Synopses-Driven Data Integration and Federated Learning

Eros Fabrici @ Athena Research Center, eros.fabrici@athenarc.gr

Overview

Data Integration (DI) is the set of processes to gather and bridge data from heterogeneous sources together in order to have a unified view. It has been studied for for more that 50 years and it is gaining more importance thanks to the raise of Big Data. In particular, privacy in data science is an increasing concern and a lot of research effort is being carried out for improving privacy guarantees in all aspects of data science and management.

Federated Learning (FL) is a machine learning technique that aims to train an algorithm across multiple decentralized edge devices or servers holding local data samples, without moving the data. FL assumes the data to be aligned in order to proceed with the model training, which is not the case in real-world scenarios.

The aim of this PhD is to design and develop a DI an effective data integration solution tailored to the FL data preparation step, with a big concern on privacy and efficiency. The main approaches that will be studied are synopses, differential privacy and machine learning

Data Integration

Data Integration aim to create an unified view of different data sets. It involves three steps: *schema alignment*, *record linkage* and *data fusion*. In this PhD we will focus on the first two.

Schema alignment is the technique of identifying objects which are semantically related. In other words, schema matching is a method of finding the correspondences between the concepts of different distributed, heterogeneous data sources. Recently, the Universal Schema approach has revolutionized this aspect, which is based of matrix factorization or recurrent neural networks.

Record Linkage (RL) is the task of finding records in a data set that refer to the same entity across different data sources. The main concerns of modern record linkage research include making it scalable and guaranteeing privacy. Privacy Preserving Record Linkage is a sub-field of RL where preventing information leakage is the main concern.

Differential Privacy

Differential Privacy is a technique for publicly sharing information about a dataset by describing the patterns of groups within the dataset while withholding information about individuals in the dataset. It is *de facto* standard in the field of privacy. It is used in querying, synthetic data generation and machine learning model training.

Federated Learning

The main idea of Federated Learning is to build machine learning models based on data sets that are distributed across multiple devices while preventing data leakage. It can be distinguished in two categories, *Horizontal Federated Learning*, where the datasets share the feature space and but different in samples, and *Vertical Federated Learning* where the datasets share the samples space (different records in different datasets referring to the same entity) but have different features.

Our work will focus on Vertical FL as it implies a data preparation phase where the data needs to be aligned and linked, therefore there is the need to move the data in order to apply the comparisons.

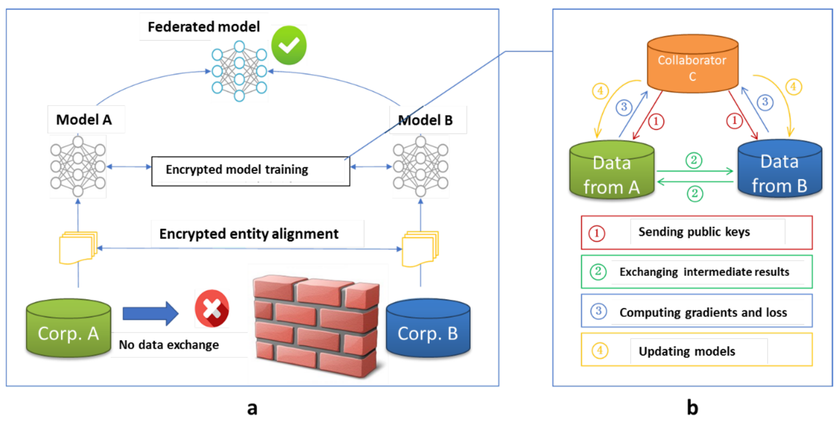

Figure 1: High Level Vertical FL Architecture

Future Goals

The objective of the project is analyze the current state-of-the-art of Privacy Preserving Schema Alignment and Record Linkage and propose a solution, tailored to the FL scenario, based on synopses, differential privacy and machine learning and finally compare it with the current approaches.