Ada Popa & Nickolai Zeldovich, only 27% slower
Homomorphic Encryption in Cloud

- Is expected to play a significant role in cloud computing allowing companies to store encrypted data in a public cloud and take advantage of the cloud providers’ analytic services.
MapReduce Parallel Homomorphic Encryption Scheme

- Generate a key, \( K \)
- Encryption: \( \text{Enc} (m, K) \rightarrow C \)
- Token generation takes a function \( f: \text{Tok} (f, K) \rightarrow t \)
- Parser creates sequence of input pairs:
  \[ \text{Par} (t, C) \rightarrow (x_i, v_i)_i \]
- Mapper takes input pairs, returns a sequence of intermediate pairs
  \[ \text{Map} (x, v) \rightarrow (\alpha_j, \beta_j)_j \]
- Shuffle takes intermediate pairs, and returns \( h \) in space \( H \), label \( \delta \), and partition \( P \)
- This will be input for Reducer, which returns output \( (\alpha, \zeta) \)
- Merge takes \( (\alpha_t, \zeta_t)_t \) pairs and returns a ciphertext \( C' \).
Key Management

- The keys that we used needs to be protected.
- We can store the keys in a secure key management system.
- In parallel programming or distributed programming like Hadoop, key needs to be retrieved by different workers as part of their job.
- The key is transferred from the key management system to MapReduce task by each request.

- In the MapReduce Parallel Homorphism, key generator generates two keys: Private Key and Public Key.
MapReduce Parallel Homomorphic Encryption Controller Job

- Controller in cloud:
  - Receives a ciphertext C’ and a function f
  - Generates n jobs, assign the tasks to the workers running in parallel

- The result of all workers’ job is a ciphertexts’ output (C’)

Owner of data

Controller

Ciphertext Ĉ

function f
Conclusions

- A good method to protect hospitality data in the cloud is using the homomorphism encryption.

- The future work will be to have the experimental results of different encryption on cloud and compare their accuracy with parallel homomorphic encryption on Hospitality big data.