Primitives for secure data analytics on SparkFHE

Soumyakanti Das

Advisor: Dr. Peizhao Hu
Motivation

- **Long Term**: Data analytics on private encrypted data
- Understand the inner workings of Apache Spark core and MLlib libraries
- Contribute to the ongoing SparkFHE project by integrating parts of MLlib with Homomorphic Encryption

Source: https://www.cs.rit.edu/~ph/PrivateComputation
Apache Spark

- Unified data analytics engine for processing large-scale data
- In memory data processing engine
- Different levels of abstractions - RDD, Dataframe, Dataset
- Libraries - Spark core, MLlib, Streaming, GraphX

Apache Spark Ecosystem
Source: https://databricks.com/spark/about

Apache Spark Cluster Overview
Source: https://spark.apache.org/docs/latest/cluster-overview.html
Homomorphic Encryption

- A cryptographic technique that enables us to transform encrypted data, without decrypting it first
- Data is encrypted by the user, sent to cloud computing engines (Spark), and then the result can be decrypted by the user.
- Computationally heavy

Source: https://www.cs.rit.edu/~ph/PrivateComputation
SparkFHE

sparkfhe-examples (.jar or .zip)
spiritlab.sparkfhe.examples.*
(examples and tests)

sparkfhe-addon
spiritlab.sparkfheaddon.*

spark-submit

Spark
Apache Spark Core
Apache Mesos / Yarn, Kubernetes
Distributed File System

Cloudlab.us cluster

libSparkFHE
primitives and algorithms

sparkfhe-plugin.jar
org.apache.spark.*

- MLlib
- SQL
- Streaming
- GraphX

sparkfhe-api-java

sparkfhe-api-python

libfhe (HElib)
libseal (SEAL)
libothers (others)

Source: https://www.cs.rit.edu/~ph/PrivateComputation
Proposed Work

- Understand design, architecture, and implementation of Spark core and MLlib libraries
- Integrate various MLlib APIs to use HE
- Write test cases and demo examples for the APIs
- Fix bugs in the existing source code
Results and Evaluation

- Significant contributions to the MLlib APIs - integrating HE, testing, and writing examples
- Understanding basics of Homomorphic Encryption
- Writing a report/paper at the end with details of my findings and implementations
Thanks

Questions?