# FLUID - A SECURE AI PLATFORM

## Introduction

The need for applying AI across a variety of industry use-cases is growing at an unprecedented pace world-over, however there are a few key challenges preventing this adoption within large and medium sized enterprises / businesses. At the top of this list of challenges are, the lack of quality AI programming talent, and the lack of clarity on provisioning data for such AI projects.

In order to solve for the lack of talent, enterprises are actively seeking out AI startups or service providers that are developing unique solutions for different types of use-cases. Additionally, rates of innovation and adoption of AI in businesses, academia and government are increasing exponentially, as they are beginning to learn the immense value that can be derived from AI. The two most important requisites for AI are data and the AI algorithm itself. The costs and the efforts to collect new data is limited to organizations that can afford to run data collection campaigns. Therefore development of AI is limited to data that is easily obtainable. This easily obtainable data is usually what the research institutes have chosen to publish online, or what the people and the organizations can consider to safely share.

However, massive amounts of valuable data remains hidden, this data that holds the secrets to curing diseases, achieving better governance, better distribution of resources, and understanding the nature of our species. This data remains hidden because there are no incentives for the data owners to share it, and sometimes this data is sensitive in nature thus limiting new developments in AI that can directly benefit the people.

The traditional AI paradigm does not accommodate this need for privacy of the digital assets and trust between multiple parties, which are paramount drivers of new collaborations. The lack of a framework that allows for effective collaboration between such parties, while completely preserving the privacy and ownership of their digital assets, leads to a fragmentary use of economies of scale.

Eder Labs aims at redesigning the AI ecosystem to enable data providers to have complete ownership of their data, therefore enabling organizations to extract intelligence from people's / users' data without infringing their privacy and enabling research organizations to create valuable AI from new unearthed data. With capabilities to deploy the AI model at the edge, we can now restrict the data movement to cloud for AI inference. This not only reduces the infrastructure needed to handle streams of real time load, but also reduces the security risks by cutting down the movement of data.

As the future gets more connected digitally, there will be more services to support the digital age. This will generate more data from myriad sources which can help optimize many of our existing capabilities and unravel new innovations.

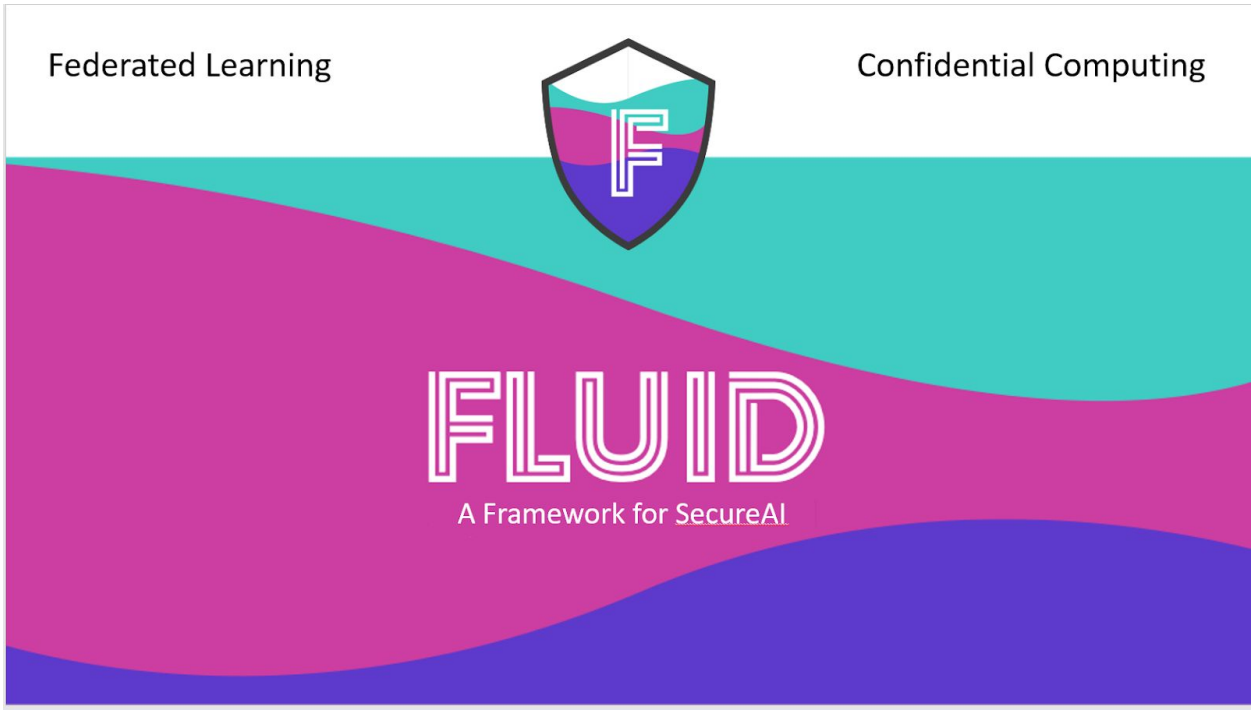
Eder Labs' believes, guaranteeing data privacy and providing proof of confidentiality of the same will enable and further incentivize organizations to trust and innovate together. In the overall market for AI solutions, across use-cases for all industries, the Eder Labs approach combines two key, recent technological developments: federated learning and confidential computing. This approach is further bolstered by the evolving data-protection paradigm world-over.

# Why Fluid?

According to a survey conducted by IAB, more organizations are transitioning into being 'data-centric' by using audience data for better marketing and strategy. The major challenge they are facing is the lack of talent in the domain of analysis and AI. Hence, the data owners would want to engage with AI service providers to achieve their data-centric goals. It would involve sharing sensitive data in some form with the service provider. Along with privacy concerns, the mobility of data also becomes a roadblock for a fruitful engagement when compliance laws like GDPR and upcoming privacy frameworks are factored in.

In the Secure Machine Learning landscape, some of the most popular algorithms that ensure privacy-preserving machine learning are Homomorphic Encryption, Secure Multi-party Computation (MPC), and Differential Privacy. They do not guarantee complete privacy by themselves but when implemented in combination at different places of the Machine Learning pipeline. Hence, the implementation of such algorithms in a manner that they are agnostic to the ML pipeline is challenging. Finally, incorporating them is computationally expensive, and the accuracy often suffers.

The motivation behind Fluid is to not only ensure data privacy and secure computations but also to ensure there are minimal changes to the ML workflow.




# Fluid is an end-to-end platform for machine learning with complete data privacy

Fluid is a platform that enables multiple non-trusting organizations / parties to collaborate together and create unique AI from sensitive data while maintaining absolute data privacy. Instead of sending data from multiple clients to the central AI model for computation and training, the AI model is securely distributed to the multiple client's infrastructure where the model performs both the training and prediction inference locally without the need to move data from the provisioned secure run-time environment.

Eder Labs uses federated learning with trusted execution environments to ensure complete data and model privacy. Federated Learning ensures that the data remains encrypted and in control of the owners while the learnings are shared to the global model. Trusted Execution Environments ensure that the computation is done on encrypted data inside the black-box, as well as provide proof of confidentiality each time the computation is invoked.

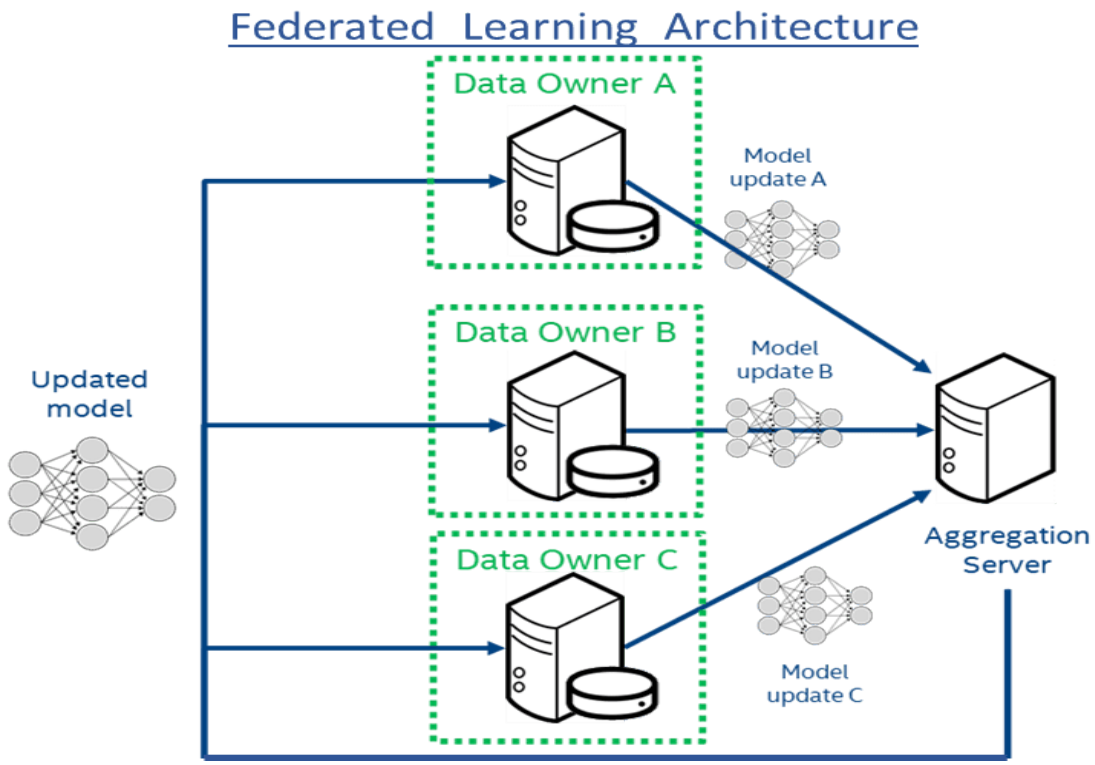
All data contributions to the central intelligence as well as access to the intelligence are measured and the ownership of the model is distributed with respect to the contributions made by the participating organizations.

# Federated learning

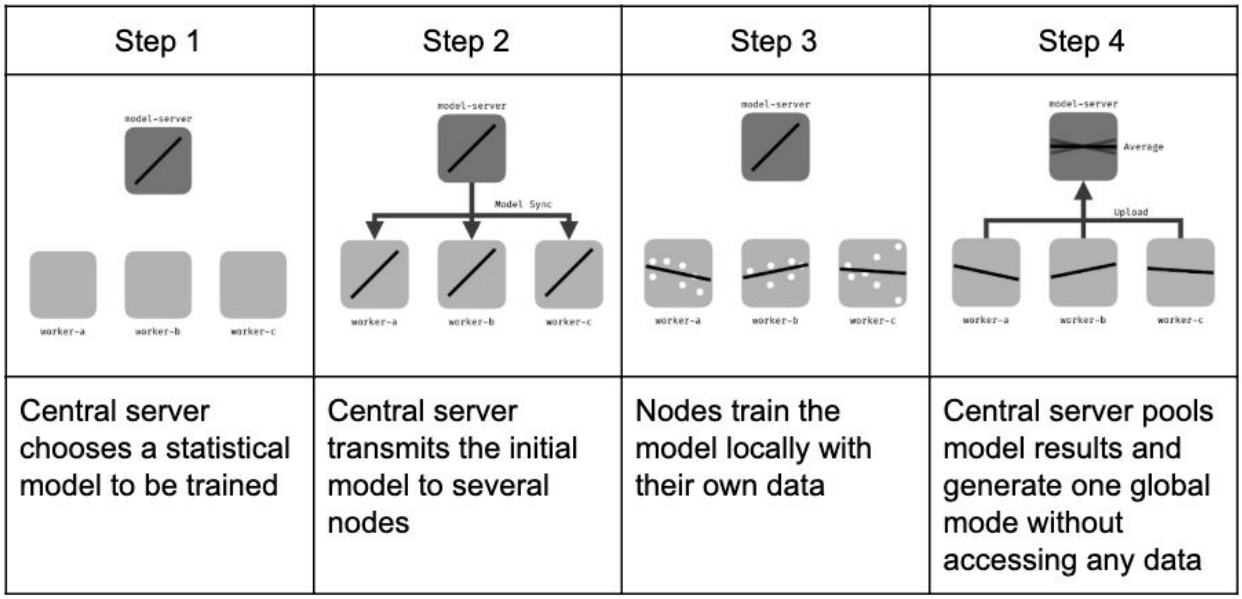
Federated Learning is akin to Distributed learning, where instead of accumulating data from all sources to a central server for training the global model, the global model is distributed to the multiple edge devices capable of performing learning tasks. The learning from the multiple different models is then accumulated to a central server, where they are aggregated into a globally updated model. This updated global model is then sent back to all edge nodes. This cycle is then repeated.

If the edge devices are *general purpose* and capable of model inferencing, they can also be programmed to do machine learning locally, albeit at a lower pace. But, since the data collected by one federated learning node will be smaller compared to the entire corpus, the learning will also take much lesser time to update the model locally.

*The following diagram represents the workflow for federated aggregation with 3 different data owners training a local model and sharing the model updates with a central secure aggregation server :*



*The following diagram depicts the various steps involved in a federated learning setting with 3 different data owners training a federated ML model :*



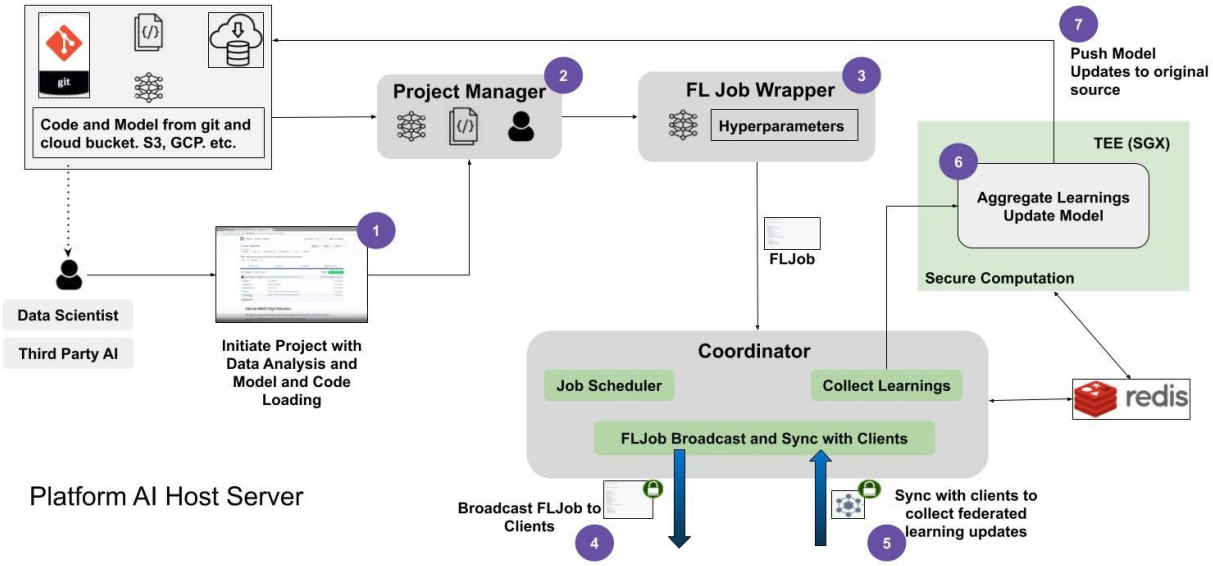

*Step 1: involves the initialization of the global model*

*Step 2: deploys the global model to the client nodes*

*Step 3: indicates the learning curves for the local models at each of the three data owners after training on their respective data points securely*

*Step 4: depicts the overall learning curve for the aggregated model*

*The following flow diagram depicts the userflow for the AI vendor*:



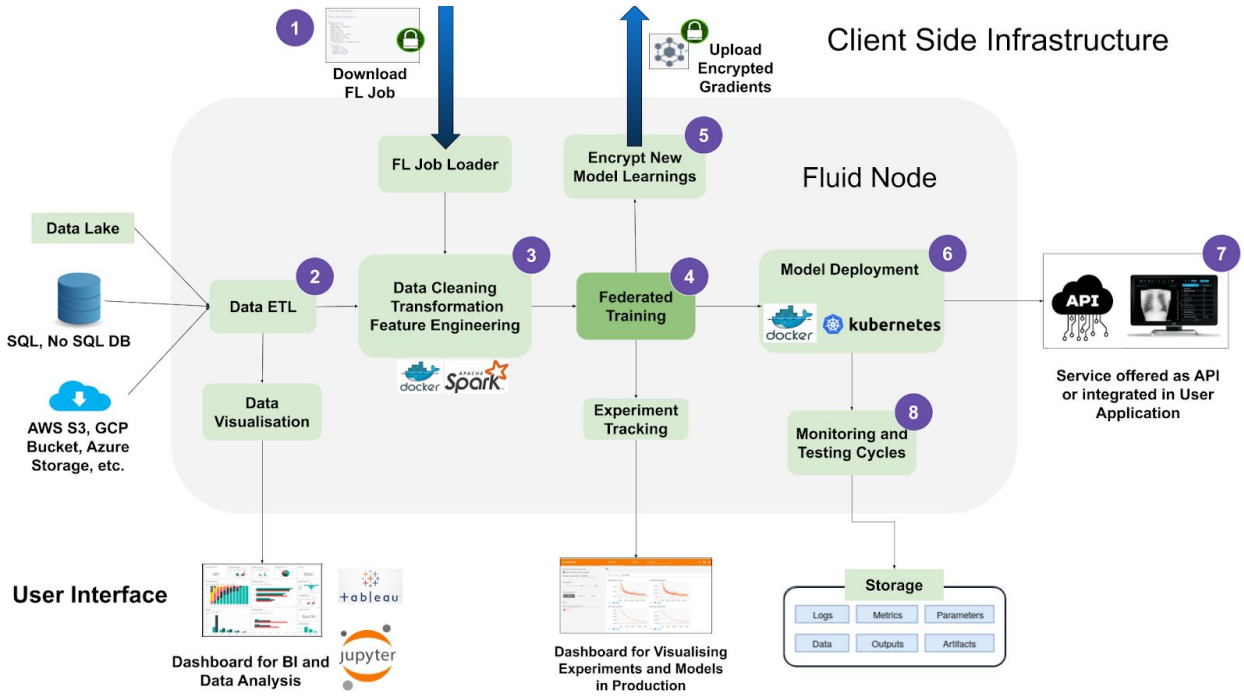

**Step 1 - 3** : Initiate a new project by linking the code and the model repositories, load it into the project manager and extract a Federated Learning Job (FLJob) from it. FLjob is the set of instructions needed to run

**Step 4** : The Coordinator is responsible for managing connectivity with multiple active clients. The working of clients is discussed in the diagram below. The coordinator in this step, will broadcast the FLJob to all it's clients.

**Step 5** : The clients after training will send encrypted learning (model gradients) to the coordinator, the coordinator will receive these learnings and store them for future processing.

**Step 6 - 7**: Once enough updates are collected to fulfill the criterion. These encrypted learnings will be first loaded in an intiated TEE enclave whose integrity will be verified each time it is loaded. The encrypted learnings will then be aggregated inside the Trusted Execution Environment.
The newer model update will go through validation tests and checks, and if the results show improvement the model will be updated on the source.

*The following flow diagram highlights the processes followed at the data owner's end to setup and start the federated learning project :*



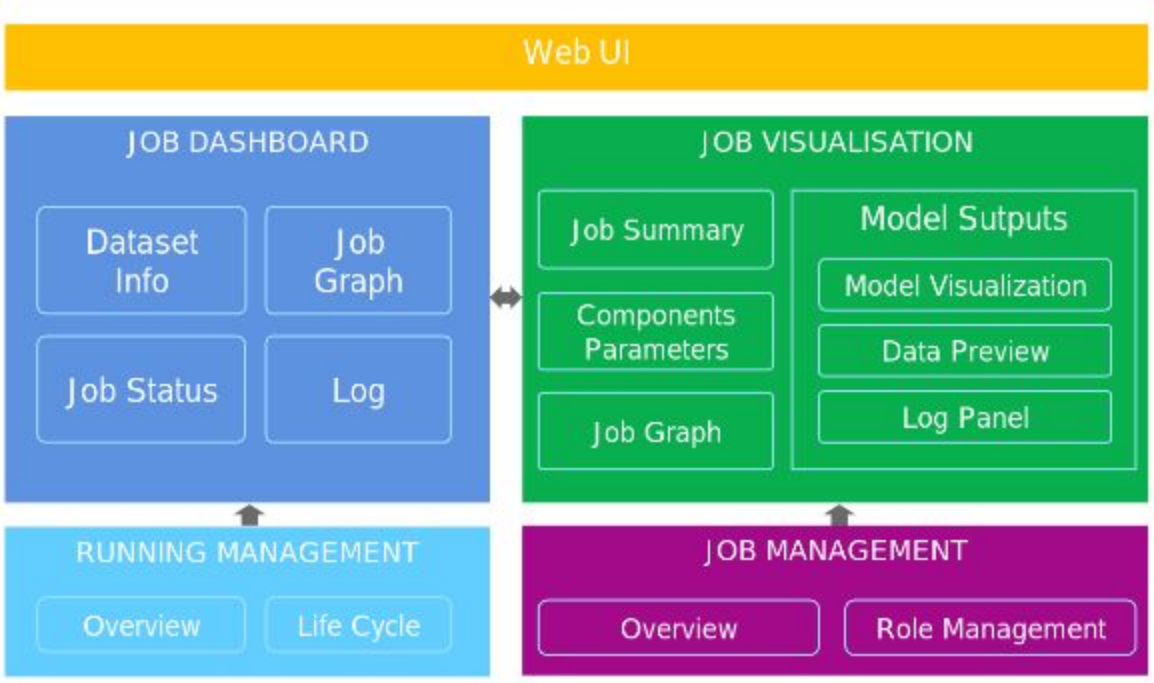

The client's infrastructure will first be setup and fluid node will be connected to the relevant DB for carrying out the process further

**Step 1 - 3** : The client will authenticate with the host coordinator and download FLJob. The instructions will be extracted for Data ETL, pre-processing, feature engineering and other data transformations.

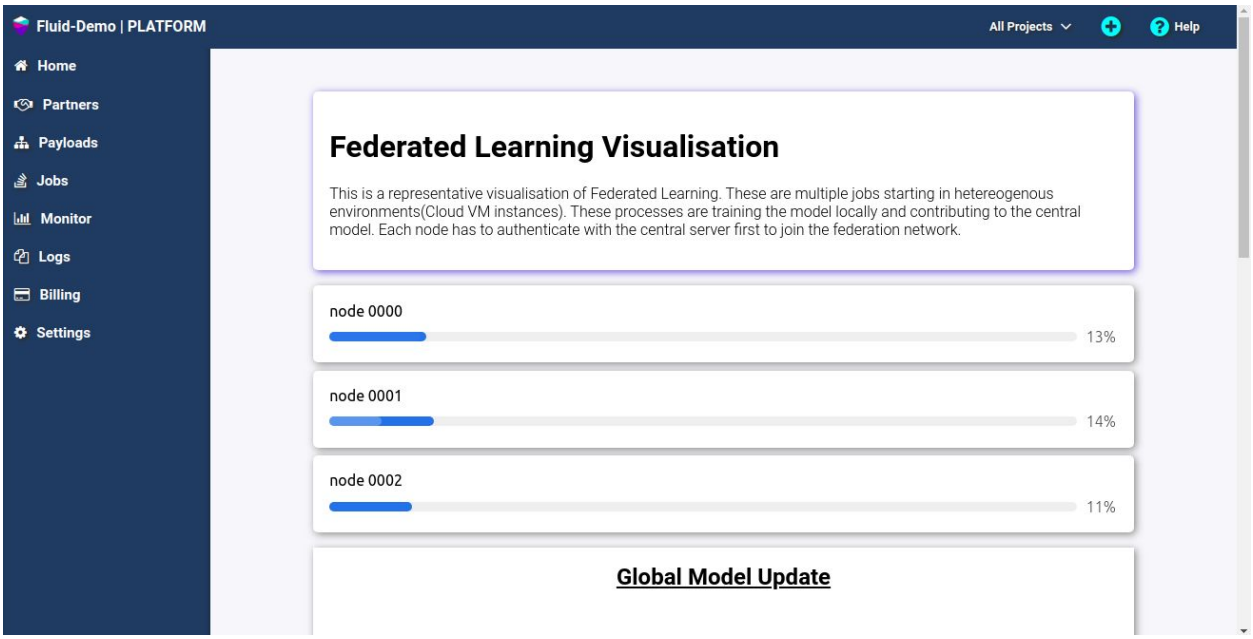
**Step 4 - 5 :** Other training hyper-parameters will be loaded from the FLJob and the federated training session will start. The training will happen locally in the client's environment, and the learnings from the new data will be encrypted and sent back to the coordinator.

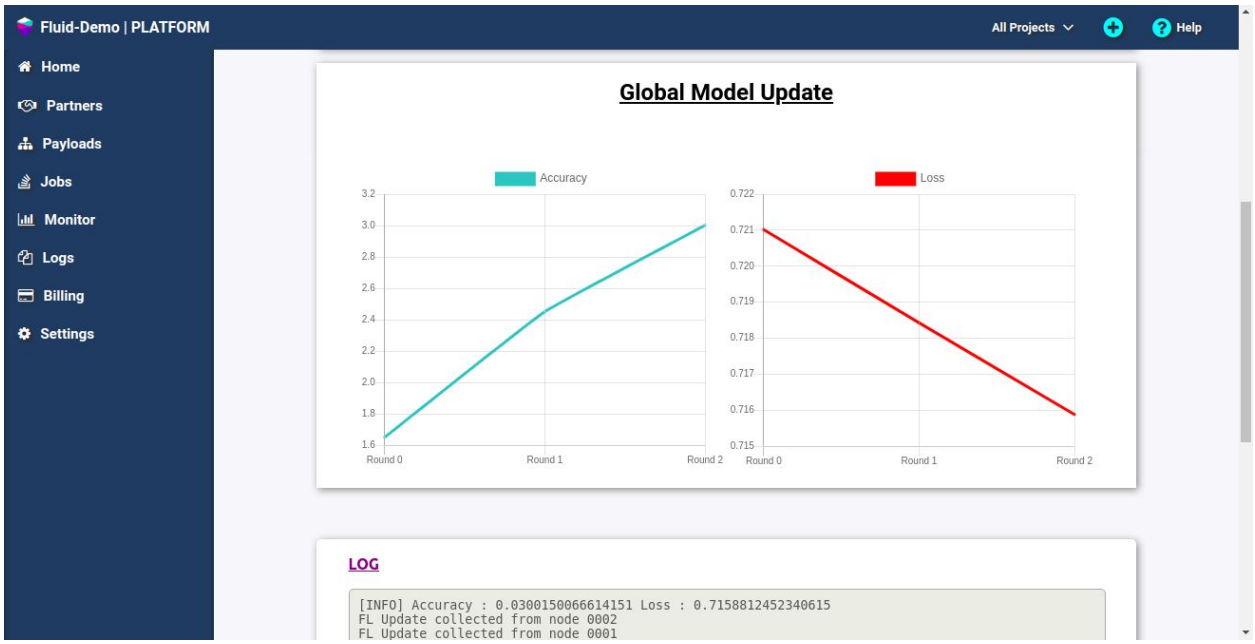
**Step 6 - 8** : The latest available model update to the client will then be loaded into the deployment environment, with procedures for complete model lifecycle management and logging the post deployment metrics for continuous active learning.

*The architecture of the UI dashboard*:




*The following are screens from Fluid visualizing the dashboard for the federated learning setting*




Fluid-Demo | PLATFORM
All Projects
Help
Home
Partners
Payloads
Jobs
Monitor
Logs
Billing
Settings
Global Model Update
Accuracy
Loss
3.2
3.0
2.8
2.6
2.4
2.2
2.0
1.8
1.6
0.722
0.721
0.720
0.719
0.718
0.717
0.716
0.715
Round 0
Round 1
Round 2
LOG
[INFO] Accuracy : 0.0300150066614151 Loss : 0.7158812452340615
FL Update collected from node 0002
FL Update collected from node 0001

# Confidential computing

To enable secure collaboration between the data owners and the AI service providers, many aspects of cryptography need to be baked into the existing AI software paradigm. Such a secure framework can be achieved through the use of cryptographic methods such as homomorphic encryption, secure multiparty computation or hardware based trusted/isolated execution environments, where the data and the AI algorithm remain secure and private throughout the process of execution.

While homomorphic encryption and secure multiparty computation (MPC) can be used to keep both data and the AI models private, they incur large computational costs and lack efficiency, especially for more complex AI models containing millions of parameters. In contrast, trusted execution environments provide a more feasible way to achieve the same as they are several orders of magnitude faster and support general-purpose computation i.e. not just arithmetic operations as in the case of MPC.

Fluid uses a runtime that enables privacy preserving AI through the use of trusted execution environments (TEE) in the cloud & hybrid cloud platforms. Fluid trusted currently supports Intel SGX, which uses the TEE in the Intel processors. Fluid secures both the data and the model at the runtime using the TEE and assures security to the involved parties through the use of remote attestation.

Attestation is the process of demonstrating that a software component that it is running on top of a trusted hardware platform with legitimate code and data. Attestation can be both local and remote. Eder Labs uses Intel's Attestation Service to attest a service provider's application to ensure it is running on a verified and trusted machine.

### Trusted Execution Environments

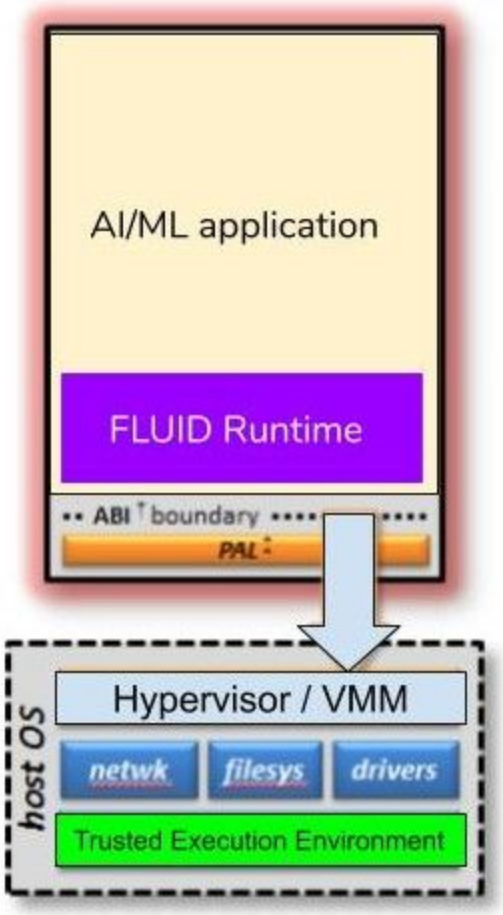
A Trusted Execution Environment (TEE) is an environment in the processor for executing computational logic, in which those executing the code can have high levels of trust in the environment which is separate from the main operating system and is not accessible from host OS's BIOS or hypervisor once initialized. It ensures that data is stored, processed and protected in a completely secure manner.

### Intel SGX

Intel's Software Guard Extensions (SGX) is a set of extensions to the Intel architecture that aims to provide integrity and confidentiality guarantees for secure computation by creating enclaves* Enclaves are stored in a hardware guarded memory called Enclave Page Cache (EPC), which is

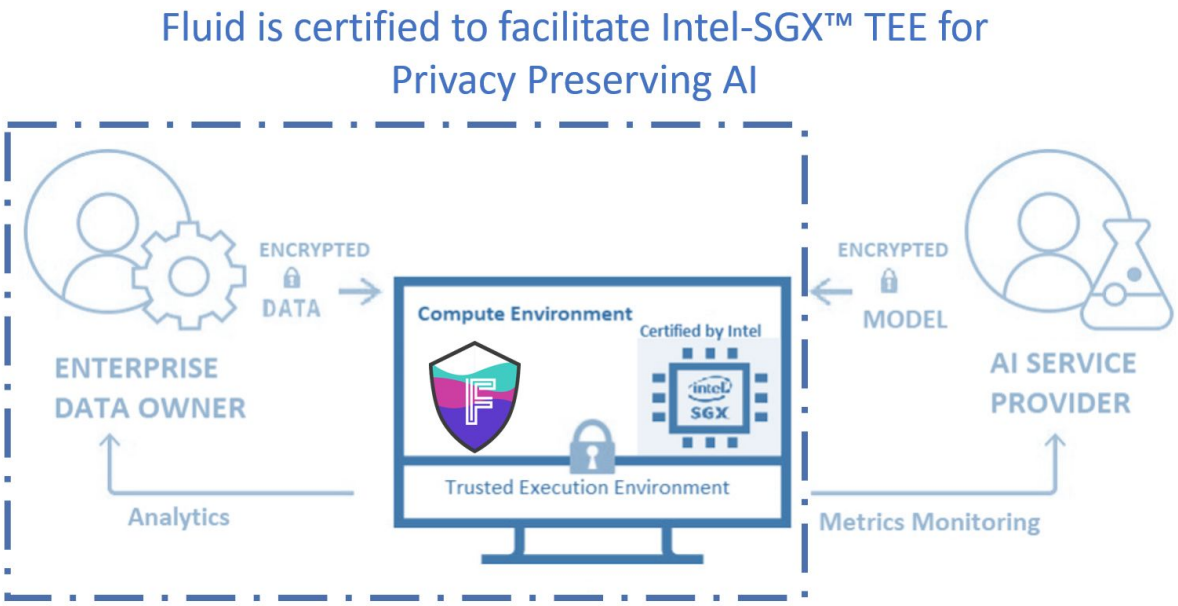
currently limited to 128MB with only 90 MB for the application. They are designed to execute programs and handle secrets in the trusted/isolated execution environment without exposing data to any malicious agent, even if they may have privileged access to the machine, such as through BIOS or Hypervisor. SGX enables this trust between the engaging parties through the use of remote attestation.

**Fluid Runtime Stack**



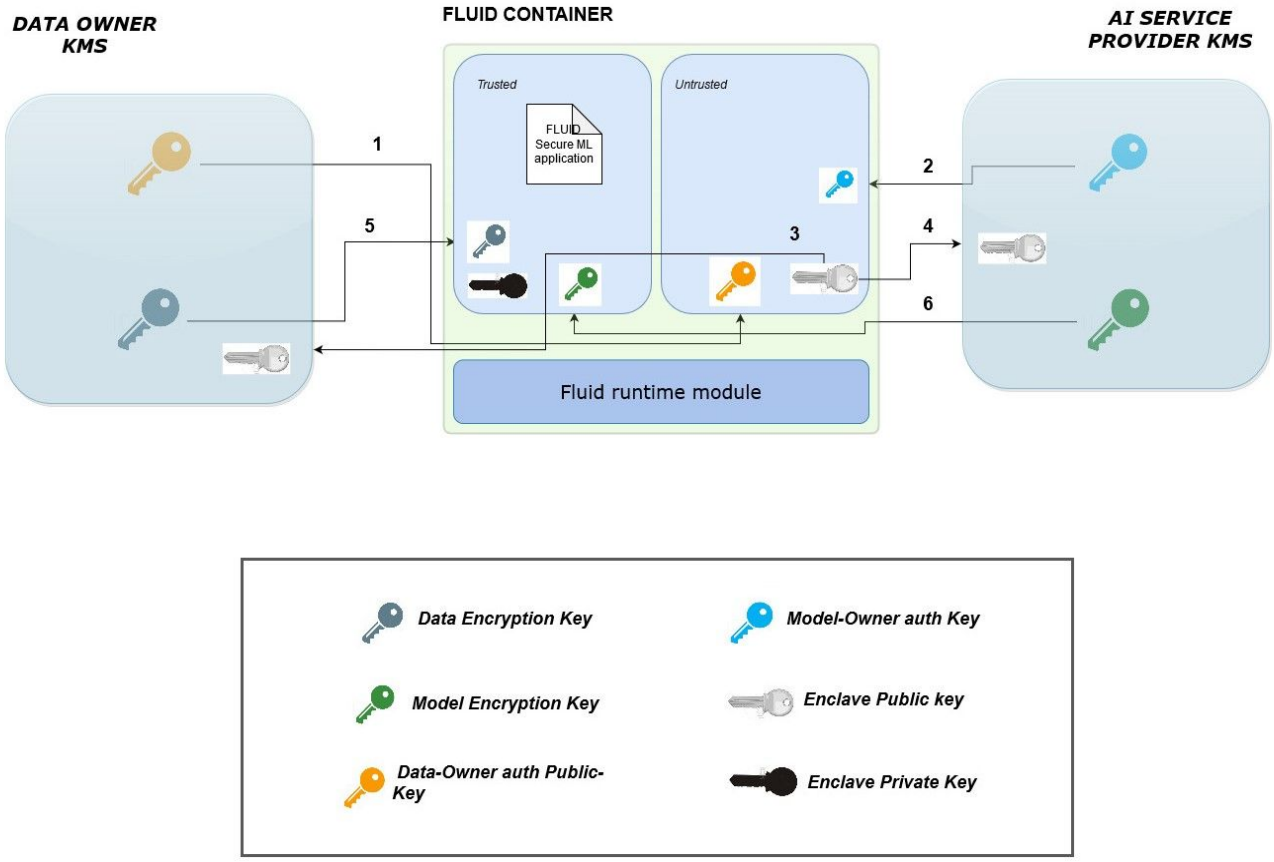

TEE compatible Fluid-runtime is deployed with the microservices running the intended application of the AI service provider. Fluid uses a picoprocess based sandboxing technique to package and deploy AI applications securely at scale.

## Encrypted ML Workflow diagram:




Fluid provides the engaging parties with simple APIs to secure all parts of the AI pipeline.

The following diagram indicates the detailed security/crypto workflow for Fluid for a sample AI engagement -




1. Data owner's public key is transferred into the fluid key manager once the project is initiated. This key will be later used to confidentially send messages to the data owner.

2. AI service provider's public key is similarly transferred to fluid key manager.

3. The initialized enclave once attested in the fluid platform, is then provisioned with key pair for secure communication with the registered parties during the engagement. The enclave's public key is encrypted with the shared data-owner auth key and sent to the data-owner.

4. Similarly, the enclave's public key is encrypted with AI service provider's key and shared with the AI owner.

5. The symmetric Data-encryption-key is encrypted with the enclave's public key and sent to the enclave.

6. Similarly, the AI owner transfers the AI model encryption key securely to the enclave to securely process the data.

# Fluid Attestation

Fluid uses its own secure attestation service based on Intel's SGX Data Center Attestation primitives, which can be used to securely manage and verify enclaves in environments where internet services are not accessible at workload runtime. Further it allows entities who are risk-averse in outsourcing trust decisions to other 3rdparties.

Fluid container's are deployed with launch enclaves which help authenticate and run different enclaves in either the data-owner's premises or the cloud. The attestation process occurs before provisioning any secrets to the fluid application enclave.

**Secure ML workload benchmarks:**

The following are the benchmarks comparing training times of various standard deep learning models when run with SGX using fluid runtime.

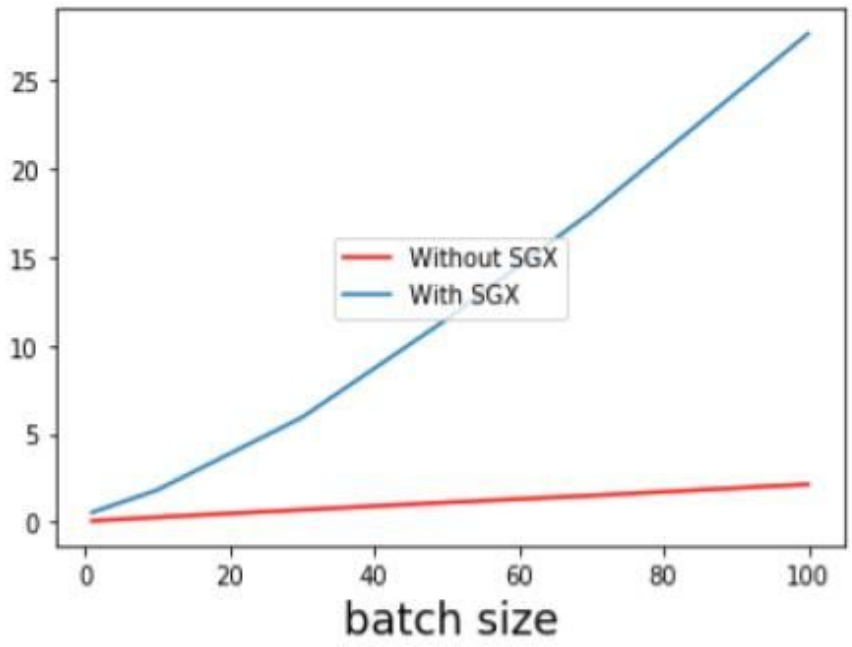
The models were trained on different training batch sizes of the IMAGENET dataset. The models benchmarked here are Mobilenet v2, Densenet and ResNet.

Y-axis: Time in seconds
X-axis: No of epochs

**Mobilenet**:

MobileNetV2 models are faster for the same accuracy across the entire latency spectrum.It is a very effective feature extractor for object detection and segmentation.

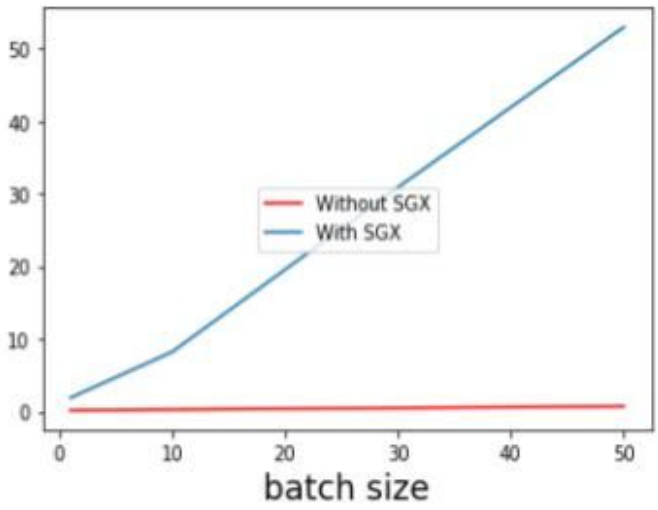
No of parameters : 5.1M
No of layers : 12




**Densenet** :

Dense Convolutional Network (DenseNet), connects each layer to every other layer in a feed-forward fashion. DenseNets have several compelling advantages: they alleviate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters.

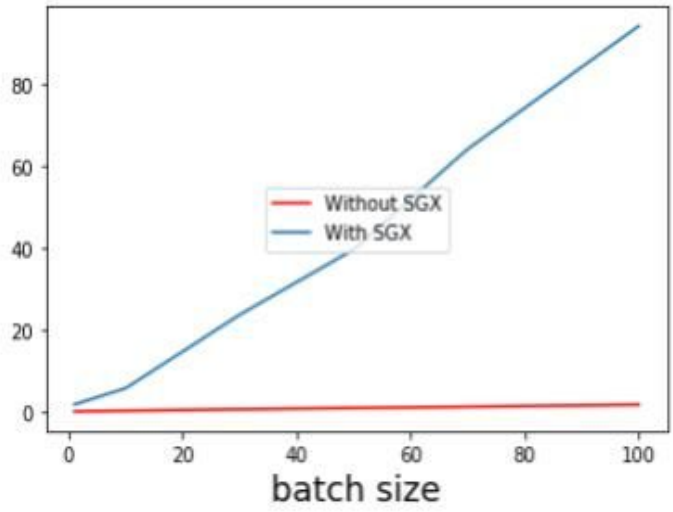
No of parameters : 20 Million
No of layers: 17




**Resnet** :

Residual Networks (ResNet in short) consists of multiple subsequent residual modules, which are the basic building block of ResNet architecture.

No of parameters: 1.7 million
Depth: 110 layers



We have benchmarked fluid runtime on an Intel(R) Xeon(R) CPU E3-1240 v5 @ 3.50GHz with 8 cores. There is a considerable overhead in training times as no optimizations were made to the model and constraints in memory size during runtime. We are working on different cross-compatible optimizers to make ML deployments using fluid runtime lighter.

## Confidential computing consortium

Established in 2019, the Confidential Computing Consortium brings together hardware vendors, cloud providers, developers, open source experts and academics to accelerate the confidential computing market; influence technical and regulatory standards; build open source tools that provide the right environment for TEE development' and host industry outreach and education initiatives. Its aims to address computational trust and security for data in use, enabling encrypted data to be processed in memory without exposing it to the rest of the system, reducing exposure to sensitive data and providing greater control and transparency for users.

General Members


Confidential Computing Consortium is a Linux Foundation project and community dedicated to defining and accelerating the adoption of confidential computing, with founding premiere members Alibaba, Arm, Google Cloud, Huawei, Intel, Microsoft and Red Hat. General members include Baidu, ByteDance, decentriq, Fortanix, Kindite, Oasis Labs, Swisscom, Tencent and VMware.

The organization aims to address data in use, enabling encrypted data to be processed in memory without exposing it to the rest of the system, reducing exposure to sensitive data and providing greater control and transparency for users. This is among the very first industry-wide initiatives to address data in use, as current security approaches largely focus on data at rest or data in transit. The focus of the Confidential Computing Consortium is especially important as companies move more of their workloads to span multiple environments, from on premises to public cloud and to the edge.

# Markets and Corresponding use cases

# Group Intelligence over same demographics across different financial organisations

In any financial ecosystem, multiple businesses/banks serve the same set of user demographics. These financial institutes run multiple AI models on fraud detection, risk assessment, credit scoring, etc. Since, they serve the same demographics in a geography it makes sense to form an alliance between these institutions so they can share their intelligence to form a central intelligence that will serve all institutes. To implement this today, these institutes will have to share their user's data, which is neither allowed nor encouraged. With Fluid's secure federated network, an alliance can be formed between these institutes to train a single model as well as use this model securely without exposing data anyhow. And each organization can claim ownership of the model on the basis of how much their data contribution improved the system overall.

# Hyperpersonalisation over same demographics across different organisations

An organization's AI model is as good as the data they have, and if the data doesn't tell complete picture the AI model does not perform. While a different organization will have the data that will complete the picture, and enhance the already existing model to much higher accuracy.

A credit rating service, or an Insurance claims processing service can serve their customers significantly better if they have access to the granular expenditure data of their user, while payment integration platforms hold exactly this data on the same user. However, even if these two parties wish to engage in alliance, they cannot exchange data due to the extremely sensitive nature of the data. Fluid can solve this problem, without any of these organizations having to move their data from their servers.

**Healthcare**

1. Genomics research models can access private genomics data across multiple healthcare institutions without moving / exposing such data.
2. Patient similarity learning.
3. Patient representation learning.
4. Predicting future hospitalizations- using EHR dataspread among various data sources/agent

5. Predicting mortality
6. ICU stay time

**Fintech**

1. Anti Money Laundering
2. Group-intelligence for fraud detection
3. ML based claims processing
4. Secure Collaborative Credit Scoring

**Industrial / Manufacturing**

1. Federated failure prediction / predictive maintenance
2. Energy management in power / manufacturing plants across geographies

**Pharma**

1. Drug discovery
2. Dynamic price negotiations without revealing pricing attributes and related IP

**Retail**

1. Dynamic Pricing intelligence across multiple store locations & online transactions
2. Smarter recommendation systems for online shoppers in different geographies

**Telecom**

1. Maintaining Quality of Service (QoS) across network nodes without having to pool data from sensors, devices, gateways to a central location
2. Hardware fault prediction across the 5G network

**Martech**

Unified customer intelligence across multiple brands is desired, however, customer intelligence could be limited to a single brand's echo chamber, mostly. Customers have diverse digital consumption patterns, and their personality on social-media is usually a projected version of their true self. In addition to all this, data / IP privacy is the new paradigm governing applied AI. Can adjacent brands

targeting / serving the same customer demographic share customer intelligence with one another while ensuring complete data privacy?

Imagine the benefits to customers from cross pollinating intelligence across companies and brands! Hyper-personalization with data privacy.

# Roadmap

2019 marked the beginning of proactive attention being given to two clear new categories in the AI domain, Confidential Computing and Federated Learning. The evolving importance of privacy in the world of data and AI is going to require the right combinations of technologies to preserve privacy in AI engagements.

Take for example, the IoT domain, with projected billions of devices coming online in this decade and over 44ZB of data expected to flow from them. This Internet of the near future, involving a variety of intercommunicating devices and systems running AI across one another, requires a secure and efficient way to track interactions, transactions, and activities of every ‘thing’ in the network. Data sovereignty and a trust framework across multiple parties / stakeholders is going to be a primary requirement with privacy-by-design becoming a fundamental requirement for all use-cases, across industry domains.

Such an extended framework will lead to a democratization of revenue potential across different types of data and AI stakeholders. Imagine such a future where edge eco-system players can discover each other, use one another’s data or applied AI services, including the possibility of micro-transaction settlements, without depending on a centralized player; where owners can control their identity and reputation. Imagine that all this is possible with the ability to leverage verifiable computation mechanisms to enforce data confidentiality and privacy at scale while provisioning regulatory and audit controls.

Eder Labs will be building and integrating additional technologies to ensure usecases for specific industries can be addressed in the most simplified manner, with limited change-management intervention.