Federated Learning

“Privacy By Design” For Machine Learning

Arushi Pandey, Senior Analyst
Nisha Krishan, Senior Analyst

Contents

Introduction

Enterprises have identified Artificial Intelligence (AI) as a quintessential enabling layer in the success of digital transformation. In 2020, more than 72% of enterprises moved beyond PoCs and embarked on their AI journey. As the AI adoption increases, so does the concern for data privacy for users. Data, an extremely crucial piece for the success of any AI implementation, has become sacrosanct, and so has data privacy.

The past few years have seen increased requirement and awareness for safeguarding data privacy among people, coupled with the need to have an agency guarding it. This propelled the rise of government guidelines, regulations, and bills aimed at increasing data privacy and ensuring governance. A few examples are the GDPR in Europe, the CCPA in the US, or the PDP in India. The concept of sovereign cloud is further propelling conversations around data privacy. These guidelines have explicit rules regarding the storage, processing, and usage of data for any purpose; and any breach of these guidelines is heavily penalized.

With this new mindset with regard to data privacy, enterprises have to now put privacy at the center of their AI strategies. This is where federated learning as a concept comes in. Broadly speaking, federated learning is a method of training machine learning models in a way that the user data does not leave its location, and hence is kept safe and private. This is different from the traditional centralized machine learning methods that require the data to be stored in a centralized location to enable training. This ensures that the data does not leave the user and the user has full control over it.

This report introduces the concept of federated learning as a solution to alleviate privacy concerns about machine learning. In this report we explore the following themes:

- Define federated learning and its key characteristics
- Establish a delineation between federated learning and centralized learning
- Detail out benefits and application areas of federated learning
- Provide an overview of ecosystem players for federated learning
- Introduce a decision-making framework for federated learning adoption

Defining federated learning

What is federated learning?

Federated learning is a newfangled mechanism of machine learning, wherein the process of learning takes place in a decentralized manner across a network of nodes / edge devices and the results are aggregated in a central server to create a unified model. It essentially comprises decoupling the activity of model training with centralized data storage. It is a combination of centralized and de-centralized machine learning.

Federated learning can happen in two distinct environments:

- **Private federated learning:** This involves private network with a few devices possessing huge amounts of data and computation power
- **Public federated learning:** This involves public network with a plethora of devices with relatively lesser data and computation power

The concept was first introduced by Google in a whitepaper in 2017 with an academic undertone and has only now begun experiencing adoption in a commercial setting. The word "Federated" implies a system wherein independent units come together to form an alliance. Let us explore that in the context of machine learning.

The mechanism of federated learning

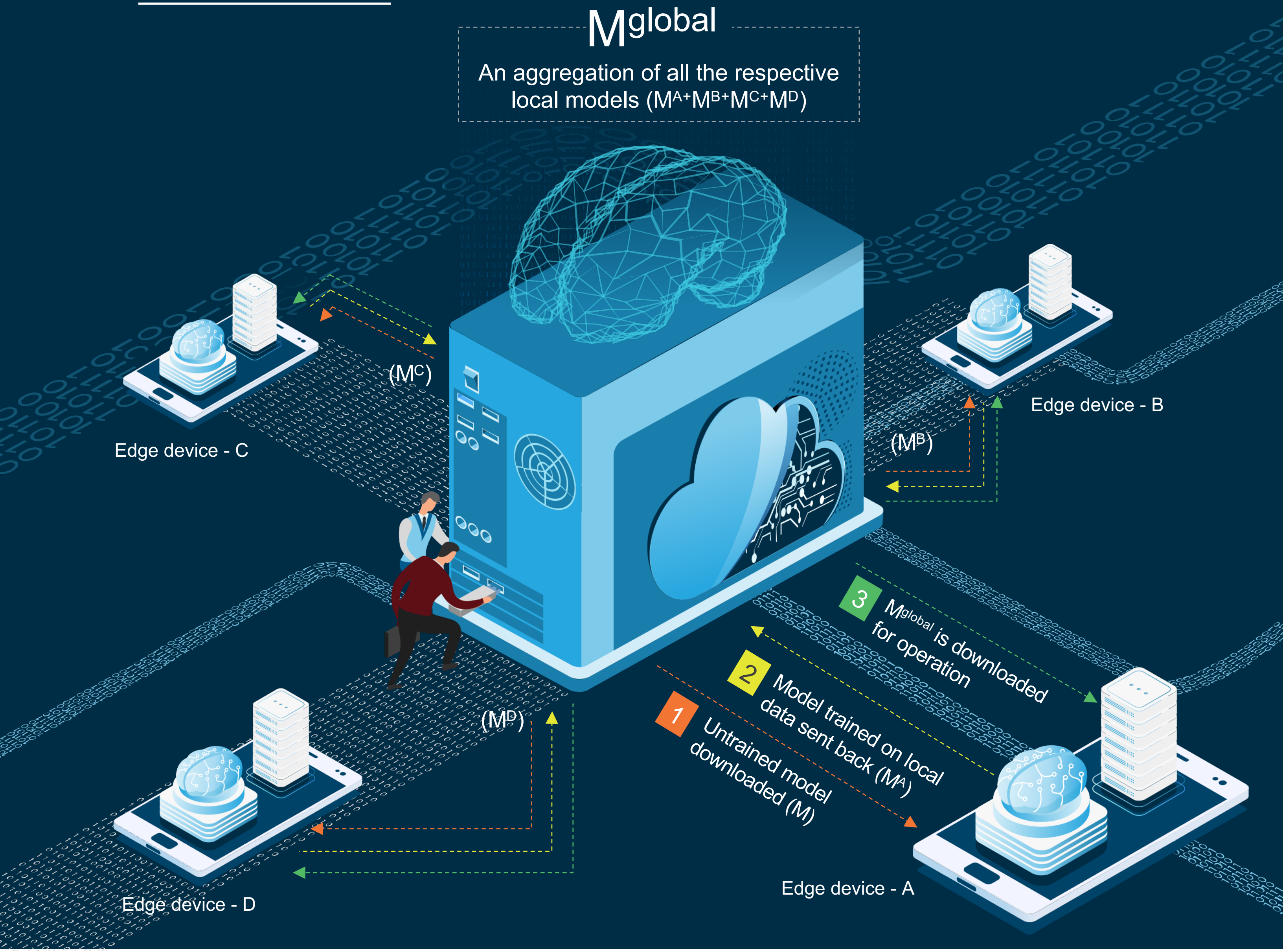
The main principle of federated learning is that a model gets trained on distributed data that resides on different devices or nodes at the edge of a network. The training happens in a de-centralized manner, first in a distributed fashion on the device and then aggregated on a centralized server.

- **On-device/edge training:** The model gets downloaded on to the various devices involved in the process and gets trained on the local data that is present on each of these devices, thus creating an individual on-device local version of the model
- **On-server training:** Once the on-device training is complete, these locally trained models are then sent back to the central servers, where they are combined to create one global consolidated version of the model

Simply put, federated learning is a model training regime that removes the need for data pooling.

The mechanism of federated learning

Source: Everest Group (2021)


M^{global}
An aggregation of all the respective local models ($M^{A}+M^{B}+M^{C}+M^{D}$)
(M^{C})
Edge device - C
(M^{B})
Edge device - B
3 M^{global} is downloaded for operation
2 Model trained on local data sent back (M^{A})
1 Untrained model downloaded (M)
(M^{D})
Edge device - D
Edge device - A

At a very broad level, federated learning is a four-step process that involves the cloud/server and the devices/nodes at the edge of the network. The main principle here being that the model is trained on distributed data that resides with the different devices/nodes of the network.

Step 1: A generic version of the model is created on the cloud/server. This generic version is then pushed to all the devices/nodes that are part of the network

Step 2: Once the devices/nodes have the generic version of the model, the training of the model takes place on the local data that is present on each individual device/node. This ends up creating multiple local versions of the generic model. The devices/nodes then upload or send back their individual version of the model to the main cloud/ server

Step 3: Once all the local versions of the model are uploaded on the cloud, an aggregation process takes place. All the local versions are combined to form one final global version of the model, which is then pushed out to the devices/nodes for operation.

The model or the algorithm that is created as a result of this training is also known as a federated averaging algorithm, as it is an average of the locally trained version of the same model.

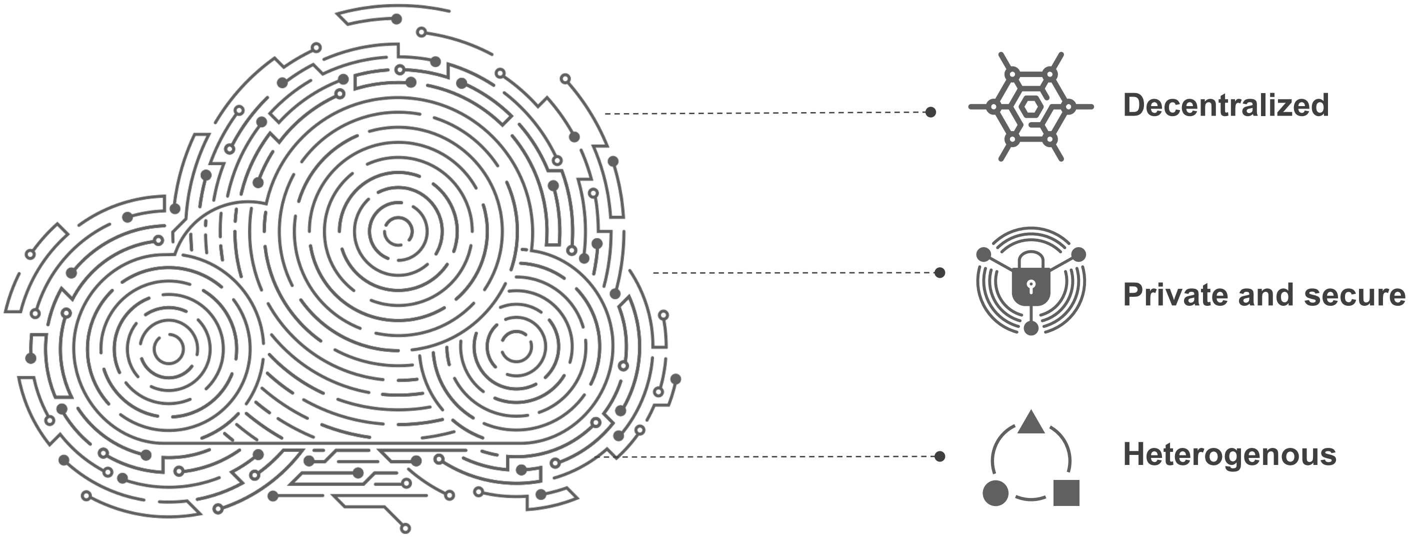
Characteristics of federated learning

- **Decentralized:** The underlying principle of federated learning is its decentralized learning mechanism where the training takes place across the different nodes/devices that participate in a network
- **Private and secure:** Privacy is baked into the mechanism of federated learning, since the training of the model takes place on the locally stored data on the node/device. Since the data resides on the device throughout, it is more secure by design
- **Heterogeneous:** There can be different types of devices in the federated learning network, hence there exist different types of processors and computing powers across these devices, making this a heterogenous training mechanism

EXHIBIT 2

The characteristics of federated learning

Source: Everest Group (2021)



A delineation with centralized learning

Federated learning and centralized learning (the traditional machine learning approach used for model training today) differ in the basic underlying principle that governs the process of training of ML models. Centralized learning methods follow a centralized learning regime, wherein all the training data is first aggregated on to a centralized server, which is then used to train and arrive at one version of the model.

EXHIBIT 3

Comparing centralized and federated learning

Source: Everest Group (2021)

Centralized learning	Federated learning
Trained on centralized data	Trained on distributed data
Data resides on the cloud or centralized server	Data resides at the various nodes in the network
Training takes place primarily in the cloud	Training happens primarily at the edge
Nodes/edge devices share local data	Nodes/edge devices share local version of the mode
Cannot operate on heterogenous data	Can operate on heterogenous data
Low user data privacy	High user data privacy

Benefits of federated learning

Everest Group take

Federated learning can deliver two levels of benefits over centralized learning, in the sense that it cannot only help overcome the shortcomings of centralized learning, but also provide additional benefits on top of it. It can help in situations where centralized learning falls short such as training in cases of data silos, data sensitivity, or heterogenous data.

Secondly, it has the potential of providing multi-faceted improvements over the traditional centralized learning mechanisms by providing lower communication and infrastructure costs, enabling offline predictions, and providing improved accuracy. But the most essential advantage that FL systems provide us with is the “privacy by design” principle.

Privacy is the key reason for federated learning adoption

Federated learning systems are designed in a way that the concept of privacy and data security is baked into its underlying principles. By definition, federated learning systems do not require data to be shared for the model training to take place. It allows the different entities to retain and store their private data. As the model is trained on the device, the data need not be moved to a centralized location. Since data is stored locally in a distributed fashion, it remains secure and private. By employing federated learning, enterprises have the option of not putting its users' data at risk. Additionally, since the data resides with the user, this gives the user complete autonomy and full control over its data. It lets the user define its own privacy policies and govern its data access rights.

Privacy is baked into federated learning as it operates on a "privacy by design" principle.

Additional benefits of federated learning

In addition to ensuring privacy and data security, federated learning offers multi-fold advantages over centralized learning in the form of lower communication costs, working across data silo, enabling offline predictions, and offering improved accuracy for the trained algorithms.

EXHIBIT 4

The benefits of federated learning

Source: Everest Group (2021)

Training across data silo

Training on heterogenous data

Lower communication costs

Enabling offline predictions

Improved accuracy

Reduced liability

- **Training across data silos:** Federated learning provides a good alternative to building a cross-enterprise and/or cross-domain ecosphere for use cases where data cannot be shared because of intellectual property rights, privacy protection, and data security concerns, and has to reside within the boundaries of the particular enterprise
- **Training on heterogenous data:** Data stored across different locations is mostly in varying formats, which needs to be standardized for centralized learning mechanisms. In the case of federated learning, since the learning is happening on the device, these data standardization activities become obsolete
- **Lower communication costs:** As the individual devices are not uploading the bulk of the data and the communication is restricted to just uploading of the local versions of the model and downloading the global version, communication costs go down significantly
- **Enabling offline predictions:** Prediction can work even when the devices are offline. So, there is no need for devices to have an internet connection. As long as a device can get input, the predictive models can be utilized to do their work
- **Improved accuracy:** Federated learning enables creation of smarter models via aggregation of many local models, and the different devices can learn from each other more frequently
- **Reduced liability:** Given that the data is not aggregated in a central location and no party or organization owns it, the liability involved with securing the data also gets reduced. This liability gets transferred to the individual nodes

Application of federated learning

Federated learning use cases

Even though federated learning is a relatively new concept, there is no dearth of potential application areas and use-cases that will significantly benefit from its application.

Improved prediction for diseases: In the current scenario, the models that individual medical institutions are developing for the prediction of a particular disease suffer from inaccuracy due to limited training data. Employing federated learning here, individual doctors can train the algorithm on their patients' data within their institution/hospital to develop a local version of the model. These local versions can be combined to create a global model whose accuracy will surpass the accuracy of the siloed models being used today.

Virtual keyboard prediction: The next-word prediction feature in our keyboards is a classic area of applications for federated learning. This can enable dynamic learning for the model as opposed to the other alternative where the model is trained on huge amounts of textual data. Based on an individual's data, the algorithm can train on these individual devices, and which will then be combined to create a more powerful, accurate, and continuously learning model for predicting the next word.

Credit risk management: Federated learning can be used to arrive at a more precise credit score by developing a model that leverages data from different financial institutions and banks, without these institutions having to share the personal and private information of individual customers. This will help in jointly generating a more comprehensive and reliable credit score for a customer without sharing her/his information.

A move toward "Edge AI"

In their quest to achieve enhanced optimization, elevated need for efficiency, transparency, and cost reduction, more and more enterprises are adopting a distributed computing architecture, which makes the edge extremely essential. Currently, we see most of the heavy lifting when it comes to data processing being done on the edge. In addition to this, we see the edge becoming a more central entity with storage, analytics, and now even training of AI models being moved to it.

Enterprises want intelligence to reside completely at the edge, not just the analytics and decision-making but even the training. This is to ensure more personalized decision-making, decreased latency, enabling offline predictions, and a more continuously trained dynamic model that will eventually provide enhanced accuracy and results.

Federated learning is best suited to empower the edge with total intelligence, with its distributed learning mechanism complementing the distributed architecture of an edge computing implementation.

Federated learning ecosystem

There exist multiple players in the market that are making significant investments for developing and scaling federated learning solutions. Given that the technology is in its nascent stages, this space is dominated by start-ups and open-source players. A few leading start-ups in this space are Owkin and Sherpa.ai, which have developed federated learning frameworks currently operational at a few enterprise locations, and are pioneering this technology.

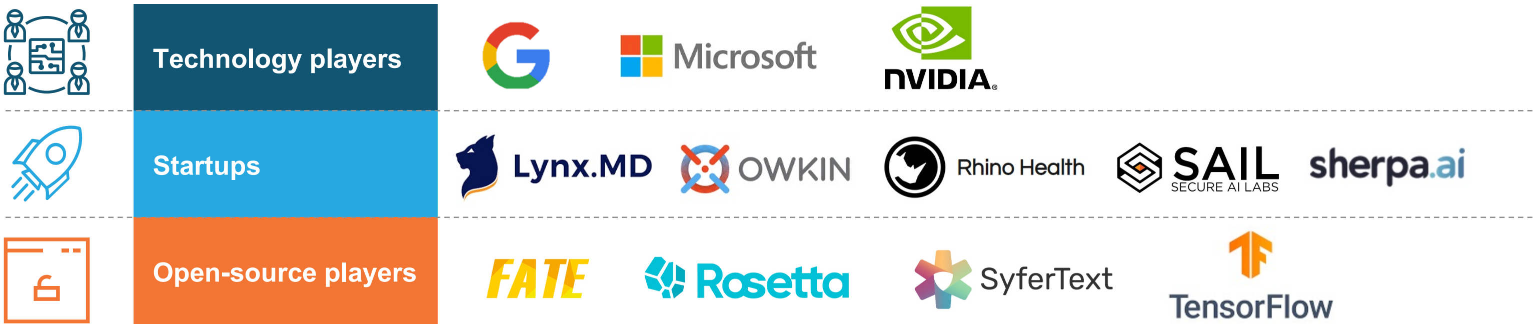
We also see some amount of activity in this space from leading technology players. In addition to Google, which is making the most contribution in this domain since introducing the concept back in 2017, we also see NVidia, Lenovo, and Microsoft making significant investments in this area.

Exhibit below indicates the various players developing solutions in the field of federated learning.

EXHIBIT 5

Federated learning ecosystem

Source: Everest Group (2021)



Technology players	Microsoft, NVIDIA
Startups	Lynx.MD, OWKIN, Rhino Health, SAIL SECURE AI LABS, sherpa.ai
Open-source players	FATE, Rosetta, SyferText, TensorFlow

Decision-making framework for federated learning

Federated learning does provide significant advantages over the traditional centralized learning mechanism, but the adoption of federated learning is extremely nuanced and needs to be well thought out before adoption.

Here, we are presenting a comprehensive framework that enterprises can use to assess the federated learning suitability for their particular use-case. The following framework highlights four key questions that enterprises should ask themselves to determine whether federated learning will help them reap significant benefits.

EXHIBIT 6

Federated learning decision making framework

Source: Everest Group (2021)

Attribute	Questions	Score
Data criticality	Does your enterprise deal with highly confidential/sensitive data?	Rate these questions on a scale of 1 to 5 based on its relevance to your use case (1 being the lowest and 5 being the highest)
Privacy requirement	Do your customers value their privacy above all else?	
Regulatory constraint	Is your industry highly regulated in terms of the guidelines defining the storage and processing of data?	
Data silo/diversity	Is your data present across different silos and is heterogeneous/diverse in nature?	
Final score (average out the scores for each question to arrive at the final score)		X

Score analysis	
0-1	Centralized learning might suffice to meet current business objectives
2-3	The business will reap significant benefits by adopting federated learning
≥4	Federated learning will become critical for business sustenance and will act as a key enabler for the enterprise business model

Future outlook

There is a lot of activity in this space among start-ups, niche vendors, or small technology players who are coming up with new use-cases that can benefit from federated learning as a methodology. However, the overall industry adoption of federated Learning remains low and limited to certain industries including healthcare & life sciences and banking & financial services. However, given the momentum perceived in the market, federated learning is headed for mainstream adoption.

This is also being accelerated by the fact that the world is becoming more privacy conscious, and people are increasingly hesitant to share their data. This could become an enormous problem for AI, which relies on humongous amounts of data for its operation and accuracy. The solution to this problem lies with federated learning, which as mentioned earlier ensures “privacy by design.”

As the ecosystem continues to evolve, new use cases emerge, data consciousness rises, and the guidelines on data privacy become more stringent, federated learning will emerge as a one-stop solution for privacy within machine learning.

Everest Group®
With you on the journey

Everest Group is a research firm focused on strategic IT, business services, engineering services, and sourcing. Our clients include leading global companies, service providers, and investors. Clients use our services to guide their journeys to achieve heightened operational and financial performance, accelerated value delivery, and high-impact business outcomes. Details and in-depth content are available at **www.everestgrp.com**.

For more information about Everest Group, please contact:

+1-214-451-3000
info@everestgrp.com

For more information about this topic please contact the author(s):

Arushi Pandey, Senior Analyst
arushi.pandey@everestgrp.com

Nisha Krishan, Senior Analyst
nisha.krishan@everestgrp.com

This document is for informational purposes only, and it is being provided "as is" and "as available" without any warranty of any kind, including any warranties of completeness, adequacy, or fitness for a particular purpose. Everest Group is not a legal or investment adviser; the contents of this document should not be construed as legal, tax, or investment advice. This document should not be used as a substitute for consultation with professional advisors, and Everest Group disclaims liability for any actions or decisions not to act that are taken as a result of any material in this publication.