A Novel Self-Serve, Customer Driven Data Standardization Scheme for Federated Learning Platform

(U.S. provisional patent number 63,149,324, Feb. 14, 2021.)

ABSTRACT

This disclosure describes an invention of a data platform that supports multiple data owners collaboratively working on a single task. Most of the existing federated machine learning frameworks assume the data is clean and preprocessed before loading into the framework, thus they focus more on the modeling and communication between each data source. But in the real world, this assumption is not valid, as the data could be dirty, unstructured and unprocessed. Hence, current federated learning frameworks cannot be applied in the large scale real world applications. To solve this problem, our invention builds a data platform which is driven and maintained by users and data transformers in our data platform, to collaboratively clean and preprocess the data. Our data platform contains a data store that saves the metadata of data owners' data, is published and searchable. The data store supports three operations: create, update, and get. Each data owner can add new metadata or update the existing metadata of a specific data in the data store. The data transformers in our data platform can pull the latest metadata from the data store and transform data owner's data accordingly.