Secure resilient distributed dataset for SparkFHE

Sai Sujith Kammari
Dr. Peizhao Hu (Advisor)
Apache Spark
Apache Spark

Introduction

• Apache Spark is an open-source, distributed processing system used for big data workloads.

• Provides high-level APIs in Java, Scala, Python and R, and supports code reuse across multiple workloads

• Started as a research project at UC Berkley’s AMP Lab in 2009 to create a framework, optimized for fast iterative processing

• Most popular big data distributed processing framework with 365,000 meetup members in 2017.
Apache Spark

How is it better?

- Spark is faster than Hadoop MapReduce as it does all processing in-memory

- Spark reuses data by using an in-memory cache to greatly speed up ML algorithms

- In-memory operations dramatically lowers the latency making multiple times faster, especially when doing ML and interactive analytics
Apache Spark

Workloads

- MLLib: Machine Learning
- Streaming: Real-time analytics
- SQL: Interactive Queries
- GraphX: Graph processing

Languages:
- R
- Python
- Scala
- Java
Security
Encryption is the process of converting normal message (plaintext) into meaningless message (Ciphertext)

Decryption is the process of converting meaningless message (Ciphertext) into its original form (Plaintext)

Training models on encrypted data predicts meaningless
Homomorphic encryption (HE) allows functions to be computed on ciphertexts without decryption.

Any function that can be represented by a low-degree polynomial with addition, subtraction, and multiplication can be supported by HE.

Data is protected in-transmit, at-rest, and in-use as results from the computations are also encrypted.
• **Encryption function** \( \text{Enc}(m) = 2m \)

• **Decryption function** \( \text{Dec}(\text{Enc}(m)) = \text{Enc}(m)/2 \)

• Plaintext addition for two numbers, \( 5 + 3 = 8 \), and homomorphic addition of two ciphertexts, \( \text{Dec}(\text{Enc}(5) + \text{Enc}(3)) = \text{Dec}(10 + 6) = 8 \).
Milestones
<table>
<thead>
<tr>
<th>Milestone#</th>
<th>Milestone Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milestone1</td>
<td>Understand the inner workings of Apache Spark core and MLlib libraries</td>
</tr>
<tr>
<td>Milestone2</td>
<td>Understand working of Homomorphic Encryption and implement in <em>some</em> of the Mlib methods</td>
</tr>
<tr>
<td>Milestone3</td>
<td>Contribute to the ongoing SparkFHR project by integrating parts of Mlib with Homomorphic encryption</td>
</tr>
</tbody>
</table>
Questions
Thank You