Doctorate Project Plan Presentation
Synopses-Driven Data Integration & Federated Learning

PhD Candidate: Eros Fabrici

Athena Research Center & Universitat Politecnica de Catalunya
Data Engineering for Data Science - ESR 3.2

Supervisor: Professor Minos Garofalakis,

November 10, 2023

Contents

1 Problem Statement
- Data Integration for Federated Learning

2 Federated Learning
- Definition
- Federated Learning Categories

3 Differential Privacy
- What is Differential Privacy
- Formal Definition

4 Data Integration
- Data Integration Categorization
- Schema Alignment
- Record Linkage

5 Goal of this PhD
- Improving FL
- Objectives & Methodology & Challenges

Contents

1 Problem Statement
- Data Integration for Federated Learning

2 Federated Learning
- Definition
- Federated Learning Categories

3 Differential Privacy
- What is Differential Privacy
- Formal Definition

4 Data Integration
- Data Integration Categorization
- Schema Alignment
- Record Linkage

5 Goal of this PhD
- Improving FL
- Objectives & Methodology & Challenges

Data Integration for Federated Learning

- Data Integration (DI) is the process of gathering data from disparate sources and fusing them in order to have an unified view.
- Big Data introduced new challenges for DI, in particular *scalability* and *guaranteeing privacy*.
- This requires techniques to guarantee privacy, computational efficiency and efficacy (correct matching results).
- Federated Learning (FL) is a machine learning technique where a federation of edge-devices aims to build a global model without moving the data to a central entity.
- Aligning and Linking the data is done manually.

Contents

1. Problem Statement
   - Data Integration for Federated Learning
2. Federated Learning
   - Definition
   - Federated Learning Categories
3. Differential Privacy
   - What is Differential Privacy
   - Formal Definition
4. Data Integration
   - Data Integration Categorization
   - Schema Alignment
   - Record Linkage
5. Goal of this PhD
   - Improving FL
   - Objectives & Methodology & Challenges

Overview

- Federated Learning was proposed recently by Google (1, 2, 3).
- Its main advantage is to be able to build a global model to be shared between a federation of data owners, without exchanging the data between them.
- Many efforts have been made to improve security and statistical challenges (4).

Contents

1 Problem Statement
- Data Integration for Federated Learning

2 Federated Learning
- Definition
- Federated Learning Categories

3 Differential Privacy
- What is Differential Privacy
- Formal Definition

4 Data Integration
- Data Integration Categorization
- Schema Alignment
- Record Linkage

5 Goal of this PhD
- Improving FL
- Objectives & Methodology & Challenges

A categorization for FL

- There are two main categories of FL
- **Horizontal** FL: same feature space, different sample space.
- **Vertical** FL: different feature space, shared sample space.

Horizontal Federated Learning

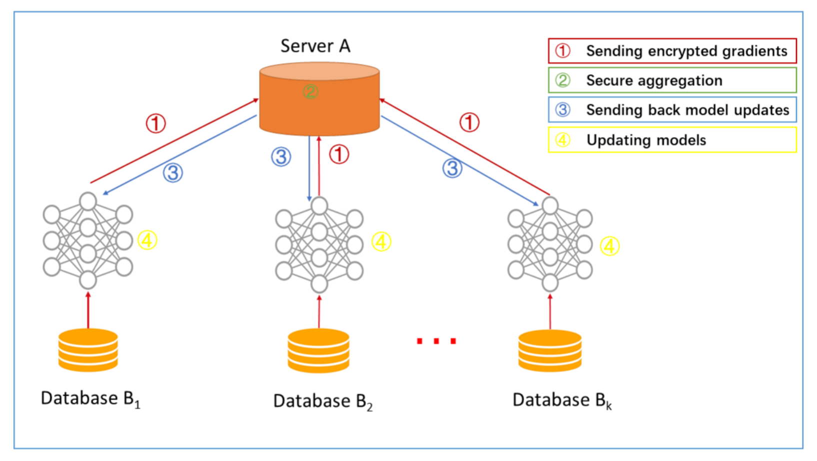

Figure: Example of Horizontal FL Architecture

Vertical Federated Learning

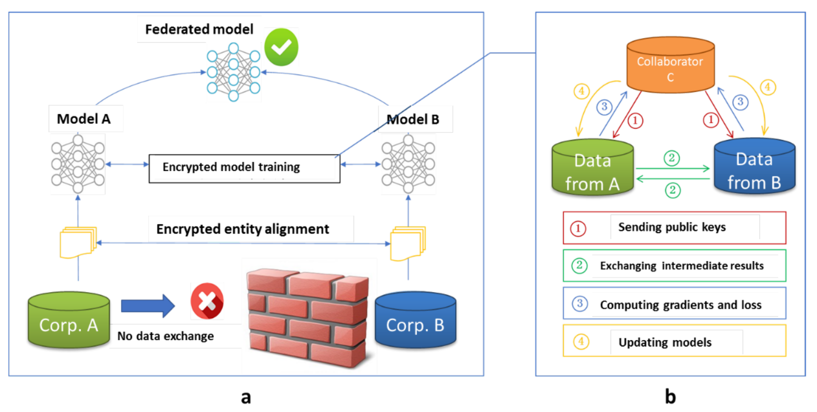

Figure: Example of a Vertical Federated Learning Architecture

Contents

1 Problem Statement
- Data Integration for Federated Learning

2 Federated Learning
- Definition
- Federated Learning Categories

3 Differential Privacy
- What is Differential Privacy
- Formal Definition

4 Data Integration
- Data Integration Categorization
- Schema Alignment
- Record Linkage

5 Goal of this PhD
- Improving FL
- Objectives & Methodology & Challenges

Differential Privacy Intuition

- Concerns of private data analysis: *membership* and *information* inference.
- Differential Privacy (DP) addresses these concerns.
- DP addresses the paradox of learning nothing about an individual while learning useful information about a population (5).
- Originally used in querying, now also for statistics, machine learning and synthetic data generation.

Contents

1 Problem Statement
- Data Integration for Federated Learning

2 Federated Learning
- Definition
- Federated Learning Categories

3 Differential Privacy
- What is Differential Privacy
- Formal Definition

4 Data Integration
- Data Integration Categorization
- Schema Alignment
- Record Linkage

5 Goal of this PhD
- Improving FL
- Objectives & Methodology & Challenges

Definition

Definition

A randomized algorithm $A : U \longrightarrow O$ is ϵ-differentially private if for $o \subseteq O$ and for all pairs of adjacent datasets $D, D' \in U$

$$\mathbb{P}[A(D) \in o] \leq e^{\epsilon}\mathbb{P}[A(D') \in o]$$

where the probability space is over the coin-flips of A

Contents

1 Problem Statement
- Data Integration for Federated Learning

2 Federated Learning
- Definition
- Federated Learning Categories

3 Differential Privacy
- What is Differential Privacy
- Formal Definition

4 Data Integration
- Data Integration Categorization
- Schema Alignment
- Record Linkage

5 Goal of this PhD
- Improving FL
- Objectives & Methodology & Challenges

Data integration

- Data Integration (DI) is the process of combining data from different sources into a single unified view.
- It is divided in three main steps: *Schema Alignment*, *Record Linkage* and *Data Fusion*
- We will focus on the first two steps.

Contents

1 Problem Statement
- Data Integration for Federated Learning

2 Federated Learning
- Definition
- Federated Learning Categories

3 Differential Privacy
- What is Differential Privacy
- Formal Definition

4 Data Integration
- Data Integration Categorization
- Schema Alignment
- Record Linkage

5 Goal of this PhD
- Improving FL
- Objectives & Methodology & Challenges

Schema Alignment

- Process that builds a mapping between data sources and a global schema or that creates a mediated schema between data sources.
- We can categorize it in three main types:
  - *Schema-level matchers*
  - *Instance-level matchers*
  - *Hybrid matchers*

Schema Alignment cont'd

- Universal Schema (6) has revolutionized schema alignment.
- It consists of inferring relations, by extracting triples (subject, predicate, object). This is done via matrix factorization, and recently via Recurrent Neural Networks (7).

Contents

1. Problem Statement
   - Data Integration for Federated Learning
2. Federated Learning
   - Definition
   - Federated Learning Categories
3. Differential Privacy
   - What is Differential Privacy
   - Formal Definition
4. **Data Integration**
   - Data Integration Categorization
   - Schema Alignment
   - **Record Linkage**
5. Goal of this PhD
   - Improving FL
   - Objectives & Methodology & Challenges

Record Linkage

- Record Linkage (RL) consists of finding records across different datasets that refer to the same real-world entity.
- It has been studied for more than 50 years [8].
- It is generally composed of three steps: (1) *blocking*, (2) *compare pair of records*, (3) *clustering records*.

Privacy-Preserving Record Linkage

- Over the last decade, the rise of Big Data introduced a new challenge for RL: **guaranteeing privacy**.
- Privacy-Preserving Record Linkage (PPRL) aims to tackle the privacy problem.
- The main challenges in PPRL are:
  - Guarantee at the same time: **scalability**, **efficacy** and **full end-to-end privacy**.
  - Moreover, most of the work is focused on PPRL between two datasets.

Contents

1. Problem Statement
   - Data Integration for Federated Learning
2. Federated Learning
   - Definition
   - Federated Learning Categories
3. Differential Privacy
   - What is Differential Privacy
   - Formal Definition
4. Data Integration
   - Data Integration Categorization
   - Schema Alignment
   - Record Linkage
5. Goal of this PhD
   - Improving FL
   - Objectives & Methodology & Challenges

FL process

- In FL research, data is assumed to be already aligned.
- In real-world scenarios aligned is done manually or by ad-hoc solutions by engineers.
- There are approaches for this problem that work on the learning algorithms (9), but not approaches that work on the data.
- Challenges:
  - Automated Schema Alignment + PPRL.
  - Perform the task in a effective and efficient manner, by extending these techniques to a multi-party scenario.
  - Ensure that the model's accuracy does not degraded excessively.

Contents

1 Problem Statement
- Data Integration for Federated Learning

2 Federated Learning
- Definition
- Federated Learning Categories

3 Differential Privacy
- What is Differential Privacy
- Formal Definition

4 Data Integration
- Data Integration Categorization
- Schema Alignment
- Record Linkage

5 Goal of this PhD
- Improving FL
- Objectives & Methodology & Challenges

Objectives

- Design and implement a synopses-driven and differentially private:
  - multi-party instance-based algorithm for Schema Alignment;
  - multi-party PPRL solution;
- Compare the algorithms proposed with the state-of-the-art solutions and analyze their computational performance and how they affect the learned FL models.

Methodology

- Study of the state-of-the-art techniques for Schema Alignment and PPRL.
- Study the applications of synopses and differential privacy for scaling DI.
- Develop algorithms for schema alignment and PPRL for FL.
- Benchmarking and Evaluation of the algorithms proposed.
- Analyze how the FL is impacted by those algorithms (time saved against accuracy loss).

Challenges

- Develop a solution that, at the same time:
  - guarantees a *good level of privacy* wrt the FL context.
  - improves *computational performance* wrt the state-of-the-art.
  - minimizes the *loss of accuracy* in the DI phase as well for the FL model.

References I

(1) Jakub Konečný et al. *Federated Optimization: Distributed Machine Learning for On-Device Intelligence*. 2016. DOI: 10.48550/ARXIV.1610.02527. URL: https://arxiv.org/abs/1610.02527.

(2) Jakub Konečný et al. *Federated Learning: Strategies for Improving Communication Efficiency*. 2016. DOI: 10.48550/ARXIV.1610.05492. URL: https://arxiv.org/abs/1610.05492.

(3) H. Brendan McMahan et al. ``Communication-Efficient Learning of Deep Networks from Decentralized Data''. In: (2016). DOI: 10.48550/ARXIV.1602.05629. URL: https://arxiv.org/abs/1602.05629.

(4) Peter Kairouz et al. ``Advances and open problems in federated learning''. In: *Foundations and Trends in Machine Learning* 14.1-2 (2021), pp. 1{210. ISSN: 19358245. DOI: 10.1561/2200000083. arXiv: 1912.04977.

(5) Cynthia Dwork and Aaron Roth. ``The algorithmic foundations of differential privacy''. In: *Foundations and Trends in Theoretical Computer Science* 9.3-4 (2013), pp. 211{487. ISSN: 15513068. DOI: 10.1561/0400000042.

(6) Sebastian Riedel et al. ``Relation extraction with matrix factorization and universal schemas''. In: *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2013, pp. 74{84.

(7) Rajarshi Das et al. ``Chains of reasoning over entities, relations, text using recurrent neural networks''. In: *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference* 1 (2017), pp. 132{141. DOI: 10.18653/v1/e17-1013. arXiv: 1607.01426.

(8) I. P. Fellegi and A. B. Sunter. ``A Theory for Record Linkage''. In: *Journal of the American Statistical Association* 64 (1969), pp. 1183{1210.

(9) Sicong Che et al. ``Federated Multi-View Learning for Private Medical Data Integration and Analysis''. In: *ACM Transactions on Intelligent Systems and Technology* 1.1 (2022), pp. 1{22. ISSN: 2157-6904. DOI: 10.1145/3501816. arXiv: 2105.01603.

Thank you for your attention.