Consilient:

FEDERATED LEARNING

A brief introduction

14/05/2024

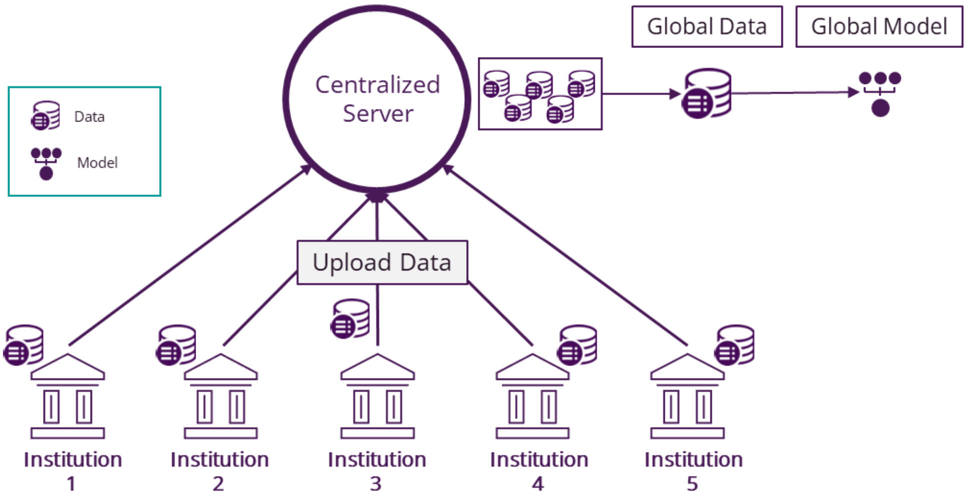
Contents

What is Machine Learning?

Machine learning is a subfield of artificial intelligence that focuses on utilizing large (centralized) datasets to develop algorithms, enabling computer systems to:

- Learn from the data.
- Identify patterns in the data.
- Make predictions based on the given dataset.

Traditionally, the approach to machine learning and data analysis has involved collecting all data on a central server, which could be in a data center or in the cloud. This means transporting and storing all the data in one place. Once centralized, machine learning algorithms can be applied to train models using the data. This has been the standard method for machine learning.



However, this centralized approach has several limitations and is not suitable for many scenarios. Specifically, it fails when data cannot be consolidated on a central server or when the data available on one server is insufficient for training an effective model.

There are several reasons why the classic centralized machine learning approach (aggregating all data in one location) is inadequate for numerous critical real-world applications, including:

Regulations: Data privacy regulations often prevent organizations from combining user data across different regions due to varying data protection laws.

User preference: Some use cases demand that data never leave an organization due to strict data security requirements.

Data volume: Combining large datasets from various sources can be prohibitively expensive in terms of processing, management, and storage. Moreover, much of the data may be irrelevant or unnecessary.

In summary, while traditional centralized machine learning has been the norm, it faces significant challenges in terms of regulatory compliance, user expectations, and data management, making it unsuitable for many important applications.

Machine Learning vs Federated Learning

Federated Learning (FL) flips the traditional machine learning approach on its head. Here’s a simple way to understand the difference:

- **Centralized machine learning:** Moves the data to the computation – **DATA MOVES.**
- **Federated learning:** Moves the computation to the data – **MODEL MOVES.**

In traditional machine learning, data is centralized for training. In contrast, Federated Learning trains models locally where the data resides, such as on local devices, servers, or infrastructure (local nodes). The data never moves, staying safe and secure. Only the trained model information, like the parameters (weights from the model's features), is shared between these local nodes at certain intervals to create a global model that all nodes share.

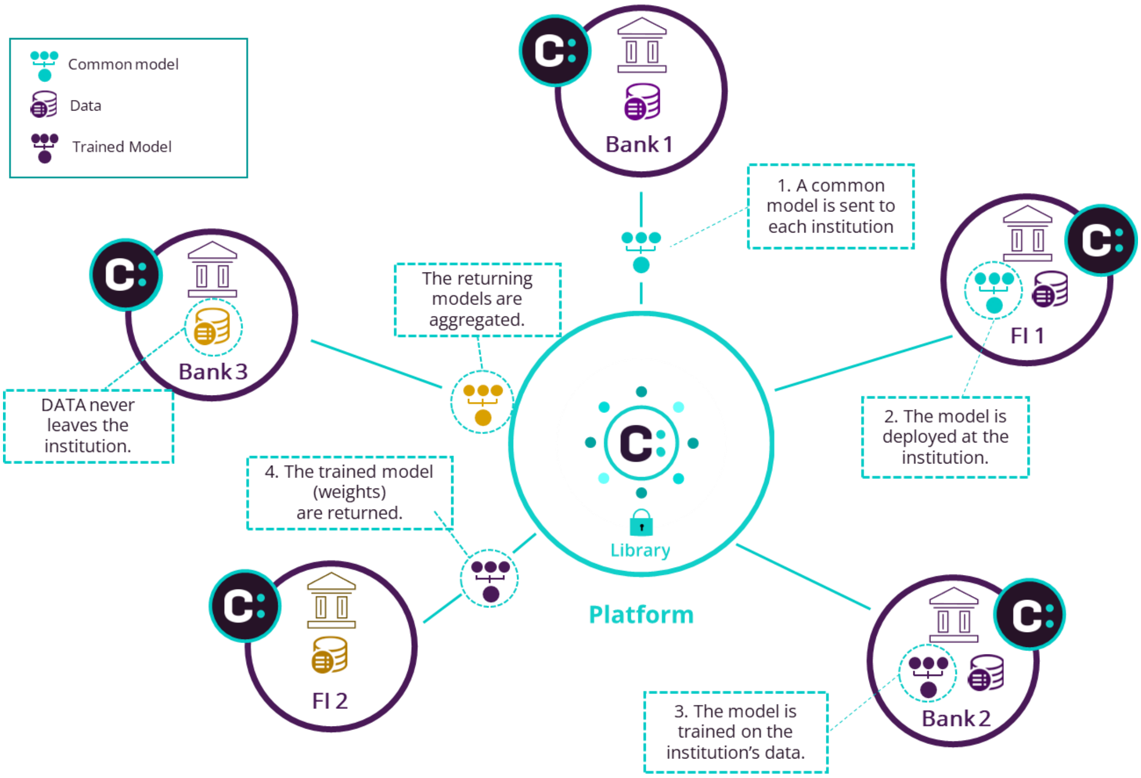
Crucially, with Federated Learning, no actual data is shared. Instead, models are trained on the largest possible datasets available locally, something that wasn't feasible before the advent of Federated Learning.

Federated Learning

Federated Learning (FL) was first introduced by Google in 2017 to tackle the challenges posed by centralized machine learning models. This shift marked a significant advancement in the field, aligning with growing concerns around data privacy and security.

As mentioned earlier, Federated Learning is a type of machine learning that "learns" by traveling across different data sets, which would be difficult to combine in one place. These data sets can be located in different organizations or in distinct parts of the same organization.

The diagram illustrates how FL works:


Common model
Data
Trained Model
Bank 1
1. A common model is sent to each institution
Bank 3
The returning models are aggregated.
FI 1
DATA never leaves the institution.
2. The model is deployed at the institution.
4. The trained model (weights) are returned.
Library
Platform
FI 2
Bank 2
3. The model is trained on the institution's data.

As the Federated Learning (FL) model travels, the data it learns from never leaves the owning organization. Instead, the model learns at each location where the data is held (in the case of Consilient, at each bank).

So how does it actually work?

Step 1: Send a Common Model to Client Nodes

A common model is sent to each organization or business unit (client nodes) to ensure that every participating node starts their local training using the same model parameters.

Step 2: Train the Model Locally

The model is then trained locally on the data of each organization, using their own local dataset to train their local model.

Step 3: Send Model Updates Back to the Server

After local training, each organization ends up with a slightly different version of the model parameters they originally received because each client node has different data examples in its local dataset. This diversity is crucial as it allows the model to get various perspectives from the different nodes. The client nodes then send these model updates back to the server.

Step 4: Aggregate Model Updates into a New Global Model

The model updates (or "gradients") are sent to the Consilient federation platform, where they are combined with updates from other datasets and validated for performance to create a "champion model."

These model gradients are constructed using mathematically rigorous privacy-enhancing technologies and contain no identifiable information about any individual customer.

This resulting model is said to be “federated” because it is trained on the combined learnings from all the models across different nodes. The champion model then travels back to the institutions and will be available for future federation.

Summary

In this way, the Federated Learning model is more accurate and powerfully predictive than any model trained at a single organization. The inherent strength of Federated Learning lies in its accuracy and robustness, making it a superior approach to traditional centralized machine learning.

Federated Learning for AML

As previously mentioned, Federated Learning (FL) is a game-changer for anti-money laundering (AML) because financial crime detection is a fringe phenomenon for most institutions. Even major banks often have small populations of known risk behaviors among their monitored entities, making it difficult to differentiate between unusual but manageable high-risk behavior and truly criminal activity. This challenge is particularly pronounced when identifying specific risk behaviors in underrepresented segments.

The Consilient models are trained on heterogeneous datasets across different entities, encoding a wide spectrum of money laundering risks. This broad exposure enables the creation of models that are not only more accurate in detecting financial crime but also more robust to variations across different jurisdictions, customer types, and behaviors. It allows for the detection of risks that may not have been observed at any specific institution but have been identified elsewhere.

By learning from multiple independent datasets, Federated Learning empowers banks to:

1) Significantly reduce the number of false positive alerts,

2) Lower operational costs,

3) Improve the detection of criminal behavior using insights from a wide array of data rather than just their own company data.

Additionally, FL provides organizations with regular updates and improvements as the model is continuously trained with data from both the bank and other participating entities.