Federated Learning & myDRE

Context

Federated learning is a machine learning approach that enables multiple parties to collaboratively train a model without sharing their raw data. Instead, each party trains a local model on its own data and then shares only the model's updates with a central server, which aggregates the updates to create a global model. This approach allows parties to benefit from each other's data while maintaining data privacy and security. Federated learning has emerged as a promising approach for training models in scenarios where data is distributed across multiple devices, such as smartphones or IoT devices, and cannot be easily centralized.

Federated learning faces several challenges, including:

1. **Heterogeneity**: Federated learning involves training models on data from multiple sources, which can be heterogeneous in terms of data format, quality, and distribution. This heterogeneity can make it challenging to develop models that generalize well to all sources.
2. **Communication and computation efficiency**: Federated learning requires frequent communication between parties, which can be costly in terms of bandwidth and energy consumption. Moreover, the local models trained by each party may have different sizes and complexities, which can make it challenging to efficiently aggregate the updates.
3. **Privacy and security**: Federated learning aims to protect the privacy of each party's data while enabling collaborative model training. However, privacy and security risks can arise if the updates are not properly secured or if malicious parties attempt to manipulate the training process.
4. **Model selection**: Federated learning often involves selecting a subset of available models to be trained collaboratively. However, selecting the right models can be challenging, particularly in scenarios where parties have different objectives or constraints.
5. **Regulatory compliance**: Federated learning involves the sharing of sensitive data, which may be subject to legal and regulatory requirements such as data protection and privacy laws. Ensuring compliance with these requirements can be challenging, particularly in cross-border collaborations.

Federated learning involves sharing sensitive data between multiple parties, which can create several risks. Some of the common risks associated with federated learning include:

1. **Privacy and security risks**: Federated learning involves sharing sensitive data, which can expose it to privacy and security risks. Malicious actors may try to intercept the data, steal it, or use it for unauthorized purposes.
2. **Data quality risks**: The data used in federated learning may be of varying quality, which can impact the accuracy of the trained model. Poor quality data can introduce biases, noise, or outliers that may degrade the model's performance.
3. **Regulatory risks**: Federated learning involves sharing data that may be subject to regulatory requirements, such as data protection and privacy laws. Failure to comply with these regulations can result in legal and financial penalties.
4. **Model performance risks**: Federated learning involves training models on data from multiple sources, which can lead to variations in the data distribution and quality. These variations can impact the model's performance and generalization ability.
5. **Coordination risks**: Federated learning involves coordinating multiple parties to train a model collaboratively. Coordination challenges, such as differences in objectives, data formats, and computing resources, can make it difficult to achieve effective collaboration.
6. **Algorithmic fairness risks**: Federated learning may inherit biases from the training data, which can perpetuate existing social, economic, or cultural biases. These biases can lead to unfair and discriminatory outcomes for certain groups or individuals.

It is important to address these risks by implementing appropriate technical, organizational, and legal measures to ensure the privacy, security, and fairness of the federated learning process.

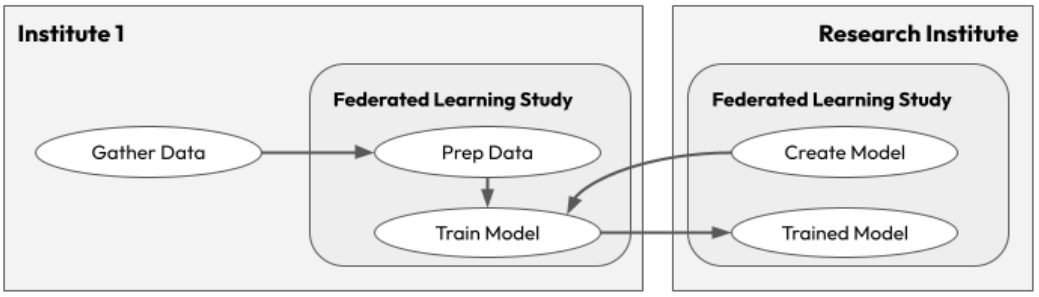
Federated Learning on myDRE

anDREa BV is currently engaged in discussions with organizations such as GO FAIR to explore potential solutions that can facilitate Federated Learning on myDRE. However, it is our belief that the basic functionalities of myDRE already provide adequate support for Federated Learning, utilizing the following mechanisms:

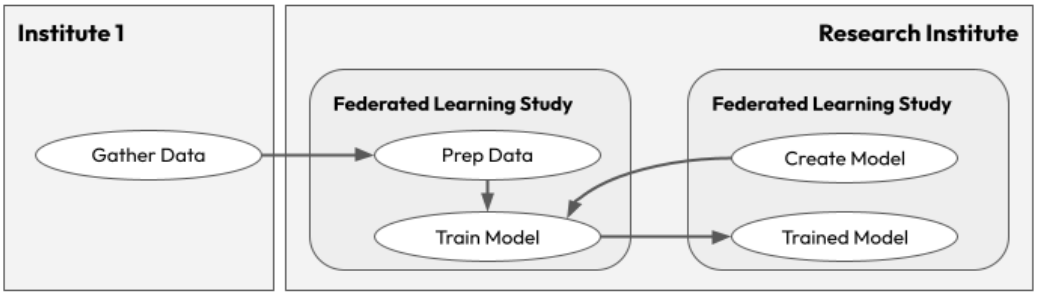
- A Workspace for creating and distributing the model, as well as receiving the trained model
- Workspace-to-Workspace communication for high-security data transfer
- Alternatively, domain whitelisting via platforms such as Github for less secure data transfer
- Workspaces for each collaborating institution that can receive and prepare data for training, receive the model via Workspace-to-Workspace or Github, and have the computational resources to train the model and export the trained model to the Research Institute or other institutions for further training.

If the collaborating institution has its own myDRE license, Scenario 1 can be utilized. Otherwise, Scenario 2 presents an alternative approach. In all cases, the data can be securely transferred to a trusted environment that allows for controlled collaboration, even with external parties. Ample storage and processing capacity is available via self-service, without the need for centralized IT support.

Scenario 1: Collaborating Institute has its own myDRE License



Scenario 2: Collaborating Institute has NOT its own myDRE License



Scenario 1: Collaboration Institute has a myDRE License

Advantages of Scenario 1:

- The collaborating institute has complete control over their data and can retract their collaboration at any time.
- Clear roles and access controls ensure that only authorized parties can access the data.
- If the training can be done by running one or more virtual machines, there are no restrictions, and easy and fast access to compute resources available on Microsoft Azure is available.
- The collaborating institute pays for all the storage and processing costs for their own data.
- Do please consult the legal departments for the following:
    - No data leaves the workspace, eliminating the need for a data transfer agreement.
    - A material transfer agreement may be needed for the trained model.

Disadvantages of Scenario 1:

- The collaborating institute needs a Microsoft Azure plan and a myDRE license, which may be expensive for participating in just one study.
- Workspace-to-Workspace transfer is not yet possible if the workspaces reside in different Microsoft Azure regions.

Scenario 2: Collaboration Institute does not have a myDRE License

Advantages of Scenario 2:

- The collaborating institute is unburdened with respect to storage and compute for the training
- Data can be transfered securely into a Workspace
- Clear roles and access controls ensure that only authorized parties can access the data.

Disadvantages of Scenario 2:

- Do please consult the legal departments for the following:
    - The research institute is technically responsible for the collaborating institute's data, which can be mitigated by making the accountable person from the collaborating institute.
    - A data processing agreement and data transfer agreement may be needed.
    - Liability and responsibility issues need to be addressed through legal agreements.
- The research institute pays for all the storage and processing costs.

Overall

Both scenarios offer secure and controlled collaboration environments for federated learning. We recommend discussing the specific advantages and disadvantages of each scenario with your legal and IT teams to determine the best approach for your research project.