Secure RDD for SparkFHE

Sai Sujith Kammari
Dr. Peizhao Hu (Advisor)
TABLE OF CONTENTS

01  Project Refresh
    What is Secure RDD

02  Accomplished Work
    HE BLAS & Secure K-means

03  Evaluation
    How was the work evaluated

04  Conclusion & Future Work
    Extension to the project
01
Project Refresh

What is Secure RDD for SparkFHE
Secure RDD for SparkFHE

- Apache Spark is the most popular open-source, distributed processing system used for big data workloads.

- Workloads in Apache Spark
  a. Core
  b. MLib
  c. Streaming
  d. SQL
  e. GraphX

- Apache Spark provides high-level APIs in Java, Scala, Python and R, and supports code reuse across multiple workloads.

- Implementing homomorphic encryption in the RDDs which are used in MLib.
Accomplished Work

Homomorphic Encrypted BLAS & Secure K-means
Secure BLAS

- Basic Linear Algebra Subprograms
  - axpy: $y = y + (a \cdot x)$
  - scal: $x = a \cdot x$
  - dspr: $A = \alpha \cdot x \cdot x^T + A$
  - dspmv: $y = \alpha \cdot A \cdot x + \beta \cdot y$
  - dswap: $x \leftrightarrow y$
  - dtpmv: $x = A \cdot x$ (or) $x = A^T \cdot x$
Secure Machine Learning

- K-Means clustering
  - Iterative unsupervised learning algorithm
  - Aims to partition data into k clusters based on the similarity of the data points
  - Formally: \( \arg\min \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 \) where \( \mu_i \) is the mean of points in \( S_i \)
- Challenges with Homomorphic Encrypted data
  - Encrypted data cannot be checked for equality
  - Encrypted values cannot be compared

Secure K-Means clustering

- Data owner participation is required
- Additional information, Updatable Distance Matrix (UDM) matrix, will be sent to cloud along with the encrypted data
- Cloud takes help of UDM to calculate the similarity
- Data owner decides when the clustering has to be stopped

Data, $D = \{ 
\{1,2\}, \\
\{4,9\}, \\
\{3,8\}\n\}$
$r_y$ where $x = 1$; $r_1 = \{1,2\}$
$m = 2$ attributes
$a_1 = 1; a_2 = 2$

UDM, $U = \{ 
\{0.0, 0.0\}, \\
\{3.0, 7.0\}, \\
\{2.0, 6.0\} 
\}, 
\{ 
\{3.0, 7.0\}, \\
\{0.0, 0.0\}, \\
\{1.0, 1.0\} 
\}, 
\{ 
\{2.0, 6.0\}, \\
\{1.0, 1.0\}, \\
\{0.0, 0.0\} 
\}\}

Explanation with $U[0]$
$x = 0$ $y = 1$
$\{0.0, 0.0\} = \{\text{abs}(1-1), \text{abs}(2-2)\}$
$\{3.0, 7.0\} = \{\text{abs}(1-4), \text{abs}(2-9)\}$
$\{2.0, 6.0\} = \{\text{abs}(1-3), \text{abs}(2-8)\}$
Secure Machine Learning

- **Secure K-Means clustering**
  - Data Owner generates UDM and encrypts data
  - Sends the UDM and encrypted data to cloud
  - Cloud populates the clusters by calculating the similarities using the UDM information
  - Cloud returns the new centroids to the data owner
  - Data owner compares the new centroids and decides to continue clustering or stop
03 Evaluation

How was my work evaluated
Evaluation

- BLAS methods

  - Operated on encrypted data and cross verified actual results with the expected

    ```
    Running axpy Example
    y ← a*x where a = 3, y = [2,2,2] and x = [1,0,1] Expected: [5 2 5]
    Actual: [5.00 2.00 5.00 ]
    Running scal Example
    x = a*x where a = 3, x = [1,1,1] Expected: [3 3 3]
    Actual: [5.00 3.00 3.00 ]
    Running depp example
    ====== Before Calculation ======
            3.00  1.00  2.00
    ====== After Calculation ======
            3.00  259.00  114.00
    ```

- K-Means clustering

  - Trained on iris data set and verified if all the flowers are in correct clusters

  - On hold as it requires data owner involvement in the process
Conclusion & Next steps

Extension to the project
Conclusion & Next steps

**BLAS and Kmeans clustering**
Other BLAS methods can be implemented. Build a new algorithm to group the data points into K clusters with less data owner involvement.

**Sparse vector**
Until now, no implementation is started for sparse vector operations.

**Secure Machine Learning**
Implement other Machine Learning algorithms on the encrypted data using the BLAS methods.
Any Questions
THANKS

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik. Please keep this slide for attribution.