

UNIVERSIDADE FEDERAL FLUMINENSE

JANAINA BORGES DE SOUSA

**Observabilidade orientada por dados de
proveniência para workflows de Aprendizado
Federado**

NITERÓI

2026

JANAINA BORGES DE SOUSA

**Observabilidade orientada por dados de
proveniência para workflows de Aprendizado
Federado**

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Orientador:

Prof. D. Sc. Marcos de Oliveira Lage Ferreira

Coorientador:

Prof. D. Sc. Daniel Cardoso Moraes de Oliveira

NITERÓI

2026

Ficha catalográfica automática - SDC/BCPV
Gerada com informações fornecidas pelo autor

D278o De Sousa, Janaina Borges
Observabilidade orientada por dados de proveniência para
workflows de Aprendizado Federado / Janaina Borges De Sousa. -
2026.
67 f.

Orientador: Marcos de Oliveira Lage Ferreira.
Coorientador: Daniel Cardoso Moraes de Oliveira.
Dissertação (mestrado)-Universidade Federal Fluminense,
Instituto de Computação, Niterói, 2026.

1. Aprendizado Federado. 2. Dados de Proveniência. 3.
Análise Visual. 4. Sistemas Distribuídos. 5. Produção
intelectual. I. Ferreira, Marcos de Oliveira Lage, orientador.
II. Oliveira, Daniel Cardoso Moraes de, coorientador. III.
Universidade Federal Fluminense. Instituto de Computação.
IV. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

JANAINA BORGES DE SOUSA

Observabilidade orientada por dados de proveniência para workflows de Aprendizado
Federado

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Aprovada em agosto de 2026.

BANCA EXAMINADORA

Prof. Marcos de Oliveira Lage Ferreira - Orientador, UFF

Prof. Daniel Cardoso Moraes de Oliveira - Co-orientador, UFF

Prof. Maria Cristina Silva Boeres, UFF

Prof. João Lucas Rulff da Costa, FGV

Niterói

2026

Dedico este trabalho, primeiramente, a Deus, por ter me guiado com saúde e fé até a conclusão da pesquisa. À minha filha Helena, por motivar cada passo da minha vida. Aos meus padrinhos e minha mãe por me apoiarem no caminho que eu decidi seguir.

Agradecimentos

Agradeço, primeiramente, a Deus, pela sabedoria e pela resiliência que me permitiram realizar este trabalho do começo ao fim.

Agradeço à minha filha, Helena, pela paciência e pelo apoio durante toda esta jornada acadêmica.

Agradeço aos meus pais, irmãos e padrinhos, que fizeram parte da minha formação e do meu desenvolvimento como indivíduo.

Agradeço aos meus amigos Paulo Henrique e Guilherme Odin, que fizeram por mim aquilo que eu precisava fazer e não conseguia, pois estava totalmente focada na execução deste trabalho. Durante o mestrado, precisei me mudar de casa, e vocês fizeram tudo acontecer como se fosse a casa de vocês. Jamais esquecerei esse gesto.

Agradeço aos amigos que conquistei ao longo dos anos como aluna da UFF e que estiveram comigo desempenhando um papel muito maior do que o de parceiros acadêmicos. Vocês foram essenciais nesta trajetória.

Às minhas amigas do pole dance, que representaram um espaço de acolhimento e terapia durante esse processo. Obrigada por ouvirem meus desabafos e por proporcionarem momentos leves e descontraídos.

Deixo também meu agradecimento aos meus amigos mais próximos, que compartilharam comigo tantos momentos ao longo dessa caminhada até chegarmos ao final.

Aos meus orientadores, Marcos e Daniel, por todo o apoio, ensinamentos e paciência durante o desenvolvimento deste trabalho.

Agradeço, ainda, a Wesley e Camila, por todo o suporte com levantamento dos requisitos que foram cruciais para a realização desta pesquisa.

Resumo

Federated Learning (FL) permite o treinamento colaborativo de modelos de Machine Learning (ML) entre múltiplos clientes distribuídos, preservando a localidade dos dados por meio de um workflow bem definido. Contudo, a natureza descentralizada da execução de FL dificulta a observabilidade desses workflows e, por consequência, o diagnóstico de desempenho, a interpretação dos resultados de treinamento e análises correlatas. Soluções de observabilidade existentes costumam ser dependentes de framework, baseando-se em logs de execução ou em telemetria proprietária que restringem a interoperabilidade e a análise entre diferentes plataformas. Defendemos que a proveniência constitui uma espinha dorsal para a observabilidade em FL, ao capturar a semântica de execução necessária para consulta, análise e exploração. Apresentamos **FedObserve**, uma abordagem de observabilidade orientada por proveniência que integra captura de proveniência em tempo de execução, um modelo de proveniência unificado e análises visuais interativas. A **FedObserve** registra dados de proveniência a nível de workflow, rodada e cliente, independentemente do framework subjacente de FL, permitindo que execuções heterogêneas sejam analisadas por meio de uma representação comum. As capacidades de visualização de proveniência suportam análises em diferentes granularidades das trilhas de execução, participação de clientes, evolução do modelo, convergência e desempenho, possibilitando que profissionais inspecionem e diagnostiquem iterativamente o comportamento do FL. Avaliamos a **FedObserve** por meio de um workflow de FL para detecção de outliers em grandes conjuntos de dados astronômicos. Um estudo com especialistas em FL demonstra que a **FedObserve** facilita a identificação de gargalos de execução, padrões de participação de clientes, heterogeneidade espacial e comportamento de convergência. Esses resultados indicam que a observabilidade orientada por proveniência aprimora a rastreabilidade, a interpretabilidade e o diagnóstico de workflows de FL, além de viabilizar o monitoramento independente de framework.

Palavras-chave: Aprendizado Federado, Dados de Proveniência, Análise Visual, Sistemas Distribuídos

Abstract

Federated Learning (FL) enables collaborative Machine Learning (ML) model training across multiple distributed clients while preserving data locality through a well-defined workflow. However, the decentralized nature of FL execution complicates workflow observability and, consequently, performance diagnosis, the interpretation of training outcomes, and related analyses. Existing observability solutions are typically framework-specific, relying on execution logs or proprietary telemetry that limit interoperability and cross-framework analysis. We argue that provenance provides a data backbone for FL observability by capturing execution semantics for querying, analysis, and decision-support exploration. We present **FedObserve**, a provenance-driven observability framework that integrates runtime provenance capture, a unified provenance model, and interactive visual analytics. **FedObserve** captures workflow-, round-, and client-level provenance data independently of the underlying FL framework, enabling heterogeneous executions to be analyzed through a common representation. Its provenance visualization capabilities support multi-granularity analysis of execution traces, client participation, model evolution, convergence, and performance, allowing practitioners to iteratively inspect and diagnose FL behavior. We evaluate **FedObserve** using a FL workflow for outlier detection in large astronomical datasets. A user study with FL experts demonstrates that **FedObserve** supports identifying execution bottlenecks, client participation patterns, spatial heterogeneity, and convergence behavior. These results show that provenance-driven, decision-support observability improves the traceability, interpretability, and diagnosis of FL workflows while enabling framework-independent monitoring.

Keywords: Federated Learning, Data Provenance, Visual Analytics, Distributed Systems.

Lista de Figuras

1	Visão geral de um workflow genérico de FL. O servidor inicializa e distribui o modelo global aos clientes, que realizam treinamento local sobre conjuntos de dados privados e retornam atualizações de modelo para agregação. O processo itera por várias rodadas até que o modelo convirja ou um critério de parada seja satisfeito.	13
2	Visão geral dos desafios no monitoramento e depuração de workflows de FL.	13
3	Conceitos centrais e relações do modelo W3C PROV.	20
4	Visão geral da arquitetura da FedObserve.	28
5	Hierarquia de dados de proveniência na FedObserve.	30
6	Esquema de proveniência orientado a relações para o workflow genérico de FL.	32
7	Interface de visão geral do workflow da FedObserve.	36
8	Interface de análise da FedObserve. (1) Round/Client Overview, (2) Resumo de estatísticas, (3) Model Training Overview, (4) Quality Overview e (5) Convergence Analysis.	37
9	Visão de Participação de Clientes.	39
10	Projeção de modelos locais e globais de clusterização em espaço bidimensional via t-SNE.	40
11	Interações entre os gráficos <i>Cluster Center Projection</i> e <i>Client Participation Overview</i>	42
12	Painel de Análise de Convergência.	43

Lista de Tabelas

1	Comparação de abordagens representativas para gestão de proveniência e visual analytics em Aprendizado Federado.	23
2	Resultados da avaliação de usabilidade, incluindo notas e observações qualitativas para as dimensões avaliadas.	52

Lista de Abreviaturas e Siglas

DLProv DLProv (sistema de proveniência para Deep Learning)

ETL Extração, Transformação e Carga

FATE FATE (Federated AI Technology Enabler)

FATE-Board FATE-Board (ferramenta de visualização para FATE)

FL Aprendizado Federado (Federated Learning)

Flower Flower (framework de Federated Learning)

FTL Aprendizado Federado por Transferência (Federated Transfer Learning)

GDPR General Data Protection Regulation (Regulamento Geral sobre a Proteção de Dados)

HetVis HetVis (sistema de caracterização da heterogeneidade em HFL)

HFL Aprendizado Federado Horizontal (Horizontal Federated Learning)

HFLens HFLens (sistema de análise visual para HFL)

LGPD Lei Geral de Proteção de Dados

ML Aprendizado de Máquina (Machine Learning)

MLflow MLflow (plataforma para gerenciamento do ciclo de vida de ML)

MLOps Machine Learning Operations (Operações de Aprendizado de Máquina)

PROV-ML Modelo PROV para Machine Learning (PROV-ML)

RGPD Regulamento Geral sobre a Proteção de Dados

TI Tecnologia da Informação

VADAF Visual Analysis and Detection of Anomalies in Federated Learning (Análise Visual e Detecção de Anomalias em Aprendizado Federado)

VFL Aprendizado Federado Vertical (Vertical Federated Learning)

Sumário

1	Introdução	12
2	Referencial Teórico	17
2.1	Fundamentos de Aprendizado Federado	17
2.2	Proveniência de Dados em Resumo	18
3	Trabalhos Relacionados	21
3.1	Captura e Análise de Proveniência	21
3.2	Visual Analytics para Aprendizado Federado	22
3.2.1	Análise Comparativa	22
4	Princípios de Design	24
5	A Abordagem FedObserve	27
5.1	Arquitetura	27
5.2	Esquema Unificado de Proveniência	29
5.3	Interface de Exploração Visual	34
5.3.1	Workflow Overview	34
5.3.2	Round/Client Overview	35
5.3.3	Model Training Overview	38
5.3.4	Model Quality/Convergence Overview	41
6	Avaliação Experimental	45
6.1	Estudo de Caso: Detecção Federada de Outliers	45

6.2	Protocolo de Avaliação	46
6.3	Feedback dos Especialistas	49
7	Conclusão e Trabalhos Futuros	53
	REFERÊNCIAS	55

1 Introdução

Na última década, o Machine Learning (ML) ([Alpaydin \(2020\)](#)) teve ampla adoção em diversos domínios de aplicação, *e.g.*, Biomedicina ([Koivu *et al.* \(2018\)](#), [Peralta, Jannin e Baxter \(2021\)](#), [Braven *et al.* \(2026\)](#) e [Lindskog *et al.* \(2026\)](#)), Astronomia ([D’Isanto *et al.* \(2016\)](#), [Alves *et al.* \(2025\)](#) e [Lopes, Ferreira *et al.* \(2025\)](#)), Finanças ([Nepomuceno *et al.* \(2025\)](#), *etc.*). A adoção de ML acelerou especialmente em 2012, após a introdução do AlexNet ([Krizhevsky, Sutskever e Hinton \(2017\)](#)), sendo em grande parte impulsionada pelos avanços em Deep Learning (DL) ([LeCun, Bengio e Hinton \(2015\)](#)). Além do DL, a disponibilidade de grandes volumes de dados, melhorias nas tecnologias de coleta e armazenamento de dados e avanços na infraestrutura de computação contribuíram para essa aceleração ([Haenlein e Kaplan \(2019\)](#)). Em particular, a ampla disponibilidade de unidades de processamento gráfico (GPUs) de alto desempenho, tanto on-premises quanto em plataformas de nuvem, possibilitou o treinamento eficiente de modelos complexos sobre conjuntos de dados em larga escala.

Apesar de os algoritmos de ML, combinados com hardware de alto desempenho, terem ampliado a aplicabilidade do ML, muitas aplicações ainda enfrentam desafios. Algoritmos de ML bem conhecidos frequentemente pressupõem que o conjunto de treinamento completo (incluindo todos os atributos e tuplas) está disponível em um único local antes do início do treinamento do modelo. Essa suposição torna-se impraticável em muitos ambientes intensivos em dados ([Oliveira, Liu e Pacitti \(2019\)](#)). Nesse contexto, dois grandes obstáculos dificultam a adoção do treinamento centralizado de ML: *(i)* o volume de dados e *(ii)* a privacidade dos dados. Do ponto de vista do volume, muitas aplicações geram conjuntos de dados grandes demais para serem transferidos, armazenados ou processados em uma única máquina ou mesmo em um pequeno cluster. Por exemplo, o próximo Legacy Survey of Space and Time (LSST) deve produzir catálogos que somam centenas de terabytes de dados por ano ([Jurić *et al.* \(2015\)](#) e [Acero-Cuellar *et al.* \(2025\)](#)). Transferir e armazenar esses dados em um repositório centralizado para treinamento de modelos pode ser inviável.

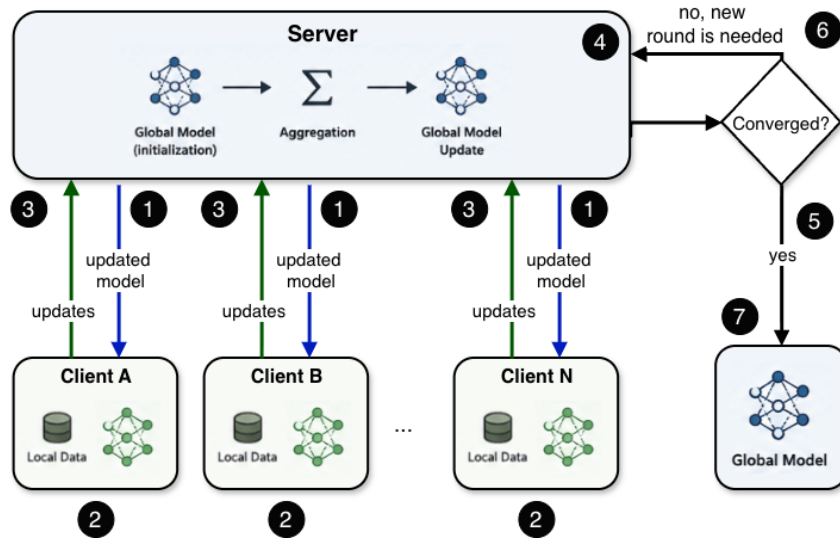


Figura 1: Visão geral de um workflow genérico de FL. O servidor inicializa e distribui o modelo global aos clientes, que realizam treinamento local sobre conjuntos de dados privados e retornam atualizações de modelo para agregação. O processo itera por várias rodadas até que o modelo convirja ou um critério de parada seja satisfeito.

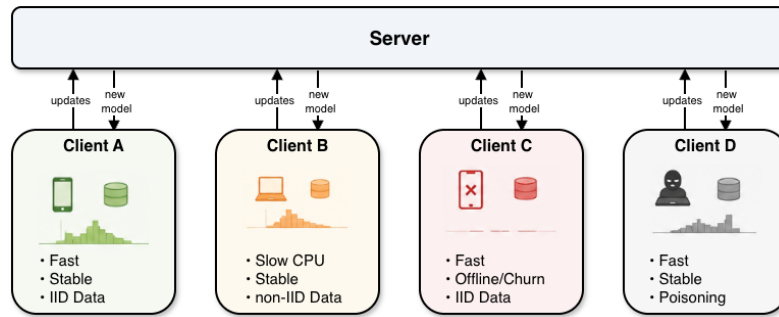


Figura 2: Visão geral dos desafios no monitoramento e depuração de workflows de FL.

A privacidade dos dados impõe um desafio de igual importância. Em muitas aplicações, os dados necessários para treinar modelos de ML contêm informações sensíveis que não podem ser compartilhadas. Alguns marcos regulatórios, como o GDPR (General Data Protection Regulation) na União Europeia ([European Parliament and Council of the European Union \(2016\)](#)) e a LGPD (Lei Geral de Proteção de Dados) no Brasil ([Presidência da República do Brasil \(2018\)](#)), impõem requisitos sobre a coleta, o armazenamento e o compartilhamento de dados pessoais e sensíveis. Assim, dados distribuídos entre múltiplas organizações estão sujeitos a restrições legais e éticas que impedem sua agregação em um único repositório. Esses desafios evidenciam a necessidade de paradigmas de ML escaláveis e preservadores de privacidade que minimizem transferências de dados ao mesmo tempo em que permitem o treinamento de modelos sobre conjuntos de dados distribuídos.

Para enfrentar os desafios acima, o Federated Learning (FL) ([McMahan *et al.* \(2017\)](#)) foi proposto como um paradigma distribuído de ML que possibilita o treinamento colaborativo de modelos entre múltiplas fontes de dados sem a troca de dados brutos. O treinamento em FL segue um workflow iterativo. Inicialmente, um servidor central envia um modelo ML inicial a todos os clientes (passo ❶ na Figura 1). Em cada rodada de treinamento, os clientes calculam atualizações dos parâmetros do modelo usando seus dados locais (passo ❷ na Figura 1) e enviam apenas essas atualizações ao servidor central (passo ❸ na Figura 1). O servidor agrega as atualizações recebidas para produzir um novo modelo global (passo ❹ na Figura 1). Esse modelo é avaliado para determinar se houve convergência ou se um critério de parada foi atendido (passo ❺ na Figura 1). Caso negativo, o servidor distribui o novo modelo global aos clientes e inicia uma nova rodada de treinamento (passo ❻ na Figura 1). Caso positivo, produz-se o modelo global final (passo ❼ na Figura 1). Ao eliminar a necessidade de transferir dados brutos, o FL preserva a localidade e a privacidade dos dados e reduz os custos de comunicação em comparação ao treinamento centralizado.

Apesar de suas vantagens, a natureza iterativa e descentralizada do workflow de FL introduz uma nova classe de desafios para compreender, monitorar e gerir o workflow. Workflows de FL distribuem o treinamento entre múltiplos clientes com distribuições de dados heterogêneas, recursos computacionais distintos, capacidades de memória e largura de banda de rede variadas. Essa execução distribuída complica diversas tarefas, *e.g.*, identificar gargalos causados por stragglers ou heterogeneidade de recursos, diagnosticar problemas de convergência decorrentes de dados não-IID (non-independent and identically distributed) ou de configurações de hiperparâmetros inadequadas, e detectar comportamento anômalo de clientes. Tais desafios tornam-se ainda mais críticos quando workflows de FL escalam para centenas de clientes, como em Cross-Device FL ([Yang *et al.* \(2019\)](#)). Conforme ilustrado na Figura 2, o comportamento do modelo global emerge da execução coordenada de clientes heterogêneos, cujas diferentes características computacionais, estatísticas, de comunicação e de participação tornam a análise de causa-raiz significativamente mais complexa do que no ML centralizado.

Dessa forma, a observabilidade de workflow ([Mattoso *et al.* \(2015\)](#)) torna-se um requisito fundamental para workflows de FL. Usuários devem ser capazes de monitorar e inspecionar o progresso do treinamento, analisar métricas do lado do cliente e do servidor, e identificar condições que demandem intervenção. A incorporação de uma abordagem de suporte à decisão para workflows ([Liu, Wilson e Gunning \(2014\)](#) e [Mattoso *et al.* \(2015\)](#)) fortalece essa capacidade ao permitir que especialistas de domínio e de ML monitorem o

processo de treinamento, interpretem resultados intermediários e promovam alterações em execuções subsequentes do workflow (ou durante a execução corrente, quando suportado pelo framework de FL), *e.g.*, ajustar hiperparâmetros, selecionar clientes participantes ou interromper execuções de workflow malsucedidas. Esse monitoramento interativo é particularmente valioso em aplicações científicas, onde a qualidade do modelo, a eficiência computacional e o uso de recursos devem ser balanceados cuidadosamente ao longo da execução do workflow.

Embora vários sistemas e frameworks de FL forneçam diferentes graus de capacidades de observabilidade, *e.g.*, HFLens ([Quan Li *et al.* \(2022\)](#)), HetVis ([Wang, Chen *et al.* \(2022\)](#)), VADAF ([Meng *et al.* \(2021\)](#)), Flower ([Beutel *et al.* \(2022\)](#)), NVIDIA FLARE ([NVIDIA Corporation \(2025\)](#)), e FATE-Board ([Fate-Board \(2026\)](#)), o suporte deles à análise de workflows de FL permanece limitado e, de certa forma, pouco explorado. Muitas abordagens dependem de logs de execução fragmentados, métricas agregadas ou telemetria específica de framework, dificultando a reconstrução do histórico completo de execução de um workflow de FL. Assim, analisar as relações entre eventos locais dos clientes e o comportamento do modelo global continua sendo desafiador, especialmente ao investigar degradação de desempenho, convergência ou resultados anômalos. Ademais, as abordagens de visualização e análise existentes costumam ser projetadas para frameworks ou tarefas de análise específicas. HFLens e VADAF acoplam fortemente seus mecanismos de observabilidade a frameworks particulares de FL, enquanto HetVis concentra-se principalmente em compreender os efeitos da heterogeneidade de dados. Essa especialização limita a aplicabilidade desses sistemas em ambientes heterogêneos de FL, onde múltiplos frameworks, arquiteturas de implantação e estratégias de execução podem coexistir.

Nesse sentido, uma limitação identificada é a falta de interoperabilidade entre ecossistemas de FL. A ausência de uma representação comum para trilhas de execução, metadados e eventos de sistema dificulta observar (e eventualmente comparar) execuções entre múltiplos frameworks, reproduzir resultados experimentais ou desenvolver ferramentas de análise agnósticas ao framework. Como consequência, usuários não podem realizar prontamente estudos comparativos, identificar padrões comuns de execução ou reutilizar conclusões de depuração e otimização entre plataformas de FL distintas. Do ponto de vista de gerenciamento de dados, isso expõe um desafio mais amplo: sistemas de FL carecem de uma representação unificada da semântica de execução que suporte consulta, integração e análise em ambientes heterogêneos. Dados de proveniência ([Herschel, Dietelkämper e Lahmar \(2017\)](#)) oferecem uma solução ao capturar dependências finas entre dados, modelos, participantes e processos computacionais. Uma representação baseada

em dados de proveniência fornece rastreabilidade de ponta a ponta, suporta interoperabilidade entre frameworks de FL e estabelece uma base comum para observabilidade por meio de consultas sobre dados de proveniência. Embora os dados de proveniência permitam consulta e rastreabilidade, consultar apenas proveniência tabular frequentemente é insuficiente para observabilidade com suporte à decisão. Soluções baseadas em consultas não são, em geral, desenhadas para suportar a natureza exploratória e iterativa da análise interativa de dados. Sistemas de visualização ([Heer, Bostock e Ogievetsky \(2010\)](#)) preenchem essa lacuna ao transformar dados de proveniência complexos e de alta dimensão em representações visuais intuitivas que permitem aos usuários interagir diretamente com a informação de execução.

Este manuscrito apresenta a **FedObserve**, uma abordagem de visual analytics orientada por dados de proveniência que fornece observabilidade interativa para workflows de FL. Diferentemente das abordagens predominantes, a **FedObserve** é construída sobre um banco de dados central de proveniência que captura metadados descrevendo dados, modelos, clientes, servidores e processos computacionais independentemente do framework subjacente de FL. Esse banco unificado possibilita uma visão holística do histórico de execução ao mesmo tempo em que suporta análise independente de framework. A **FedObserve** consulta o banco de dados de proveniência para gerar um conjunto de visualizações que facilitam a exploração de padrões de execução e a identificação de problemas, *e.g.*, clientes com mau funcionamento, comportamento anômalo no treinamento e convergência do modelo ao longo das rodadas. A interface visual dá suporte à análise interativa dos dados de proveniência. Ademais, a **FedObserve** é extensível: dados adicionais de proveniência podem ser incorporados ao banco conforme necessário, e novas visualizações podem ser integradas para atender requisitos emergentes de análise. A abordagem proposta é avaliada por meio de um estudo de caso envolvendo detecção federada de outliers para levantamentos astronômicos, utilizando um algoritmo de clusterização federada. Os resultados demonstram que a observabilidade orientada por dados de proveniência facilita a compreensão de execuções complexas de FL e que a **FedObserve** apoia efetivamente a depuração, o monitoramento e a análise interativa de workflows de FL.

O restante deste manuscrito está organizado da seguinte forma. A Seção 2 fornece o embasamento necessário e a Seção 3 revisa trabalhos relacionados. A Seção 4 apresenta os princípios de design subjacentes à **FedObserve**. A Seção 5 descreve a arquitetura do framework, o modelo de dados de proveniência e os componentes de visual analytics da **FedObserve**. A Seção 6 apresenta a avaliação experimental e discute os resultados. Finalmente, a Seção 7 conclui o manuscrito e aponta direções para trabalhos futuros.

2 Referencial Teórico

Esta seção apresenta os conceitos básicos associados à **FedObserve**. Primeiro, a Subseção [2.1](#) introduz o conceito de Aprendizado Federado. Em seguida, a Subseção [2.2](#) explica o conceito de proveniência de dados como um mecanismo para registrar o histórico de execução de workflows computacionais.

2.1 Fundamentos de Aprendizado Federado

Como mencionado anteriormente, FL ([McMahan et al. \(2017\)](#)) é um paradigma distribuído de ML que permite o treinamento colaborativo de modelos entre múltiplos clientes (cada um com seus próprios dados privados) ao evitar a consolidação dos dados brutos em um armazenamento centralizado. Em vez de transferir conjuntos de dados para um repositório compartilhado, o FL desloca o processamento para as fontes de dados, permitindo que os clientes treinem modelos locais independentemente e troquem apenas metadados intermediários de aprendizado (pesos de uma rede neural ou centróides em K-Means). Esse design descentralizado é relevante em domínios intensivos em dados onde regulamentações de privacidade, limites de propriedade ou gargalos de transferência impedem o compartilhamento direto de dados ([Hard et al. \(2018\)](#)).

Dependendo de como os dados são distribuídos entre os clientes, as abordagens de FL são comumente organizadas em diferentes categorias: *(i)* Aprendizado Federado Horizontal (HFL), *(ii)* Aprendizado Federado Vertical (VFL), e *(iii)* Aprendizado Federado por Transferência (FTL). HFL considera cenários em que os clientes mantêm conjuntos de dados com esquemas similares, mas com tuplas diferentes, permitindo treinar um modelo global explorando tuplas adicionais. Em contraste, VFL foca em cenários onde os clientes apresentam atributos de esquema distintos, exigindo coordenação entre espaços de atributos complementares para aprender um modelo global conjunto. Por fim, FTL considera cenários onde nem o espaço de atributos nem as tuplas se sobrepõem, dependendo de mecanismos de transferência de conhecimento para melhorar modelos em domínios

independentes.

O FL também pode ser categorizado pelo ambiente e pela natureza dos clientes participantes: (i) Cross-silo, (ii) Cross-device, e (iii) Híbrido. Cross-silo envolve um pequeno número de organizações com recursos estáveis, enquanto Cross-device envolve muitos clientes (smartphones, IoT) com variabilidade e conectividade instável. O híbrido combina características de ambos.

A adoção recente de FL motivou o desenvolvimento de diversos frameworks que fornecem infraestrutura para projetar, implantar e gerir workflows de FL. Esses frameworks abstraem os passos apresentados na Figura 1, *i.e.*, comunicação cliente-servidor, treinamento local, mecanismos de agregação e gerenciamento de recursos. Soluções como TensorFlow Federated (TFF) (Google Research (2024)) e Flower (Beutel *et al.* (2022)) suportam diferentes cenários de FL oferecendo ambientes de execução para coordenar o aprendizado entre clientes distribuídos.

No Aprendizado Federado, o treinamento ocorre de forma iterativa em múltiplas rodadas de comunicação, nas quais um servidor seleciona clientes, distribui o estado atual do modelo global, recebe as atualizações obtidas através do treino realizado localmente e executa a agregação para produzir a próxima versão do modelo. esse ciclo se repete até que se alcance um critério de parada, que pode estar associado à convergência ou à qualidade desejada para o modelo (McMahan *et al.* (2017) e Kairouz, McMahan, Avent *et al.* (2021)). Cenários heterogêneos são passíveis de variações de disponibilidade, capacidade computacional e distribuição dos dados entre clientes podem afetar o progresso a cada rodada, tornando a análise desse processo central para interpretar estabilidade e desempenho do treinamento federado (Tian Li *et al.* (2020)).

2.2 Proveniência de Dados em Resumo

Proveniência é definida como "(...) a origem e o histórico de um objeto ou artefato de informação"segundo Oxford University Press (2026). No contexto de workflows computacionais, como o workflow de FL, os dados de proveniência representam o caminho de derivação de um dado. A proveniência deve ser representada de forma estruturada e consultável para possibilitar sua análise ao longo do ciclo de vida dos dados (Freire *et al.* (2008)). Ao coletar, armazenar e consultar dados de proveniência, torna-se possível entender como um dado foi gerado e rastrear as transformações e modificações às quais foi submetido (Herschel, Diestelkämper e Lahmar (2017)).

Uma forma prática de entender proveniência é vê-la como um histórico estruturado da execução: quem executou uma etapa, com quais parâmetros, quais dados de entrada foram usados e quais artefatos foram produzidos. Diferentemente de logs textuais isolados, a proveniência preserva relações explícitas de derivação entre artefatos e atividades, permitindo reconstruir o caminho de geração de um resultado e responder perguntas analíticas de forma reproduzível (Freire *et al.* (2008), Herschel, Diestelkämper e Lahmar (2017)).

No contexto de workflows, a proveniência tem sido usada para suportar diversas tarefas, dentre elas a reprodutibilidade de workflows (Missier *et al.* (2016) e Rupprecht *et al.* (2020)). Além da reprodutibilidade, a proveniência tem sido explorada para escalonamento (Oliveira, Ocaña *et al.* (2012), Stockinger *et al.* (2022), Liu, Pacitti *et al.* (2017) e Wang e Yin (2013)), tolerância a falhas (Guedes *et al.* (2020), Han *et al.* (2017) e Crawl e Altintas (2008)), estimativa de recursos (Alvarenga *et al.* (2025)), caching (Heidsieck *et al.* (2021)), depuração (Oliveira, Costa *et al.* (2014)) e explicabilidade de resultados (Pina *et al.* (2025) e Zhang, Zhou e Ujcich (2024)).

Os dados de proveniência podem ser classificados segundo o tipo de metadado representado. Uma taxonomia amplamente adotada distingue *Prospective Provenance* (p-prov) e *Retrospective Provenance* (r-prov) (Freire *et al.* (2008)). A p-prov descreve a especificação do workflow (atividades, parâmetros, dependências), enquanto a r-prov registra a história efetiva da execução (valores produzidos, eventos, uso de recursos, falhas).

No contexto de FL, a p-prov descreve o plano de execução (por exemplo, número de rodadas, política de seleção de clientes e hiperparâmetros), enquanto a r-prov registra o que de fato ocorreu em cada rodada (clientes participantes, tempos de treinamento/agregação, métricas observadas e falhas). Essa separação é útil porque distingue o desenho experimental pretendido do comportamento real observado durante a execução (Freire *et al.* (2008)).

Embora vários modelos de proveniência tenham sido propostos (Moreau, Clifford *et al.* (2011), Costa *et al.* (2013) e Butt e Fitch (2020)), a recomendação W3C PROV (Moreau, Groth *et al.* (2015)) é o modelo conceitual de referência. PROV define três construtos fundamentais: *entities*, *activities* e *agents*, cujas relações padronizam a representação da proveniência e facilitam a interoperabilidade entre sistemas.

No mapeamento para W3C PROV, modelos locais e globais podem ser representados como *entities*, treinamento local e agregação como *activities*, e clientes/servidor como *agents*. Esse mapeamento padronizado facilita interoperabilidade e consulta entre execuções heterogêneas, mantendo semântica comum para análise comparativa (Moreau, Groth

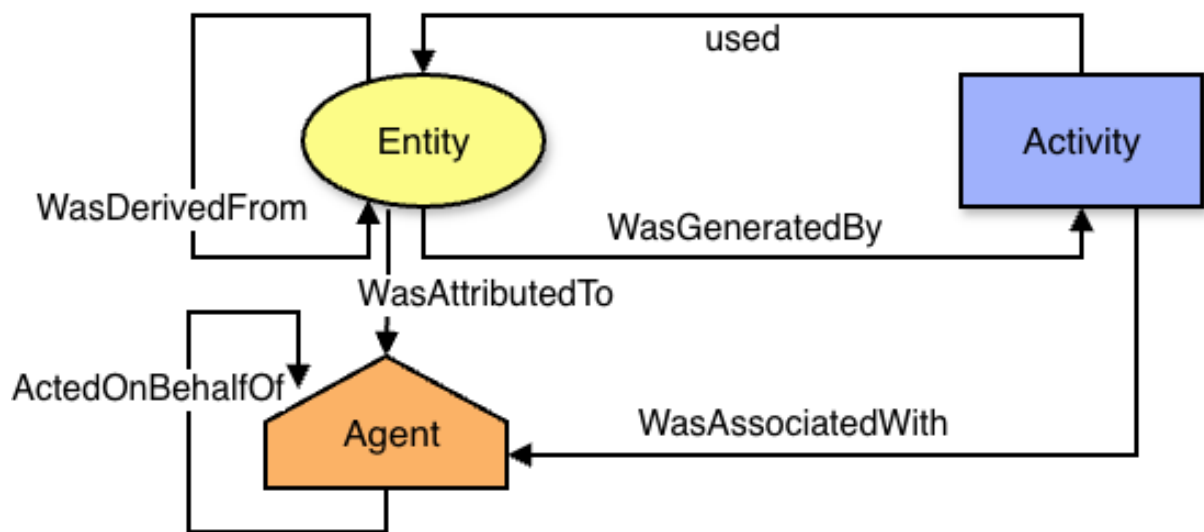


Figura 3: Conceitos centrais e relações do modelo W3C PROV.

et al. (2015), Moreau, Clifford *et al.* (2011), Costa *et al.* (2013) e Butt e Fitch (2020)).

3 Trabalhos Relacionados

A abordagem proposta neste trabalho, a **FedObserve**, relaciona-se a duas áreas complementares: (i) captura e análise de proveniência para ML e Aprendizado Federado (FL), e (ii) visual analytics para FL. Nesta seção, revisamos os trabalhos mais relevantes e discutimos como a **FedObserve** se diferencia das abordagens existentes.

3.1 Captura e Análise de Proveniência

Como discutido na Seção 2, dados de proveniência desempenham papel fundamental no gerenciamento de workflows ao permitir analisar o comportamento de execução, identificar padrões e diagnosticar problemas. Motivadas por esses benefícios, diversas abordagens investigaram captura e gerenciamento de proveniência para workflows de ML e FL.

Suporte à proveniência em ML. Renan Souza *et al.* (2022) introduz PROV-ML, um modelo de dados para workflows de ML que suporta captura de proveniência ao longo do ciclo de vida do experimento. ProvBook (Samuel, Löffler e König-Ries (2020)) captura proveniência em Jupyter Notebooks para reprodutibilidade. MLflow2PROV (Schlegel e Sattler (2023)) gera grafos PROV a partir de metadados do MLflow, embora o resultado possa conter metadados de baixo nível que obscurecem a semântica do workflow. Outros trabalhos, como DPDS (Chapman, Missier *et al.* (2022), Chapman, Lauro *et al.* (2024) e Gregori *et al.* (2024)), capturam proveniência fina durante preparação de dados para melhorar explicabilidade e reprodutibilidade.

Suporte à proveniência em FL. Abordagens direcionadas a FL incluem frameworks que combinam técnicas criptográficas com gestão de modelos para rastrear transformações ao longo do ciclo de vida (Gu, Naraparaju e Zhao (2024)), mecanismos para identificar clientes que contribuíram para um modelo global (TraceFL Gill, Anwar e Gulzar (2025)), plataformas baseadas em blockchain para gerir proveniência (Bandara *et al.* (2022)), e extensões de frameworks (p.ex., Flower-PROV Lopes, Nunes *et al.* (2023)). ProvFL (Gill, Anwar e Gulzar (2023)) propõe grafos de proveniência ao nível de neurônio para expli-

cabilidade. Apesar desses avanços, ainda existe espaço para melhorar a análise integrada da proveniência, unindo captura, consulta e exploração visual.

3.2 Visual Analytics para Aprendizado Federado

Sistemas de visual analytics já foram propostos para melhorar a compreensão de workflows de FL, apoiando análise de comportamento do modelo, participação de clientes e dinâmica de treinamento distribuído. HFLens (Quan Li *et al.* (2022)) oferece inspeção em múltiplas granularidades; VADAF (Meng *et al.* (2021)) combina detecção de anomalias com visualização; HetVis (Wang, Chen *et al.* (2022)) foca em heterogeneidade de dados. Contudo, essas ferramentas geralmente utilizam logs ou metadados específicos de framework, limitando comparações entre plataformas. Em contraste, a FedObserve adota proveniência como representação agnóstica de framework, permitindo análise comparativa e exploração integrada.

Embora não seja específico para FL, o ProWis (Carolina Veiga Ferreira de Souza *et al.* (2024)) apresenta uma abordagem relevante para este trabalho ao combinar, em uma mesma arquitetura, *provenance-oriented design*, gerenciamento de execuções e exploração visual interativa em tempo de execução.

3.2.1 Análise Comparativa

A Tabela 1 compara abordagens representativas com a FedObserve. Trabalhos de proveniência em FL focam em captura, rastreabilidade e reprodutibilidade, mas tipicamente atendem a tarefas específicas (atribuição de clientes, governança, otimização) em vez de fornecer uma base geral para análise exploratória integrada. Sistemas de visual analytics facilitam inspeções específicas, porém costumam ser acoplados a formatos proprietários. A FedObserve adota W3C PROV como representação unificada e integra captura e análise visual interativa em uma arquitetura única.

Os critérios de comparação da Tabela 1 foram definidos a partir das perguntas analíticas que orientam a observabilidade de workflows federados e que também direcionaram nossa avaliação com especialistas. Os critérios foram aprimorados ao longo do processo de comparação, onde foram consideradas as capacidades de cada uma das abordagens avaliadas. Para tornar essa comparação objetiva entre trabalhos com escopos distintos, traduzimos essas perguntas em cinco capacidades observáveis na literatura: (i) suporte explícito à proveniência, (ii) mecanismos de analytics, (iii) suporte visual para inspeção

interativa, *(iv)* independência de framework e *(v)* suporte em tempo de execução. Assim, cada coluna da tabela representa uma capacidade necessária para viabilizar as tarefas analíticas centrais de diagnóstico e monitoramento em FL.

Tabela 1: Comparação de abordagens representativas para gestão de proveniência e visual analytics em Aprendizado Federado.

Abordagem	Proveniência	Analytics	Suporte Visual	Indep. de Framework	Suporte Tempo de Execução
Gu, Naraparaju e Zhao (2024)	✓	×	×	×	×
TraceFL Gill, Anwar e Gulzar (2025)	✓	×	×	×	×
Bassa-ML Bandara <i>et al.</i> (2022)	✓	×	×	×	×
Peregrina, Ortiz e Zirpins (2022)	✓	×	×	×	×
Schlegel, Scheliga <i>et al.</i> (2026)	✓	✓	×	×	×
ProvFL Gill, Anwar e Gulzar (2023)	✓	✓	×	×	×
Flower-PROV Lopes, Nunes <i>et al.</i> (2023)	✓	✓	×	×	✓
HFLens Quan Li <i>et al.</i> (2022)	N/A	✓	✓	×	✓
VADAF Meng <i>et al.</i> (2021)	N/A	✓	✓	×	✓
HetVis Wang, Chen <i>et al.</i> (2022)	N/A	✓	✓	×	✓
ProWis Carolina Veiga Ferreira de Souza <i>et al.</i> (2024)	✓	✓	✓	×	✓
FedObserve	✓	✓	✓	✓	✓

4 Princípios de Design

Como mencionado anteriormente, o objetivo da **FedObserve** é suportar observabilidade orientada por proveniência de workflows de FL por meio de captura integrada de proveniência, visual analytics e exploração interativa. Embora os dados de proveniência forneçam um registro da execução do workflow de FL, compreender o comportamento do treinamento distribuído requer mais do que coletar trilhas de execução. Os usuários devem ser capazes de analisar as interações entre clientes, rodadas de treinamento, operações de agregação, modelos e métricas de execução, além de comparar experimentos executados sob diferentes configurações ou frameworks de FL. Defendemos que a proveniência fornece uma representação unificada capaz de suportar essas análises ao capturar explicitamente as entidades envolvidas em um workflow de FL e as relações entre elas. Esta seção apresenta os requisitos que guiaram o *design* da **FedObserve**. Os requisitos propostos foram derivados a partir dos objetivos de suportar análise orientada por proveniência de execuções individuais de FL. Esse objetivo levanta as seguintes questões:

- Quais dados de proveniência devem ser coletados para suportar a análise de um workflow de FL?
- Quais capacidades de visual analytics devem ser fornecidas para auxiliar usuários a observar e compreender execuções de FL?

Essas questões guiaram o *design* da **FedObserve**, definindo tanto os dados mínimos de proveniência que devem ser capturados durante a execução do workflow quanto as capacidades analíticas oferecidas pela **FedObserve**. Para suportar análises orientadas por proveniência, o banco de dados de proveniência deve descrever o ciclo de vida completo de uma execução de workflow FL, incluindo configuração, rodadas de treinamento, clientes participantes, modelos gerados, métricas de execução e as relações entre esses elementos. As capacidades descritas a seguir correspondem aos requisitos de design da **FedObserve**. Parte delas é implementada e avaliada neste manuscrito, enquanto outras representam

decisões arquiteturais que orientam a extensibilidade da abordagem para cenários e algoritmos adicionais. Com base nesses requisitos, identificamos os seguintes dados que devem ser coletados:

1. A configuração do workflow, incluindo o algoritmo de aprendizado, os hiperparâmetros, o número de rodadas e o ambiente de execução.
2. As rodadas executadas durante o treinamento, incluindo sua ordem, duração, operações de agregação e clientes participantes.
3. A participação de cada cliente em cada rodada, incluindo status de execução local, modelos locais gerados e estatísticas de treinamento.
4. As métricas de execução produzidas ao longo do treinamento, incluindo tempo de treinamento, tempo de comunicação, tempo de agregação, acurácia do modelo, perda e outras métricas definidas pelo usuário.
5. As relações de proveniência que conectam workflows, rodadas, clientes, modelos, métricas, conjuntos de dados e atividades.
6. A evolução temporal da execução do workflow, permitindo correlacionar eventos de proveniência com mudanças no comportamento do treinamento ao longo do tempo.

Os quatro primeiros requisitos capturam os principais componentes de um workflow de FL juntamente com suas características de execução. As relações de proveniência descrevem explicitamente como as entidades do workflow estão conectadas, permitindo que os usuários reconstruam o caminho de derivação dos dados e investiguem trajetórias de derivação. Por fim, as informações temporais permitem analisar a evolução do workflow ao longo do treinamento distribuído. Coletivamente, esses requisitos fornecem os dados de proveniência necessários para suportar a análise de execuções de FL.

Uma vez que os dados de proveniência tenham sido coletados e representados no padrão W3C PROV, os usuários devem ser capazes de explorá-los por meio de visualizações que expõem perspectivas complementares da execução do workflow. Em vez de focar em análises predefinidas, a **FedObserve** deve fornecer capacidades visuais interativas de propósito geral que permitam aos usuários investigar o comportamento do workflow, identificar padrões de execução, *etc.* Com base nas tarefas analíticas identificadas nos trabalhos relacionados e nas necessidades comuns de monitoramento em FL, definimos as seguintes capacidades de visual analytics:

1. Visão Geral do Workflow: fornecer um resumo de alto nível de uma execução de workflow de FL, incluindo a configuração do workflow, os clientes participantes, a duração da execução e o desempenho do treinamento.
2. Visão Geral por Rodada: permitir a inspeção das rodadas de comunicação visualizando a evolução das métricas de execução, o comportamento de agregação e a participação dos clientes ao longo do processo de treinamento.
3. Visão Geral por Cliente: suportar a investigação do comportamento individual de cada cliente por meio de métricas de treinamento local, tempos de execução, frequência de participação e contribuição ao modelo global.
4. Visão Geral do Modelo: fornecer feedback visual em tempo real sobre os resultados do workflow, permitindo que os usuários verifiquem como o modelo está convergindo (ou não), resultados anômalos e comportamentos de execução inesperados durante o treinamento.
5. Visão Geral de Qualidade/Convergência: fornecer visualizações da convergência e qualidade do modelo por meio da evolução das métricas de treinamento e avaliação ao longo das rodadas de comunicação, suportando a avaliação do comportamento de aprendizado e a comparação de execuções de workflow.

Essas capacidades se complementam ao permitir que os usuários analisem workflows de FL a partir de diferentes perspectivas, desde resumos de alto nível até exploração detalhada de proveniência. O uso coordenado dessas visualizações permite aos usuários correlacionar dados de proveniência com métricas de execução, compreender como eventos do workflow influenciam o desempenho do modelo e analisar execuções em diferentes frameworks de FL usando uma representação comum de proveniência. Os requisitos apresentados nesta seção guiaram diretamente o *design* da **FedObserve**. O modelo de proveniência introduzido na Seção 5 captura as informações necessárias para representar workflows de FL independentemente do framework de FL subjacente, enquanto os componentes de visual analytics propostos fornecem as capacidades descritas anteriormente, permitindo exploração e monitoramento orientados por proveniência de execuções de workflows de FL.

5 A Abordagem FedObserve

A **FedObserve** é um framework de observabilidade orientado por proveniência projetado para monitorar e analisar workflows de FL. Sua arquitetura compreende múltiplos módulos independentes e interoperáveis, cada um responsável por um aspecto específico de captura, armazenamento, análise ou visualização de proveniência. Em conjunto, esses módulos fornecem uma infraestrutura para compreender o comportamento de execução de workflows de FL por meio de dados de proveniência. Nesta seção, apresentamos os componentes arquiteturais do framework (Subseção 5.1), o esquema de proveniência (Subseção 5.2) e a interface de análise visual (Subseção 5.3).

5.1 Arquitetura

A Figura 4 apresenta a arquitetura conceitual da **FedObserve**, mostrando como seus módulos interagem para prover observabilidade de ponta a ponta ao longo do ciclo de vida de workflows de FL. A arquitetura segue um design modular composto por três módulos principais: (i) o FL Framework, responsável por executar o workflow de FL e gerar dados de proveniência; (ii) o Provenance Ingestion & Query Service, que captura, armazena, indexa e expõe dados de proveniência por meio de uma interface de consulta unificada; e (iii) o Analytics & Exploration, que permite aos usuários analisar dados de proveniência por meio de visualizações. Essa separação de responsabilidades favorece a extensibilidade, permitindo que cada módulo evolua de forma independente.

A execução da **FedObserve** começa com o FL Framework, que executa o workflow de FL e gera dados de proveniência ao longo de seu ciclo de vida. A **FedObserve** suporta duas formas alternativas de interagir com frameworks de FL. A primeira baseia-se em instrumentação em tempo de execução, na qual o framework de FL é complementado com decorators de código ([Gamma et al. \(1994\)](#)) que capturam eventos de proveniência de forma transparente. A segunda utiliza dados de proveniência já expostos pelo framework por meio de APIs ou logs de execução. No protótipo implementado e avaliado

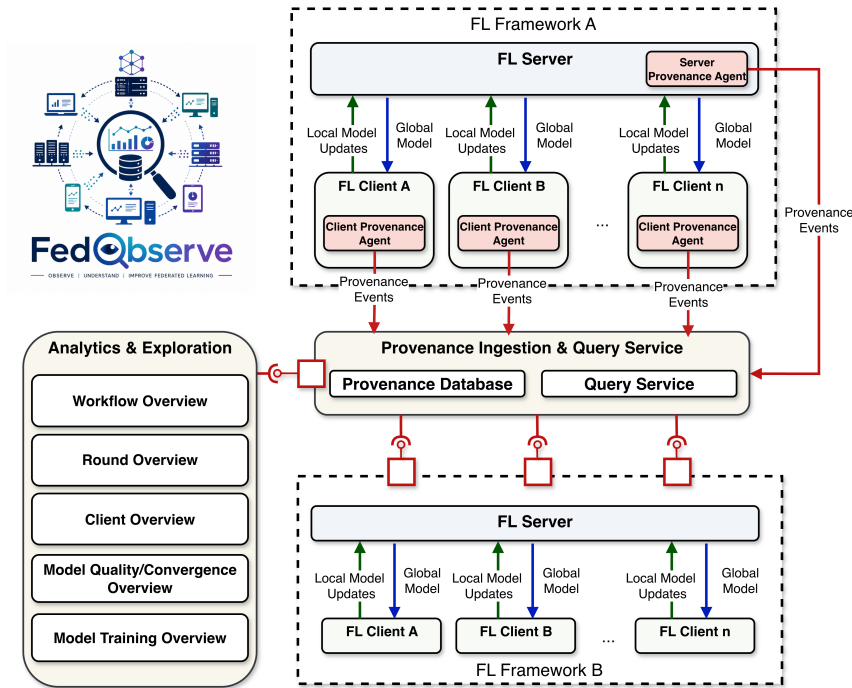


Figura 4: Visão geral da arquitetura da FedObserve.

neste manuscrito, adotamos a primeira estratégia. Independentemente da forma como a FedObserve interage com os frameworks de FL, todos os registros de proveniência capturados são enviados ao Provenance Ingestion & Query Service.

A principal responsabilidade do Provenance Ingestion & Query Service é mapear os dados de proveniência coletados para o esquema unificado adotado pela FedObserve (detalhado na Subseção 5.2). Este módulo compreende dois componentes principais. O primeiro é o Provenance Database, que armazena os dados de proveniência e pode ser implementado em sistemas de gerenciamento de banco de dados relacionais (RDBMS) como MonetDB, DuckDB ou PostgreSQL. O segundo é o Query Service, que fornece uma interface para acessar os dados de proveniência independentemente da implementação do banco de dados subjacente. Ao desacoplar o armazenamento da proveniência do seu acesso, essa camada simplifica o desenvolvimento de aplicações de observabilidade e permite utilizar diferentes backends de armazenamento conforme os requisitos de desempenho e escalabilidade.

Por fim, o módulo Analytics & Exploration consome dados de proveniência por meio do Query Service para fornecer capacidades visuais interativas (detalhadas na Subseção 5.3). Este módulo expõe uma interface de visual analytics que permite aos usuários monitorar a execução de workflows de FL em múltiplos níveis de granularidade, desde resumos em nível de workflow até análises detalhadas por cliente e por rodada. Por

meio dessa interface, os usuários podem inspecionar a execução do workflow, analisar o comportamento individual de cada cliente e identificar padrões de execução anômalos e gargalos de desempenho. Ao integrar captura de proveniência, armazenamento, consulta e visual analytics em um framework unificado, a **FedObserve** fornece uma abordagem de observabilidade orientada por proveniência para frameworks de FL heterogêneos. O código-fonte da **FedObserve**, incluindo o esquema de proveniência e a interface de analytics, está disponível publicamente em nosso repositório GitHub: <https://github.com/UFFeScience/FedObserve>.

5.2 Esquema Unificado de Proveniência

Como discutido na Subseção 5.1, o esquema de proveniência é a espinha dorsal de dados da **FedObserve**. Ele fornece a representação unificada sobre a qual a **FedObserve** constrói suas capacidades de observabilidade, viabilizando a coleta, integração e análise de dados de proveniência ao longo da execução do workflow de FL. Ao capturar metadados de execução, o esquema suporta as funcionalidades de visual analytics apresentadas na Subseção 5.3, permitindo que praticantes de FL inspecionem a execução do workflow, identifiquem possíveis anomalias e compreendam o comportamento dos processos distribuídos de FL. Para prover observabilidade, o esquema de proveniência deve capturar informações abrangendo todos os estágios do ciclo de vida de FL. Como os workflows de FL são inerentemente descentralizados, a proveniência é gerada independentemente pelo servidor central e por múltiplos clientes participantes, cada um mantendo uma visão parcial da execução global. Consequentemente, o problema de gestão de proveniência é em si naturalmente distribuído, exigindo que o esquema represente de forma consistente a proveniência produzida em ambientes de execução heterogêneos e geograficamente distribuídos, preservando as relações entre eventos do servidor e dos clientes.

Um requisito fundamental de projeto do esquema de proveniência da **FedObserve** é a extensibilidade. Na prática, os requisitos de observabilidade evoluem ao longo do tempo à medida que os usuários identificam métricas, parâmetros ou eventos de execução adicionais que devem ser monitorados. Portanto, o esquema deve suportar a adição de novos atributos de proveniência *a posteriori* sem comprometer a integridade dos dados de proveniência já armazenados nem exigir refatorações disruptivas do esquema. Essa extensibilidade permite que a **FedObserve** acomode diferentes frameworks de FL, domínios de aplicação e requisitos de monitoramento definidos pelo usuário, mantendo compatibilidade retroativa. A Figura 5 apresenta os três níveis complementares de proveniência considera-

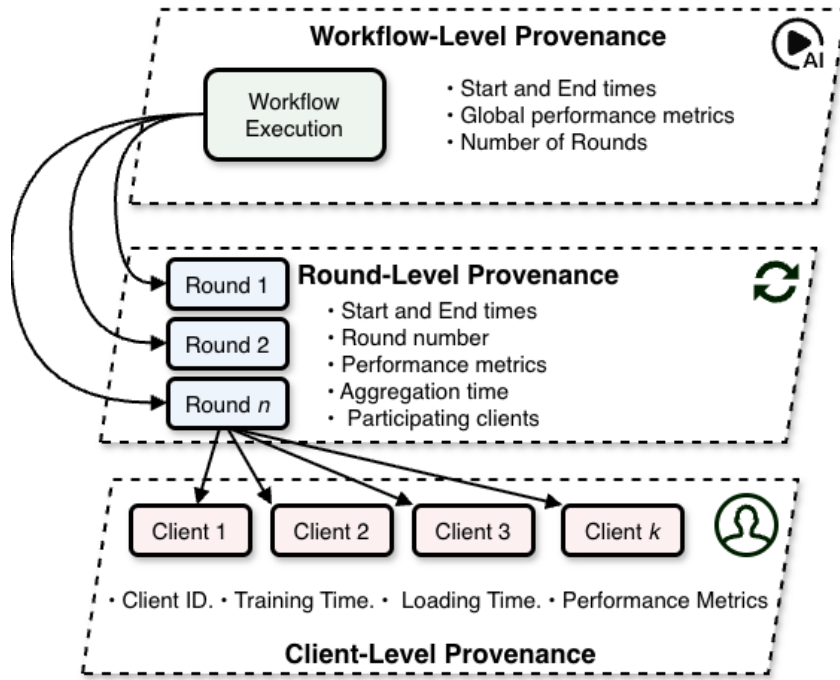


Figura 5: Hierarquia de dados de proveniência na FedObserve.

dos pela FedObserve. No nível mais alto, a Proveniência em Nível de Workflow descreve a execução de um workflow de FL como um todo, incluindo parâmetros de configuração globais e métricas agregadas, *e.g.*, o número de rodadas, o desempenho do modelo global e o status de execução do workflow. O segundo nível, a Proveniência em Nível de Rodada, captura metadados associados a cada rodada de treinamento, incluindo seus timestamps de início e fim, o número de clientes participantes e indicadores intermediários de desempenho. Por fim, a Proveniência em Nível de Cliente registra o comportamento de cada cliente participante durante cada rodada de treinamento, incluindo tempo de treinamento local, métricas de desempenho do modelo e outros metadados de execução específicos do cliente. Organizar os dados de proveniência segundo esses três níveis hierárquicos permite à FedObserve suportar observabilidade multi-granular. Os usuários podem navegar de uma visão global do workflow de FL inteiro até níveis progressivamente mais detalhados, chegando à inspeção da execução de um cliente específico em uma rodada específica.

Para definir um esquema comum de proveniência capaz de acomodar diferentes frameworks de FL, a FedObserve adota dois princípios fundamentais de projeto. Primeiro, o modelo de proveniência é definido em relação a uma especificação genérica de workflow de FL, e não aos detalhes de implementação de nenhum framework específico. Segundo, a proveniência é capturada por meio de APIs dos frameworks de FL ou de um mecanismo de instrumentação agnóstico de framework baseado em decorators, permitindo que

os dados de proveniência sejam coletados de forma transparente durante a execução do workflow, sem exigir modificações no modelo de proveniência subjacente. Em conjunto, essas escolhas de projeto garantem que a **FedObserve** permaneça portátil entre diferentes ecossistemas de FL enquanto fornece uma representação unificada de proveniência.

A especificação genérica de workflow de FL foi derivada por meio da abstração dos workflows descritos na literatura de FL ([McMahan *et al.* \(2017\)](#), [Tian Li *et al.* \(2020\)](#) e [Kairouz, McMahan, Avent *et al.* \(2021\)](#)). Embora os frameworks de FL existentes difiram em seus detalhes de implementação, geralmente seguem uma sequência comum de atividades lógicas. A **FedObserve** modela, portanto, um workflow de FL como a execução das seguintes atividades: *(i)* Preparação de Dados, onde os conjuntos de dados são adquiridos e pré-processados; *(ii)* Inicialização do Workflow, durante a qual o modelo global e a configuração de treinamento são definidos; *(iii)* Carregamento de Dados, responsável por distribuir e preparar os conjuntos de dados locais; *(iv)* Seleção e Inicialização de Clientes, na qual os clientes participantes são selecionados e inicializados para a rodada atual; *(v)* Treinamento do Cliente, onde os clientes otimizam independentemente seus modelos locais; *(vi)* Agregação de Modelos, durante a qual as atualizações dos modelos locais são combinadas para produzir um novo modelo global; *(vii)* Avaliação Global, onde o modelo agregado é validado usando métricas de avaliação predefinidas; e *(viii)* Finalização do Workflow, que conclui o processo de treinamento e armazena os artefatos resultantes e os metadados de execução.

Com base nessa especificação genérica de workflow de FL, a **FedObserve** adota o modelo de proveniência orientado a relações proposto por [Ogasawara *et al.* \(2011\)](#). Neste modelo, cada atividade do workflow é representada como uma transformação que consome uma relação de entrada e produz uma relação de saída. Consequentemente, cada atividade do workflow de FL é associada a um par de relações de proveniência que capturam tanto o contexto de execução quanto os produtos de dados gerados naquele estágio. Cada relação armazena os atributos necessários para suportar monitoramento, análise e diagnóstico subsequentes da execução do workflow de FL. A Figura 6 apresenta o esquema de proveniência utilizado pela **FedObserve**. As relações de entrada são representadas por retângulos verdes e identificadas pelo prefixo *i*, seguido da atividade correspondente do workflow (*e.g.*, *iClientTraining*). De forma análoga, as relações de saída são representadas por retângulos azuis e utilizam o prefixo *o*. Cada relação de entrada está conectada à sua relação de saída correspondente por uma aresta rotulada com a atividade do workflow de FL que transforma a entrada na saída.

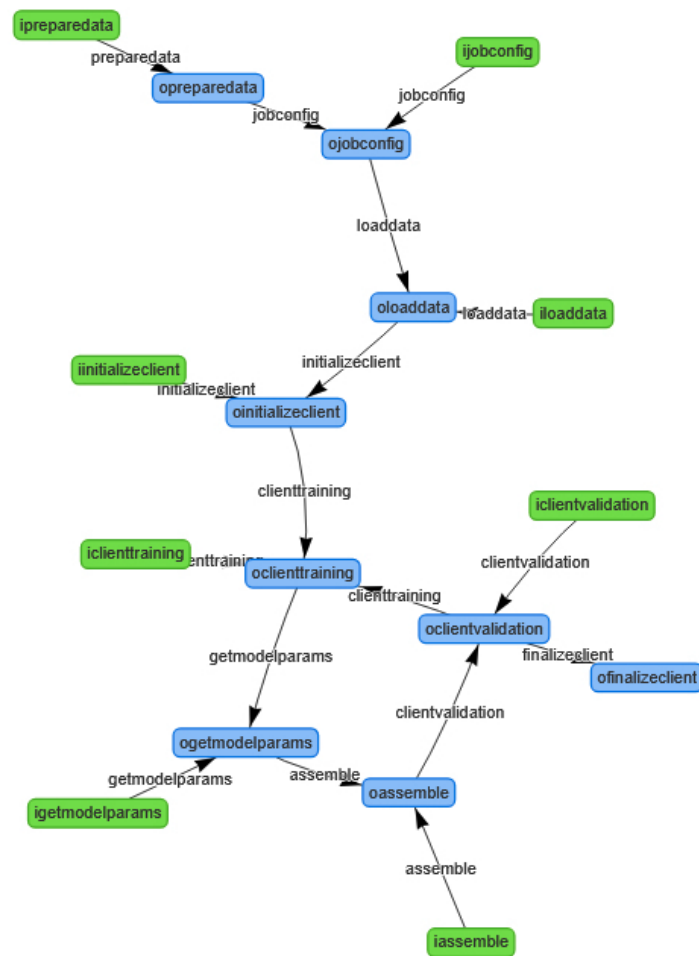


Figura 6: Esquema de proveniência orientado a relações para o workflow genérico de FL.

Uma vantagem importante desse design orientado a relações é sua independência de framework. Como o esquema de proveniência modela as atividades lógicas de um workflow de FL, e não os detalhes de implementação de um framework específico, ele pode acomodar proveniência gerada por diferentes sistemas de FL sem exigir modificações estruturais. Diferenças específicas de cada framework são refletidas naturalmente nas relações populadas. Por exemplo, se um framework não possui uma atividade explícita de preparação de dados, as relações correspondentes (`iDataPreparation` e `oDataPreparation`) simplesmente permanecem vazias, enquanto o restante do esquema de proveniência é preservado. Essa flexibilidade permite à **FedObserve** manter uma representação unificada de proveniência entre frameworks de FL heterogêneos.

Embora a proveniência forneça a base semântica para integração e rastreabilidade entre frameworks, ela não elimina a necessidade de escolhas analíticas e visuais dependentes da tarefa. Em outras palavras, a interoperabilidade é garantida no nível dos dados de proveniência, enquanto a interpretação eficaz desses dados ainda requer mapeamentos visuais alinhados ao tipo de modelo e ao objetivo de análise.

Uma vez definido o esquema de proveniência, ele deve ser populado com dados de proveniência coletados durante a execução dos workflows de FL. Na **FedObserve**, consideramos dois mecanismos alternativos para coletar dados de proveniência: (i) utilizar dados de proveniência já expostos por um framework de FL por meio de APIs ou logs de execução, e (ii) instrumentar o framework de FL para capturar proveniência em tempo de execução. No primeiro cenário, a **FedObserve** baseia-se em *cartridges* específicos de cada framework (Birsan (2005)), responsáveis por extrair dados de proveniência dos formatos suportados por cada framework (*e.g.*, logs em texto, arquivos JSON ou APIs) e transformá-los em registros de proveniência compatíveis com o esquema adotado. Embora essa abordagem viabilize a integração de frameworks existentes sem modificar seu código-fonte, exige a implementação e manutenção de um cartridge dedicado para cada framework suportado, o que pode se tornar trabalhoso à medida que o número de frameworks aumenta.

O segundo cenário consiste em instrumentar o framework de FL quando seu código-fonte está disponível, *e.g.*, Flower e NVIDIA FLARE. Neste manuscrito, adotamos essa estratégia usando a biblioteca DfAnalyzer (Silva, Oliveira *et al.* (2018) e Silva, Campos *et al.* (2020)), que fornece um mecanismo leve de instrumentação baseado em decorators (*i.e.*, injeção de código (Gamma *et al.* (1994))). Decorators são inseridos em pontos selecionados de execução do framework de FL para capturar dados de proveniência e enviá-los

ao banco de dados de proveniência em tempo de execução. Especificamente, os usuários anotam o código do framework de FL com chamadas do DfAnalyzer para registrar tanto *p-prov* quanto *r-prov*, que são automaticamente mapeados para as relações correspondentes do esquema de proveniência. O Listing ?? apresenta a definição de uma transformação de workflow usando o DfAnalyzer. Neste exemplo, a transformação `PrepareData` é associada a uma relação de entrada correspondente ao conjunto de dados bruto e dois atributos de saída representando o conjunto de dados pré-processado e o número de erros de pré-processamento, respectivamente. Uma vez definida a transformação, a biblioteca de captura de proveniência associa automaticamente os dados de proveniência gerados durante a execução às relações correspondentes no esquema de proveniência.

5.3 Interface de Exploração Visual

Como introduzido na Subseção 5.1, o módulo Analytics & Exploration é responsável por traduzir os dados de proveniência capturados pela `FedObserve` em representações visuais acionáveis. A interface foi projetada em torno do princípio de revelação progressiva: os usuários partem de uma visão de proveniência em nível de workflow e podem aprofundar-se até o nível de rodada e cliente. Ela compreende quatro funcionalidades principais: (i) Workflow Overview (*i.e.*, Resumo de Execução na Figura 7), que reporta a configuração e a escala de uma execução selecionada de workflow de FL; (ii) Round/Client Overview, que expõe como os clientes ingressam, treinam e deixam o pool de clientes ao longo de múltiplas rodadas; (iii) Model Training Overview, que visualiza o modelo treinado e as métricas de desempenho associadas; e (iv) Model Quality/Convergence Overview, que acompanha as métricas de convergência e qualidade do modelo (no contexto deste manuscrito, um modelo de clusterização). Todas as visualizações são interativas, construídas com a biblioteca D3.js, e populadas sob demanda por meio do Provenance Query Service.

5.3.1 Workflow Overview

O uso da `FedObserve` começa com a interface apresentada na Figura 7, que fornece uma visão geral dos dados de proveniência coletados de múltiplas execuções de workflows de FL. Na parte inferior da interface, fileiras de cartões compactos (área ❶ na Figura 7) resumem a proveniência em nível de workflow para cada execução. Para cada execução de workflow de FL, a `FedObserve` exibe metadados-chave, incluindo o nome do workflow, o número total de clientes participantes, o horário de início da execução, a duração total

(em minutos) e o status atual (*e.g.*, PENDING, COMPLETED, FAILED, *etc.*). O número de clientes refere-se ao conjunto de clientes envolvidos em toda a execução do workflow de FL, uma vez que o subconjunto de clientes participantes pode variar entre as rodadas.

À medida que o número de execuções de workflow pode crescer significativamente ao longo do tempo, inspecionar diretamente uma grande coleção de execuções pode se tornar desafiador. Para resolver esse problema, a **FedObserve** fornece uma visualização de coordenadas paralelas (área ❷ na Figura 7) que permite aos usuários filtrar e comparar interativamente execuções de workflows de FL segundo diferentes dimensões relacionadas ao desempenho. Atualmente, os critérios de filtragem suportados incluem makespan do workflow (tempo total de execução), tempo de treinamento, tempo de carregamento de dados, tempo de avaliação e a métrica de avaliação inerente do modelo em uso (*e.g.*, o silhouette score).

Embora essa visão se concentre em dados de proveniência de alto nível, ela fornece uma caracterização inicial da escala, do comportamento de execução e da estabilidade do workflow de FL. Por exemplo, os usuários podem rapidamente distinguir uma execução em larga escala e estável envolvendo 20 clientes ao longo de 30 rodadas de comunicação de uma execução menor envolvendo apenas 2 clientes e 5 rodadas, antes de realizar uma inspeção mais detalhada dos registros de proveniência subjacentes.

5.3.2 Round/Client Overview

O Round/Client Overview (Figura 8) fornece proveniência em nível de rodada e de cliente. Cada faixa horizontal representa um cliente participante ou o servidor central, enquanto o eixo horizontal codifica as rodadas de treinamento (área ❶ na Figura 8 e Figura 9 com mais detalhes). Cada faixa é segmentada em fatias — uma por rodada em que o cliente ou servidor correspondente esteve ativo —, de modo que lacunas em uma faixa revelam imediatamente churns, ingressos tardios ou rodadas puladas devido a políticas de seleção de clientes. Dois modos de codificação, selecionáveis por um menu suspenso, controlam como cada fatia é renderizada. No modo *largura*, a espessura de uma fatia codifica a duração da atividade correspondente, *i.e.*, tempo de treinamento local para clientes ou tempo de agregação para o servidor, de modo que faixas mais largas correspondem a rodadas mais lentas. A duração pode ser normalizada de duas formas: (i) *Global*, que escala as larguras usando as durações mínima e máxima observadas entre todos os clientes e todas as rodadas, permitindo comparar o comportamento de um cliente em relação a todo o workflow; e (ii) *Por rodada*, que normaliza apenas em relação aos clientes que

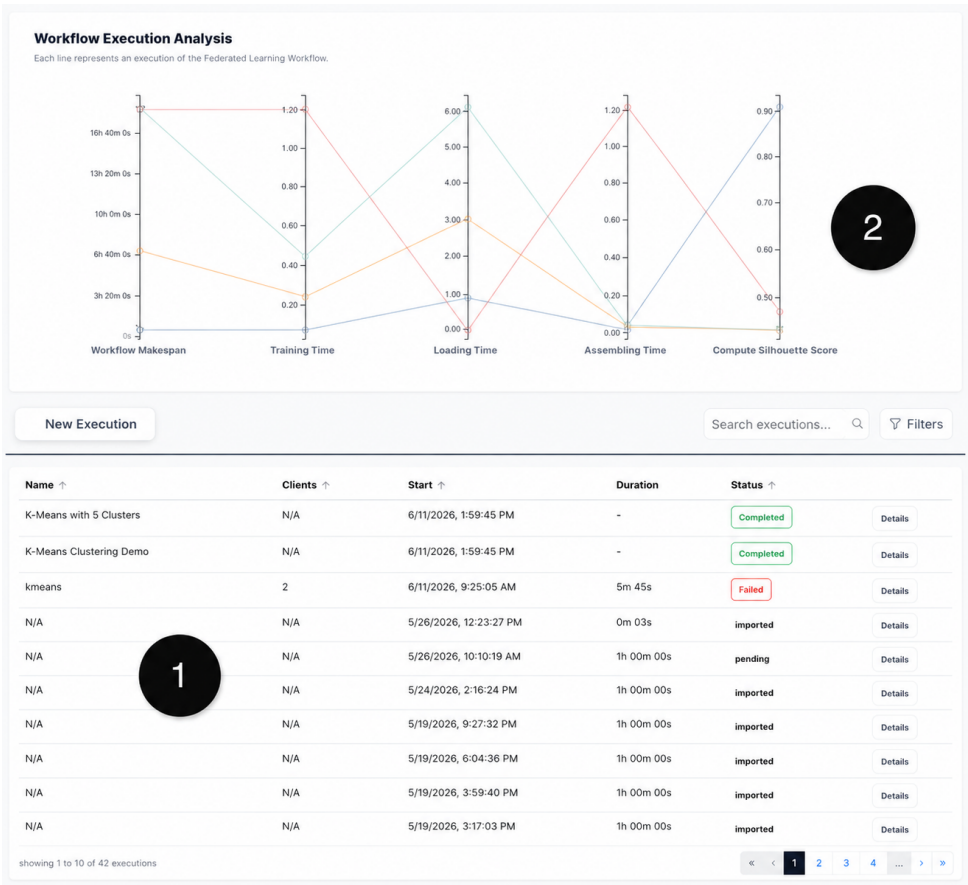


Figura 7: Interface de visão geral do workflow da FedObserve.

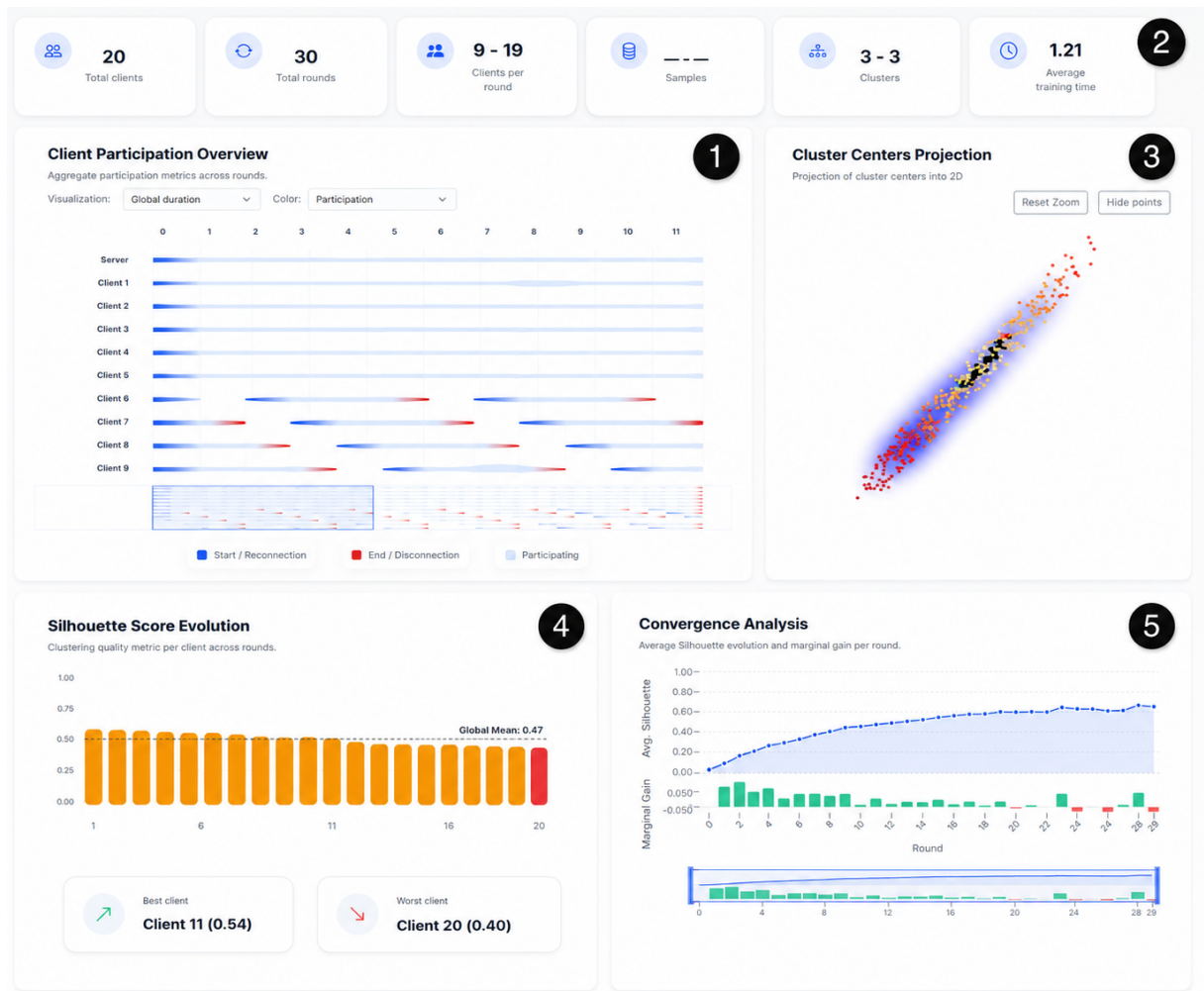


Figura 8: Interface de análise da FedObserve. (1) Round/Client Overview, (2) Resumo de estatísticas, (3) Model Training Overview, (4) Quality Overview e (5) Convergence Analysis.

participaram dessa rodada específica de treinamento, isolando stragglers relativos dentro de uma rodada, independentemente da distribuição de tempo de execução do workflow inteiro.

No modo *cor*, as fatias são coloridas de acordo com uma métrica analítica. Dois esquemas de cor são atualmente suportados na FedObserve: (i) *Participação*, que marca o início da participação de um cliente (ou sua reconexão após um churn) em azul escuro e o fim de sua participação em vermelho; e (ii) *Silhouette Score*, que mapeia cada fatia ao valor de qualidade de clusterização da rodada usando uma escala sequencial de vermelho a verde, permitindo que degradações de desempenho sejam apresentadas diretamente na linha do tempo de participação.

Com relação à exploração de metadados, os usuários podem selecionar interativamente

uma rodada individual, um cliente específico, múltiplas rodadas, múltiplos clientes ou combinações arbitrárias de clientes e rodadas diretamente a partir das faixas horizontais (área ❶ na Figura 8). A cada seleção, os metadados de resumo exibidos no topo da interface (área ❷ na Figura 8) são atualizados para refletir o subconjunto selecionado de dados de proveniência. As estatísticas exibidas incluem o número de clientes selecionados, o número de rodadas, o tempo médio de treinamento local, *etc.* Essa interação coordenada permite aos usuários transitar de uma visão global do workflow de FL para análises focadas de subconjuntos específicos de execução. Ao consultar dinamicamente estatísticas resumidas do banco de dados de proveniência para a seleção atual, a *FedObserve* facilita análises comparativas entre clientes e rodadas.

A figura 9 mostra a visão de Participação de Clientes. Cada faixa horizontal representa o servidor ou um cliente participante, e cada fatia corresponde a uma rodada de treinamento. A linha do tempo fornece proveniência em nível de rodada e cliente, revelando participação de clientes, churn, ingressos tardios e rodadas puladas. As fatias podem ser codificadas por largura para representar a duração da atividade (normalizada global ou por rodada) ou por cor para visualizar eventos de participação ou métricas analíticas como o silhouette score.

Adicionalmente, para suportar workflows com muitas rodadas, o painel inclui um *minimap* horizontal posicionado abaixo da visão principal (parte inferior da área na Figura 9). O *minimap* renderiza uma visão comprimida de toda a linha do tempo e expõe uma janela de seleção arrastável e redimensionável. Ao selecionar essa janela, a visão principal é atualizada para focar nas rodadas selecionadas, seguindo um padrão de visão geral mais detalhe que mantém o padrão global de participação visível enquanto o analista faz zoom em um intervalo específico.

5.3.3 Model Training Overview

O Model Training Overview (área ❸ na Figura 8 e Figura 10 com mais detalhes) fornece uma representação visual de um dos aspectos mais importantes do FL: a relação entre os modelos treinados localmente e o modelo global agregado. Essa visão permite aos usuários avaliar o quanto o modelo local de cada cliente diverge do modelo global durante o treinamento, fornecendo insights sobre convergência do modelo, heterogeneidade dos clientes e a eficácia geral do processo de otimização federada. Dependendo do subconjunto selecionado de clientes e rodadas, a interface compara dinamicamente os modelos locais correspondentes com o modelo global associado, permitindo verificar se o processo de trei-

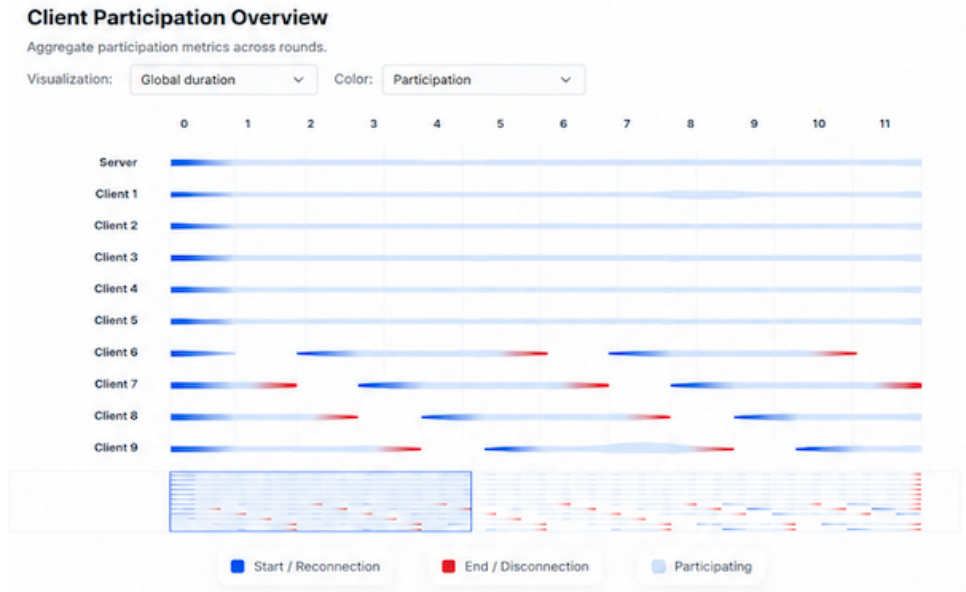


Figura 9: Visão de Participação de Clientes.

namento evolui conforme o esperado ou se clientes específicos apresentam comportamento anormal.

Fornecer tais comparações é desafiador porque os frameworks de FL suportam uma ampla variedade de modelos de ML, cada um com sua própria representação (*e.g.*, pesos de redes neurais, árvores de decisão, centróides de clusters ou vetores de embedding). Consequentemente, o componente de visualização é intencionalmente projetado para ser extensível, permitindo que a estratégia de comparação seja customizada de acordo com o modelo de aprendizado empregado pelo workflow de FL analisado, preservando um paradigma de interação consistente entre aplicações.

Embora o esquema de proveniência da FedObserve seja independente de framework e possa ser acoplado a diferentes workflows de FL, as visualizações analíticas não são totalmente neutras ao tipo de algoritmo. Em particular, parte das visualizações apresentadas neste trabalho foi instanciada para um cenário de aprendizado não supervisionado (clusterização federada), no qual a comparação entre modelos locais e global é realizada sobre centróides e a qualidade é resumida pelo *silhouette score*. Em workflows supervisionados, nos quais o comportamento do modelo é normalmente acompanhado por métricas como acurácia, perda e precisão, os mesmos mapeamentos visuais podem não preservar a melhor semântica analítica. Dessa forma, a portabilidade entre frameworks é obtida no nível de captura e integração da proveniência, mas a camada de visual analytics requer especializações por família de algoritmo e por objetivo de análise.



Figura 10: Projeção de modelos locais e globais de clusterização em espaço bidimensional via t-SNE.

O exemplo mostrado na Figura 10 corresponde a uma aplicação de clusterização federada. Nesse caso, a visualização compara os centróides de clusters produzidos pelo treinamento local de cada cliente com os centróides do modelo global agregado. Para cada rodada de treinamento, as matrizes de centróides dos modelos local e global (de tamanho $N \times M$, onde N é o número de clusters e M é o número de atributos) são achatadas em vetores de comprimento fixo. Esses vetores são subsequentemente projetados em um espaço 2D usando t-SNE, que preserva as relações de vizinhança da representação original de alta dimensão. A visualização resultante permite aos usuários avaliar intuitivamente a proximidade entre modelos locais e globais, facilitando a identificação de clientes outliers, comportamentos de treinamento divergentes e padrões de convergência ao longo do processo de aprendizado federado.

O gráfico de dispersão resultante (Figura 10) renderiza a trajetória do modelo global do servidor como uma spline suave de Catmull-Rom conectando suas posições sucessivas rodada a rodada, enquanto os centróides individuais dos clientes são desenhados como pontos ao redor dessa trajetória. A distância cliente-servidor é redundantemente codificada por dois canais: uma escala de cores sequencial amarelo-laranja-vermelho, na qual pontos mais escuros e avermelhados indicam clientes cujo modelo local diverge mais fortemente do modelo agregado; e clientes cuja distância ao centróide do grupo excede $2,5 \times \sigma$,

onde σ é o desvio padrão das distâncias cliente-centróide naquela rodada, são sinalizados como outliers estatísticos e renderizados com raio aumentado e contorno destacado, chamando atenção para clientes que podem ser afetados por dados não-IID, stragglers, etc.

Algumas alterações entre os gráficos *Cluster Center Projection* e *Client Participation Overview* são possíveis. Ao clicar em fatias do gráfico *Cluster Center Projection* ocorrem filtros no gráfico *Client Participation Overview*. A área ❶ da figura 11 mostra o gráfico *Client Participation Overview* sem nenhuma seleção. Já a área ❷ da figura 11 mostra a seleção de um único cliente no gráfico *Client Participation Overview* enquanto as áreas ❸ e circledblack da figura 11 mostra seleções aleatórias no *Client Participation Overview*. Todas as seleções fazem com que apenas os pontos que correspondem apenas aos clientes/rounds selecionados se destaquem no gráfico *Cluster Center Projection*.

A Figura 11 ilustra a interação coordenada entre os gráficos *Client Participation Overview* e *Cluster Center Projection*. Essa interação permite que seleções realizadas no *Client Participation Overview* filtrem o subconjunto de elementos destacados no *Cluster Center Projection*, facilitando a análise conjunta entre padrões de participação e a posição relativa dos modelos locais em relação ao modelo global. Na área ❶, apresenta-se a visualização sem qualquer seleção, exibindo o conjunto completo de clientes e rodadas. Na área ❷, observa-se a seleção de um único cliente no *Client Participation Overview* em todos os rounds em que participou, fazendo com que apenas os pontos correspondentes a esse cliente permaneçam destacados no *Cluster Center Projection*. Já as áreas ❸ e ❹ mostram seleções envolvendo diferentes subconjuntos de clientes e rodadas, evidenciando que o destaque no *Cluster Center Projection* é atualizado de acordo com o contexto selecionado. A seleção é realizada com um clique no rótulo do cliente para selecionar todos os pontos do cliente, com um clique no rótulo do round para selecionar todos os clientes que participaram do round selecionado ou com um clique em fatias aleatórias que representam clientes em um round específico. Em conjunto, essa coordenação entre visualizações permite relacionar eventos de participação, churn e recorrência temporal com padrões de convergência, divergência e possíveis comportamentos anômalos dos clientes ao longo do workflow de FL.

5.3.4 Model Quality/Convergence Overview

Enquanto o Model Training Overview aborda *onde* os modelos dos clientes estão posicionados em relação ao modelo global do servidor (e.g., o espaço de centróides), ele não

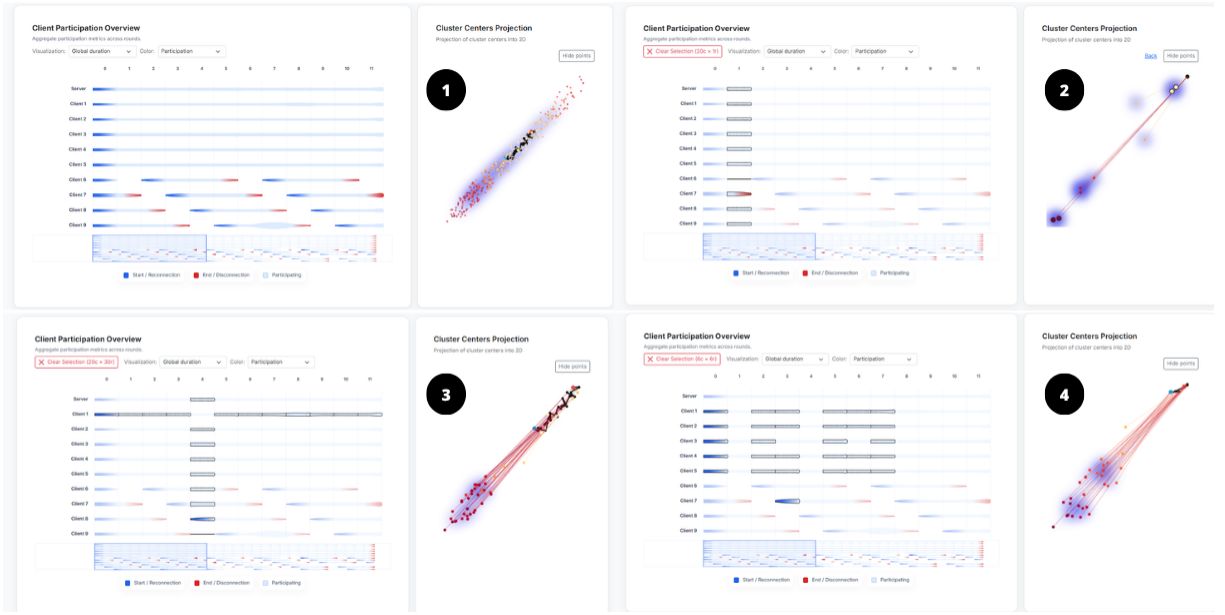


Figura 11: Interações entre os gráficos *Cluster Center Projection* e *Client Participation Overview*

responde uma pergunta mais simples mas igualmente importante: *quão bom* é o modelo resultante (*e.g.*, o cluster) e ele está melhorando ao longo do tempo? A *FedObserve* responde a essa pergunta por meio de dois painéis complementares que compartilham o silhouette score como métrica comum de qualidade, mas segmentam os dados de proveniência ao longo de dimensões ortogonais. O painel *Silhouette Score Evolution* (área ④ na Figura 8) é *centrado no cliente*: agrega o desempenho de cada cliente ao longo de todo o workflow e classifica os clientes entre si, suportando perguntas como “quais clientes estão com desempenho abaixo do esperado e em que medida?”. O painel *Convergence Analysis* (área ⑤ na Figura 8 e Figura 12 com mais detalhes) é, por sua vez, *centrado na rodada*: agrega o desempenho entre os clientes para cada rodada e o acompanha ao longo do eixo temporal do workflow, suportando perguntas como “o pool de clientes está convergindo e quando se estabilizou?”.

Painel de Evolução de Qualidade. Este painel (área ④ na Figura 8) suporta a avaliação comparativa da qualidade do modelo em nível de cliente (qualidade de clusterização na Figura 8). Cada barra vertical representa o silhouette score médio de um cliente ao longo de todas as rodadas em que participou, com os clientes ordenados ao longo do eixo horizontal pelo identificador do cliente e, ao passar o mouse, re-ordenáveis por pontuação. Uma linha de referência horizontal tracejada marca o silhouette score médio global do workflow, facilitando a identificação de clientes com desempenho acima ou abaixo da média da federação. O painel também exibe dois indicadores resumidos, *Melhor cliente*

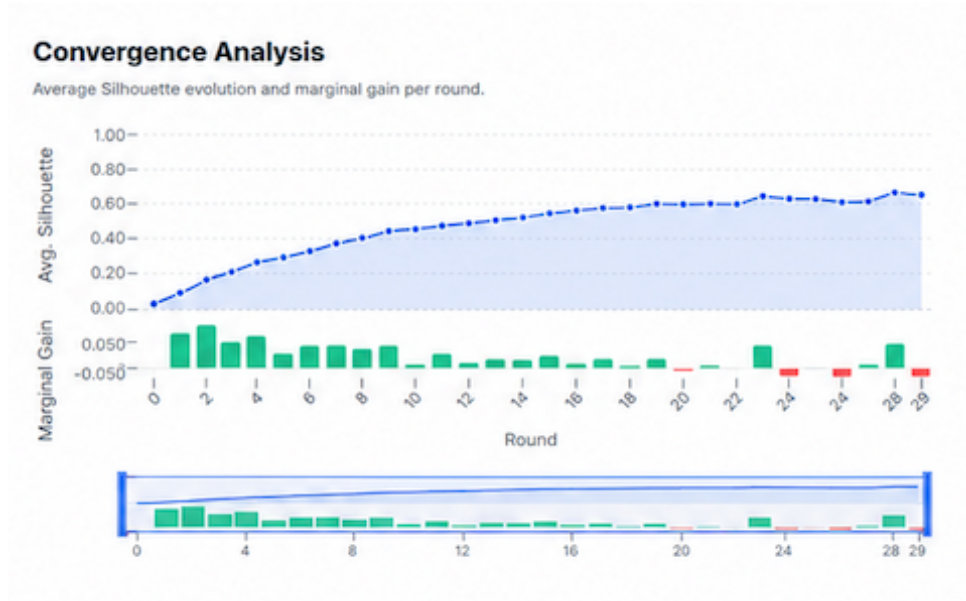


Figura 12: Painel de Análise de Convergência.

e *Pior cliente*, extraídos diretamente dessa classificação, fornecendo aos usuários uma âncora imediata para investigações mais aprofundadas e específicas por cliente.

Painel de Análise de Convergência. Este painel (Figura 12) é composto por dois gráficos empilhados verticalmente e sincronizados no tempo. O gráfico superior plota a evolução do silhouette score médio por rodada como um gráfico de linha e área, expondo a tendência global de melhora (ou degradação) da qualidade do modelo ao longo do workflow de FL. O gráfico inferior reporta o ganho marginal rodada a rodada, $\Delta_n = score_n - score_{n-1}$, como um gráfico de barras no qual deltas positivos são renderizados em verde e deltas negativos em vermelho. Essa decomposição separa duas dimensões analíticas complementares: (i) o *nível absoluto de qualidade* do modelo agregado (gráfico superior) e (ii) sua *estabilidade de aprendizado* (gráfico inferior). Na prática, rodadas consecutivas com Δ_n próximo de zero indicam que o treinamento se estabilizou, enquanto oscilações persistentes de sinal ou deltas de grande magnitude sugerem um modelo que ainda não convergiu ou é sensível à heterogeneidade dos clientes. A leitura conjunta de ambos os gráficos evita interpretações ambíguas que uma única métrica produziria isoladamente: um silhouette score médio elevado pode coexistir com instabilidade local (deltas oscilantes), enquanto uma pontuação moderada pode, de fato, representar um processo bem convergido e estável quando os deltas se aproximam de zero.

De forma análoga ao Round/Client Overview, o painel de Análise de Convergência integra um mecanismo sincronizado de *minimap* e zoom para suportar a exploração de execuções de longa duração. O usuário primeiro inspeciona o padrão global de conver-

gência no *minimap* comprimido e, em seguida, seleciona um intervalo de interesse, *e.g.*, rodadas 1–29 na parte inferior da Figura 12, para aplicar zoom simultâneo nos gráficos de evolução do silhouette e de ganho marginal. Essa sincronização permite verificar se uma aparente estabilização observada na tendência agregada é um platô genuíno ou apenas um artefato visual da linha do tempo comprimida.

6 Avaliação Experimental

Esta seção apresenta a avaliação da **FedObserve**. O objetivo é avaliar a eficácia do framework proposto em suportar observabilidade orientada por proveniência de workflows de FL. Em particular, investigamos se a **FedObserve** permite que os usuários monitorem a execução do workflow, identifiquem gargalos de desempenho, analisem a participação dos clientes, inspecionem a convergência do modelo e comparem diferentes execuções de workflow usando dados de proveniência capturados automaticamente durante a execução. Para tanto, especialistas em FL avaliaram a **FedObserve** sob duas perspectivas complementares. A primeira focou no monitoramento em tempo de execução de uma execução de workflow de FL em pequena escala, enquanto a segunda investigou a análise retrospectiva de uma coleção de workflows de FL em larga escala executados anteriormente. Em conjunto, esses cenários avaliam a capacidade da **FedObserve** de suportar tanto a observabilidade em tempo de execução quanto a exploração *post-mortem* de proveniência. O restante desta seção está organizado da seguinte forma. A Subseção 6.1 apresenta o estudo de caso e o workflow de FL alvo. A Subseção 6.2 detalha o protocolo de avaliação adotado no estudo de usuário. Por fim, a Subseção 6.3 discute os feedbacks dos especialistas.

6.1 Estudo de Caso: Detecção Federada de Outliers

A avaliação da **FedObserve** foi conduzida usando um workflow federado de clusterização *k*-means como estudo de caso para clusterização de objetos astronômicos em catálogos de grande escala, seguindo a abordagem proposta por [Lopes, Ferreira *et al.* \(2025\)](#). O workflow é executado no NVIDIA FLARE e instrumentado para capturar dados de proveniência automaticamente, que são então armazenados no banco de dados de proveniência da **FedObserve**. Cada cliente armazena um fragmento do catálogo astronômico DES DR2 ([Abbott, Adamów *et al.* \(2021\)](#)) correspondente a uma região diferente do céu, enquanto o servidor coordena o workflow de FL sem acessar os dados astronômicos brutos subjacentes.

Cada cliente pré-processa independentemente seu conjunto de dados local selecionando os atributos fotométricos `mag_auto_g_dered`, `mag_auto_r_dered`, `mag_auto_i_dered`, `mag_auto_z_dered` e `mag_auto_y_dered`, e calculando as features de cor $g - r$, $r - i$ e $i - z$. Essas features definem o espaço de entrada para a detecção de outliers baseada em clusterização. Como cada cliente processa apenas sua região local do céu, nem os objetos astronômicos nem as features derivadas são trocados ao longo da execução.

Neste estudo de caso, o algoritmo K-Means ([MacQueen \(1967\)](#)) é empregado como algoritmo não supervisionado subjacente. O treinamento local consiste em estimar centróides de clusters a partir do conjunto de dados do cliente, após o qual apenas os parâmetros do modelo (centróides de clusters e metadados associados) são enviados de volta ao servidor. O servidor agrega as atualizações recebidas para produzir um novo modelo global, que é redistribuído aos clientes para a próxima rodada de treinamento. Após a convergência, cada cliente aplica o modelo global final de clusterização aos seus dados locais. Cada objeto astronômico é atribuído ao centróide de cluster mais próximo e sua distância ao centróide correspondente é calculada. Objetos com distâncias inusualmente grandes são considerados candidatos a outliers, pois estão em regiões esparsas do espaço de features, em vez de estarem dentro de clusters densos que representam as populações astronômicas dominantes.

6.2 Protocolo de Avaliação

Como mencionado anteriormente, esta avaliação investiga se a `FedObserve` permite que os usuários monitorem a execução do workflow, explorem dados de proveniência em diferentes níveis de granularidade, identifiquem gargalos de execução, analisem a participação dos clientes, inspecionem a convergência do modelo e comparem múltiplas execuções de workflows de FL. A avaliação foca nas capacidades de observabilidade fornecidas pela `FedObserve`, e não no desempenho ou na acurácia do workflow de FL subjacente. A avaliação foi projetada em torno de duas análises complementares. A primeira avalia a `FedObserve` como ferramenta de monitoramento em tempo de execução, onde a proveniência é capturada e visualizada continuamente enquanto um workflow de FL está em execução. A segunda avalia suas capacidades para análise retrospectiva, na qual a proveniência coletada de execuções anteriores de workflows em larga escala é explorada para investigar o comportamento de execução e identificar padrões relevantes. Vale ressaltar que a avaliação com K-Means federado constitui uma prova de conceito (proof of concept) da abordagem; a generalização para outros algoritmos de FL, como modelos baseados em

deep learning, é deixada como trabalho futuro.

Cenário 1 – Monitoramento em Tempo de Execução. O primeiro cenário avalia a capacidade da **FedObserve** de suportar o monitoramento em tempo de execução de um workflow de FL em pequena escala. Os avaliadores iniciaram uma nova execução de um workflow de FL e monitoraram a execução por meio do dashboard de visual analytics à medida que os dados de proveniência se tornavam disponíveis. Durante a execução, os avaliadores foram incentivados a inspecionar a evolução da proveniência em nível de workflow, de rodada e de cliente. O workflow executado neste cenário foi intencionalmente configurado com uma escala reduzida para facilitar a inspeção da proveniência à medida que novas rodadas eram concluídas. A execução usou três clusters, cinco rodadas e dois clientes.

No cenário de monitoramento online, a atualização do dashboard foi configurada com periodicidade de 1 minuto, de modo que os painéis eram recalculados e re-renderizados a cada nova janela temporal de coleta. O monitoramento era encerrado quando ao menos um dos seguintes critérios de parada era satisfeito: *(a)* término do workflow (**COMPLETED**), *(b)* falha da execução (**FAILED**), ou *(c)* alcance do número máximo de rodadas configurado para o experimento.

Cenário 2 – Análise Retrospectiva do Workflow. O segundo cenário avalia as capacidades exploratórias da **FedObserve** usando proveniência coletada de um workflow de FL em larga escala executado anteriormente. Ao contrário do cenário anterior, os avaliadores analisaram registros históricos de proveniência após a conclusão do workflow, permitindo exploração irrestrita do histórico de execução. Este cenário investiga se a **FedObserve** suporta a compreensão do comportamento do workflow, a comparação de execuções dos clientes, a inspeção da evolução do modelo e a identificação de padrões de execução anômalos. O workflow analisado foi configurado com três clusters, 30 rodadas e 20 clientes participantes, produzindo um banco de dados de proveniência substancialmente maior, adequado para análise exploratória.

Avaliadores. A avaliação envolveu dois especialistas de domínio com ampla experiência em FL e familiaridade com os dados de proveniência gerados durante a execução de FL. Recrutar avaliadores para este estudo foi desafiador porque a avaliação exigiu indivíduos que possuísem simultaneamente expertise em algoritmos de FL, análise de modelos e interpretação de dados de proveniência. Porém, nenhum dos avaliadores tinha experiência prévia com sistemas de visual analytics orientados por proveniência. Consequentemente, o objetivo do estudo não foi obter resultados estatisticamente representativos, mas sim

coletar feedback qualitativo aprofundado de usuários especializados capazes de avaliar criticamente a **FedObserve** em cenários realistas. Essa abordagem é consistente com recomendações na pesquisa de visualização e IHC, onde avaliações baseadas em especialistas são consideradas apropriadas quando a população-alvo de usuários é pequena, especializada e o foco está em avaliar utilidade e relevância de domínio, em vez de generalização estatística (Lam *et al.* (2012)). Antes das sessões de avaliação, os avaliadores receberam um tutorial introduzindo a arquitetura da **FedObserve**, o modelo de proveniência adotado e as visualizações disponíveis. O tutorial focou exclusivamente em explicar a funcionalidade da **FedObserve** e não incluiu nenhuma orientação sobre as análises esperadas durante a avaliação.

Tarefas de Avaliação. Os avaliadores foram solicitados a completar duas tarefas correspondentes aos cenários de avaliação: (i) iniciar um novo workflow Federado de k -Means por meio da **FedObserve** e monitorar sua execução enquanto os dados de proveniência eram continuamente coletados e incorporados ao dashboard de visual analytics; e (ii) analisar um workflow Federado de k -Means em larga escala executado anteriormente usando a interface de visual analytics e reportar observações relevantes sobre execução do workflow, comportamento dos clientes, desempenho de execução, rodadas de comunicação e convergência do modelo. A segunda tarefa adotou intencionalmente um protocolo exploratório aberto em vez de prescrever uma sequência fixa de interações. Essa escolha de design reflete a natureza exploratória da análise de proveniência, na qual os analistas formulam hipóteses iterativamente e navegam entre múltiplos níveis de abstração ao investigar o comportamento do workflow.

Crítérios de Avaliação. Durante o experimento, os avaliadores foram solicitados a investigar se a **FedObserve** lhes permitia identificar características relevantes de execução, incluindo: (i) gargalos de execução do workflow, como aumentos no tempo de treinamento do cliente, tempo de agregação do servidor, tempo de carregamento de dados ou tempo de avaliação; (ii) padrões de participação de clientes, incluindo participação intermitente, participação persistente e assimetrias de comunicação entre rodadas; (iii) heterogeneidade de clientes, comparando características de execução e comportamento do modelo entre diferentes clientes participantes; (iv) convergência do modelo, inspecionando a relação entre modelos locais e o modelo global correspondente ao longo das rodadas de comunicação; e (v) diferenças entre execuções de workflow por meio da análise de resumos de proveniência em nível de workflow e estatísticas de execução. Em vez de exigir que os avaliadores respondessem perguntas predefinidas, a avaliação incentivou a exploração livre dos dados de proveniência.

Coleta e Análise de Dados. Cada sessão de avaliação foi gravada em tela para capturar as interações dos avaliadores com a **FedObserve**. Os avaliadores foram incentivados a verbalizar seu raciocínio ao longo da avaliação usando um protocolo think-aloud, ao mesmo tempo em que documentavam os principais insights obtidos a partir dos dados de proveniência. Após as sessões de avaliação, os vídeos gravados e os relatórios dos avaliadores foram analisados conjuntamente para verificar se a **FedObserve** suportava observabilidade orientada por proveniência de workflows de FL. Em particular, a análise focou em verificar se os avaliadores conseguiam identificar gargalos de execução, padrões de participação dos clientes, heterogeneidade de clientes, comportamento de convergência do modelo e diferenças de execução exclusivamente por meio das informações de proveniência capturadas e visualizadas pelo framework proposto. Os resultados desta avaliação são apresentados e discutidos na subseção a seguir.

6.3 Feedback dos Especialistas

Esta seção discute os achados obtidos nos dois cenários de avaliação descritos na Seção 6.2. A discussão considera primeiro o cenário de monitoramento online, seguido pela análise retrospectiva de acordo com as perspectivas analíticas introduzidas na Seção 6.2, e conclui com observações de usabilidade e implicações para a observabilidade orientada por proveniência em workflows de FL. Ao longo desta seção, os dois avaliadores são referidos como E1 e E2.

Monitoramento Online do Workflow. O primeiro cenário de avaliação investigou se a **FedObserve** suporta monitoramento orientado por proveniência enquanto um workflow de FL está em execução. Embora as execuções do workflow tenham sido concluídas com sucesso, os dados de proveniência chegaram de forma assíncrona, resultando em pequenas lacunas temporais no Round/Client Overview. Essas lacunas são esperadas porque os eventos de proveniência são coletados independentemente da janela de atualização da visualização. Apesar dessa limitação, a **FedObserve** permaneceu responsiva ao longo da execução, demonstrando que a proveniência pode ser incorporada incrementalmente sem impedir os avaliadores de inspecionar resultados parciais de execução.

Inspeção em Nível de Workflow. Ambos os avaliadores usaram o Workflow Overview como ponto de entrada inicial para investigar o comportamento do workflow. A visualização expôs imediatamente diferenças no makespan entre execuções com configurações similares, permitindo que os avaliadores priorizassem execuções que mereciam inspeção

mais aprofundada. Um avaliador identificou ainda duas execuções com tempo total de execução quase idêntico, mas com distribuições marcadamente diferentes de custos de carregamento, treinamento e agregação. Em contraste, outra execução alcançou um silhouette score comparável com menos da metade do tempo de execução. Essas observações indicam que o custo de execução por si só é insuficiente para explicar a qualidade do modelo e ilustram a importância de inspecionar conjuntamente a proveniência em nível de workflow e as métricas de qualidade ao comparar execuções de FL. Ambos os avaliadores concordaram que o Workflow Overview suporta efetivamente a triagem inicial do workflow antes de avançar para análises de granularidade mais fina. Os principais problemas de usabilidade reportados estavam relacionados à interação, e não à capacidade analítica. Em particular, a ausência de uma legenda no gráfico de coordenadas paralelas dificultava associar polilinhas individuais às suas execuções correspondentes quando o visual se tornava denso, enquanto a falta de indicação explícita de se as métricas de tempo representavam valores cumulativos ou por rodada ocasionalmente introduzia ambiguidade durante a avaliação.

Comportamento de Participação dos Clientes. Ambos os avaliadores identificaram com sucesso a participação heterogênea dos clientes no workflow de 30 rodadas e 20 clientes. O Round/Client Overview distinguiu claramente os clientes que participaram ao longo de toda a execução daqueles com disponibilidade intermitente, refletindo a participação heterogênea tipicamente observada em implantações de FL. Além dos padrões de participação, a visualização possibilitou a identificação de anomalias de desempenho localizadas. Um avaliador observou que o cliente 7 consistentemente exigiu tempos de treinamento mais longos do que os demais clientes participantes, enquanto vários picos isolados apareceram em combinações específicas de cliente-rodada sem se propagar para rodadas ou clientes vizinhos. Em contraste, o tempo de agregação do servidor permaneceu estável ao longo da execução do workflow. A visualização coordenada permitiu, portanto, que os avaliadores distinguíssem anomalias de execução localizadas de gargalos sistêmicos sem inspecionar logs de execução brutos. Embora ambos os avaliadores tenham chegado a conclusões similares sobre a heterogeneidade de participação dos clientes, sugeriram melhorias para reduzir a ambiguidade interpretativa. Em particular, observaram que o Round/Client Overview deveria indicar explicitamente se os clientes exibidos correspondem ao pool completo de clientes ou apenas ao subconjunto disponível dentro da janela de análise selecionada.

Análise do Treinamento do Modelo. O Model Training Overview permitiu que os avaliadores investigassem as relações entre os modelos locais e globais ao longo da execução.

Ambos os avaliadores identificaram um pequeno subconjunto de clientes que permaneceu consistentemente distante do centróide global durante múltiplas rodadas. Curiosamente, esses clientes apresentaram simultaneamente melhorias graduais no silhouette score, sugerindo que sua divergência estrutural originou-se de características intrínsecas de seus conjuntos de dados locais, e não de comportamento de treinamento instável. As visões coordenadas também incentivaram o raciocínio entre perspectivas. Um avaliador investigou se os clientes que apresentavam maiores desvios de centróide também correspondiam àqueles com atrasos de execução identificados no Round/Client Overview. A associação observada entre essas duas perspectivas sugere que a heterogeneidade espacial e o desempenho de execução representam características em grande parte dependentes. Isso ilustra um dos principais benefícios de diferentes níveis de proveniência, permitindo que os usuários não apenas formulem hipóteses, mas também confirmem ou refutem potenciais relações entre diferentes perspectivas.

Avaliação de Convergência e Qualidade. A FedObserve suportou consistentemente a análise de convergência do modelo. Ambos os avaliadores observaram um aumento suave no silhouette score global ao longo das rodadas de comunicação, acompanhado de deltas de convergência predominantemente positivos e apenas pequenas flutuações locais. A ausência de oscilações abruptas indicou convergência estável do workflow avaliado. No nível de cliente, o ranking de silhouette score destacou um pequeno subconjunto de clientes participantes com qualidade média inferior ou maior variabilidade em relação ao restante do pool de clientes, permitindo a identificação de clientes potencialmente instáveis para investigação mais aprofundada. Mais importante, os avaliadores enfatizaram que combinar a tendência de convergência com a evolução da qualidade em nível de cliente facilitava distinguir um treinamento em melhoria ativa de um já estabilizado. A principal preocupação de usabilidade nesta análise envolveu a visualização de ranking, cujo eixo horizontal representa posição de ranking em vez de identificadores de cliente. Os avaliadores reportaram que exibir diretamente os identificadores de cliente simplificaria a interpretação dos rankings individuais.

Feedback de Usabilidade. De modo geral, ambos os avaliadores consideraram a FedObserve eficaz para a observabilidade orientada por proveniência de workflows de FL. A Tabela 2 resume a avaliação quantitativa de usabilidade. As limitações reportadas referem-se principalmente a refinamentos de interação, e não a capacidades analíticas ausentes. Em ambas as avaliações, os avaliadores sugeriram consistentemente adicionar uma legenda ao Workflow Overview, ampliar as regiões sensíveis ao hover em visualizações densas, esclarecer a semântica das métricas de tempo, substituir posições de ranking

por identificadores de cliente, habilitar destaque coordenado entre visualizações, corrigir os controles de visibilidade da projeção de centróides de cluster, suportar visualização em tela cheia para gráficos densos e permitir a exportação de dados de proveniência para análises externas.

Em geral, a avaliação demonstra que os avaliadores conseguiram identificar gargalos de workflow, padrões heterogêneos de participação, diferenças estruturais entre clientes, comportamento de convergência e trade-offs de execução exclusivamente por meio dos dados de proveniência capturados e visualizados pela **FedObserve**. Embora várias melhorias de interação tenham sido identificadas, elas representam refinamentos incrementais, e não limitações da abordagem proposta, indicando que a **FedObserve** já fornece suporte efetivo para a observabilidade orientada por proveniência de workflows de FL.

Tabela 2: Resultados da avaliação de usabilidade, incluindo notas e observações qualitativas para as dimensões avaliadas.

Dimensão	Nota	Observação
Facilidade de navegação	4/5	A interface foi geralmente fácil de navegar. O principal problema de usabilidade foi o pequeno alvo de hover no gráfico de coordenadas paralelas no Workflow Overview.
Clareza das métricas	3/5	Algumas métricas e rótulos eram ambíguos, particularmente a distinção entre valores cumulativos e por rodada e a interpretação dos rótulos de ranking.
Confiança nas conclusões	5/5	As visualizações coordenadas forneceram evidências suficientes para suportar conclusões analíticas com alta confiança.
Utilidade para investigação	4/5	A FedObserve permitiu a identificação eficiente de achados relevantes que teriam sido substancialmente mais difíceis de obter inspecionando registros de proveniência diretamente.

7 Conclusão e Trabalhos Futuros

Este manuscrito apresentou a **FedObserve**, uma abordagem de observabilidade orientada por proveniência que combina captura de proveniência, gestão de dados e visual analytics para suportar a análise de workflows de FL. Ao contrário de abordagens existentes que se baseiam em logs específicos de framework ou mecanismos isolados de monitoramento, a **FedObserve** adota a proveniência como representação unificada da execução de FL, permitindo que eventos de workflow, comportamento dos clientes, métricas de execução e evolução do modelo sejam integrados em um modelo de dados comum e consultável. Ao combinar dados de proveniência com visual analytics, a abordagem permite que os usuários transitem fluidamente de resumos em nível de workflow para investigações detalhadas em nível de rodada e cliente, preservando a linhagem de execução necessária para análise reprodutível.

Do ponto de vista de gestão de dados, a principal contribuição deste manuscrito é demonstrar que os dados de proveniência podem servir não apenas como um registro de execução, mas também como a base para observabilidade em FL distribuído. O esquema de proveniência proposto captura a estrutura lógica dos workflows de FL independentemente de um framework de execução específico, enquanto a arquitetura modular desacopla captura, armazenamento, consulta e visualização de proveniência. Essa separação viabiliza extensibilidade entre diferentes ecossistemas de FL e fornece um substrato analítico comum para monitoramento, diagnóstico e análise comparativa de execuções federadas.

A avaliação experimental demonstrou a utilidade prática da **FedObserve**. Por meio da exploração orientada por proveniência, os avaliadores conseguiram identificar gargalos de execução, padrões de participação dos clientes, heterogeneidade espacial e comportamento de convergência diretamente a partir dos dados de proveniência coletados. A avaliação também mostrou que os usuários conseguiram navegar efetivamente entre visualizações complementares para correlacionar eventos de execução locais com o comportamento global do workflow, suportando o diagnóstico de fenômenos complexos de treinamento distribuído. Embora a avaliação tenha envolvido um número limitado de participantes, os

resultados fornecem evidências encorajadoras de que o visual analytics orientado por proveniência pode melhorar substancialmente a interpretabilidade e a explicabilidade das execuções de FL.

Embora este manuscrito demonstre a **FedObserve** usando um workflow de clusterização federada k -means, o modelo de proveniência subjacente é independente do algoritmo de aprendizado e pode acomodar metadados de execução adicionais e métricas específicas de modelo. Trabalhos futuros investigarão, portanto, o suporte a algoritmos de FL baseados em deep learning, visualizações mais ricas com suporte a proveniência e estudos de usuário em maior escala envolvendo praticantes com diferentes níveis de expertise. Também planejamos incorporar análises de proveniência em tempo real, incluindo detecção online de anomalias e geração de alertas, permitindo que a proveniência suporte não apenas a análise retrospectiva, mas também a tomada de decisão em tempo de execução em cenários operacionais de FL. De forma mais ampla, acreditamos que este manuscrito destaca o potencial da proveniência como uma abstração de primeira classe para observabilidade em sistemas distribuídos de IA. À medida que o FL continua a evoluir em direção a implantações cada vez mais heterogêneas e em larga escala, a gestão de dados orientada por proveniência fornece uma base principiada para construir ferramentas independentes de framework que suportem monitoramento, diagnóstico, reprodutibilidade e análise interativa ao longo do ciclo de vida completo de workflows de FL.

Como limitação, destacamos que, embora a integração de novos frameworks seja viabilizada por ajustes na captura de proveniência, parte das visualizações ainda depende da semântica do algoritmo e deve ser adaptada para cenários que possuem arquiteturas de modelo distintas.

REFERÊNCIAS

ABBOTT, T. M. C.; ADAMÓW, M. *et al.* The Dark Energy Survey Data Release 2. **The Astrophysical Journal Supplement Series**, American Astronomical Society, v. 255, n. 2, p. 20, jul. 2021. ISSN 1538-4365. DOI: [10.3847/1538-4365/ac00b3](https://doi.org/10.3847/1538-4365/ac00b3). Disponível em: <http://dx.doi.org/10.3847/1538-4365/ac00b3>.

ACERO-CUELLAR, Tatiana *et al.* **RTN-095: The Vera C. Rubin Observatory Data Preview 1**. [S. l.]: NSF-DOE Vera C. Rubin Observatory, 2025. DOI: [10.71929/RUBIN/2570536](https://doi.org/10.71929/RUBIN/2570536). Disponível em: <https://www.osti.gov/servlets/purl/2570536>.

ALPAYDIN, Ethem. **Introduction to Machine Learning**. 4. ed. [S. l.]: MIT Press, 2020.

ALVARENGA, Luis C. R. *et al.* Optimizing Resource Estimation for Scientific Workflows in HPC Environments: A Layered-Bucket Heuristic Approach. **Concurr. Comput. Pract. Exp.**, v. 37, n. 4-5, 2025. DOI: [10.1002/CPE.8381](https://doi.org/10.1002/cpe.8381). Disponível em: <https://doi.org/10.1002/cpe.8381>.

ALVES, A.A. *et al.* Deep learning identification of asteroids interacting with g-s secular resonances. **Planetary and Space Science**, v. 258, p. 106062, 2025. ISSN 0032-0633. DOI: <https://doi.org/10.1016/j.pss.2025.106062>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0032063325000297>.

BANDARA, E. *et al.* Bassa-ML — A Blockchain and Model Card Integrated Federated Learning Provenance Platform. *In*: IEEE 19th Annual Consumer Communications & Networking Conference (CCNC). [S. l.: s. n.], 2022. p. 753–759.

BEUTEL, Daniel J. *et al.* **Flower: A Friendly Federated Learning Research Framework**. [S. l.: s. n.], 2022. arXiv: [2007.14390](https://arxiv.org/abs/2007.14390) [cs.LG]. Disponível em: <https://arxiv.org/abs/2007.14390>.

BIRSAN, Dorian. On Plug-ins and Extensible Architectures. en. **ACM Queue**, Association for Computing Machinery (ACM), v. 3, n. 2, p. 40–46, mar. 2005.

BRAVEN, Tehmi E. den *et al.* Machine learning for precision medicine: promoting value considerations through perspective-taking hypothetical group design exercises. **AI**

Ethics, v. 6, n. 1, p. 127, 2026. DOI: [10.1007/S43681-025-00973-5](https://doi.org/10.1007/S43681-025-00973-5). Disponível em: <https://doi.org/10.1007/s43681-025-00973-5>.

BUTT, Anila Sahar; FITCH, Peter. ProvONE+: A Provenance Model for Scientific Workflows. *In: WEB Information Systems Engineering – WISE 2020*. Cham: Springer International Publishing, 2020. (Lecture Notes in Computer Science). p. 431–444.

CHAPMAN, Adriane; LAURO, Luca *et al.* Supporting Better Insights of Data Science Pipelines with Fine-grained Provenance. **ACM Transactions on Database Systems**, fev. 2024. Just Accepted. ISSN 0362-5915. DOI: [10.1145/3644385](https://doi.org/10.1145/3644385). Disponível em: <https://dl.acm.org/doi/10.1145/3644385>. Acesso em: 28 mar. 2024.

CHAPMAN, Adriane; MISSIER, Paolo *et al.* DPDS: Assisting Data Science with Data Provenance. *en. PVLDB*, v. 15, n. 12, p. 3614–3617, 2022. DOI: [10.14778/3554821.3554857](https://doi.org/10.14778/3554821.3554857). Disponível em: <https://vldb.org/pvldb/vol15/p3614-torlone.pdf>.

COSTA, Flavio *et al.* Capturing and querying workflow runtime provenance with PROV: a practical approach. *In: GUERRINI, Giovanna (ed.). Joint 2013 EDBT/ICDT Conferences, EDBT/ICDT '13, Genoa, Italy, March 22, 2013, Workshop Proceedings*. [S. l.]: ACM, 2013. p. 282–289. DOI: [10.1145/2457317.2457365](https://doi.org/10.1145/2457317.2457365). Disponível em: <https://doi.org/10.1145/2457317.2457365>.

CRAWL, Daniel; ALTINTAS, Ilkay. A Provenance-Based Fault Tolerance Mechanism for Scientific Workflows. *In: FREIRE, Juliana; KOOP, David; MOREAU, Luc (ed.). Provenance and Annotation of Data and Processes, Second International Provenance and Annotation Workshop, IPAW 2008, Salt Lake City, UT, USA, June 17-18, 2008. Revised Selected Papers*. [S. l.]: Springer, 2008. v. 5272. (Lecture Notes in Computer Science), p. 152–159. DOI: [10.1007/978-3-540-89965-5_17](https://doi.org/10.1007/978-3-540-89965-5_17). Disponível em: https://doi.org/10.1007/978-3-540-89965-5_17.

D'ISANTO, A. *et al.* An analysis of feature relevance in the classification of astronomical transients with machine learning methods. **Monthly Notices of the Royal Astronomical Society**, Oxford University Press (OUP), v. 457, n. 3, p. 3119–3132, fev. 2016. ISSN 1365-2966. DOI: [10.1093/mnras/stw157](https://doi.org/10.1093/mnras/stw157). Disponível em: <http://dx.doi.org/10.1093/mnras/stw157>.

EUROPEAN PARLIAMENT AND COUNCIL OF THE EUROPEAN UNION. **Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)**. [S. l.: s. n.], 2016. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>. Official Journal of the European Union, L119, 4 May 2016.

FATE-BOARD. **FATE-Board: A Suite of Visualization Tools for Federated Learning**. [S. l.: s. n.], 2026. GitHub repository. Accessed: 2026-07-20. Disponível em: <https://github.com/FederatedAI/FATE-Board>.

FREIRE, Juliana *et al.* Provenance for Computational Tasks: A Survey. **Computing in Science & Engineering**, v. 10, n. 3, p. 11–21, 2008. DOI: [10.1109/MCSE.2008.79](https://doi.org/10.1109/MCSE.2008.79).

GAMMA, Erich *et al.* **Design Patterns: Elements of Reusable Object-Oriented Software**. 1. ed. [S. l.]: Addison-Wesley Professional, 1994. ISBN 0201633612. Disponível em: http://www.amazon.com/Design-Patterns-Elements-Reusable-Object-Oriented/dp/0201633612/ref=ntt_at_ep_dpi_1.

GILL, Waris; ANWAR, Ali; GULZAR, Muhammad Ali. ProvFL: Client-Driven Interpretability of Global Model Predictions in Federated Learning. **CoRR**, abs/2312.13632, 2023. DOI: [10.48550/ARXIV.2312.13632](https://doi.org/10.48550/ARXIV.2312.13632). arXiv: [2312.13632](https://arxiv.org/abs/2312.13632). Disponível em: <https://doi.org/10.48550/arXiv.2312.13632>.

GILL, Waris; ANWAR, Ali; GULZAR, Muhammad Ali. TraceFL: Interpretability-Driven Debugging in Federated Learning via Neuron Provenance. *In: PROCEEDINGS of the IEEE/ACM 47th International Conference on Software Engineering*. Ottawa, Ontario, Canada: IEEE Press, 2025. (ICSE '25), p. 2264–2276. ISBN 9798331505691. DOI: [10.1109/ICSE55347.2025.00128](https://doi.org/10.1109/ICSE55347.2025.00128). Disponível em: <https://doi.org/10.1109/ICSE55347.2025.00128>.

GOOGLE RESEARCH. **TensorFlow Federated: Machine Learning on Decentralized Data**. [S. l.: s. n.], 2024. <https://www.tensorflow.org/federated>. Accessed: 2026-07-20.

GREGORI, Luca *et al.* Design and Development of a Provenance Capture Platform for Data Science. *In: PROCS. 3rd DATAPLAT workshop, co-located with ICDE 2024*. Utrecht, NL: IEEE, mai. 2024.

GU, Michael; NARAPARAJU, Ramasoumya; ZHAO, Dongfang. **Enhancing Data Provenance and Model Transparency in Federated Learning Systems – A Database Approach**. [S. l.: s. n.], 2024. arXiv: [2403.01451](https://arxiv.org/abs/2403.01451) [cs.CR]. Disponível em: <https://arxiv.org/abs/2403.01451>.

GUEDES, Thaylon *et al.* Provenance-based fault tolerance technique recommendation for cloud-based scientific workflows: a practical approach. **Clust. Comput.**, v. 23, n. 1, p. 123–148, 2020.

HAENLEIN, Michael; KAPLAN, Andreas. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. en. **Calif. Manage. Rev.**, SAGE Publications, v. 61, n. 4, p. 5–14, ago. 2019.

HAN, Xueyuan *et al.* FRAPpuccino: Fault-detection through Runtime Analysis of Provenance. *In: LARA, Eyal de; SUNDARARAMAN, Swaminathan (ed.). 9th USENIX Workshop on Hot Topics in Cloud Computing, HotCloud 2017, Santa Clara, CA, USA, July 10-11, 2017.* [S. l.]: USENIX Association, 2017.

Disponível em:

<https://www.usenix.org/conference/hotcloud17/program/presentation/han>.

HARD, Andrew *et al.* Federated Learning for Mobile Keyboard Prediction, 2018. arXiv preprint arXiv:1811.03604. DOI: [10.48550/arXiv.1811.03604](https://doi.org/10.48550/arXiv.1811.03604).

HEER, Jeffrey; BOSTOCK, Michael; OGIEVETSKY, Vadim. A Tour through the Visualization Zoo. **ACM Queue**, v. 8, n. 5, p. 20–30, 2010. DOI: [10.1145/1805974.1805978](https://doi.org/10.1145/1805974.1805978).

HEIDSIECK, Gaëtan *et al.* Cache-aware scheduling of scientific workflows in a multisite cloud. **Future Gener. Comput. Syst.**, v. 122, p. 172–186, 2021. DOI:

[10.1016/J.FUTURE.2021.03.012](https://doi.org/10.1016/J.FUTURE.2021.03.012). Disponível em:

<https://doi.org/10.1016/j.future.2021.03.012>.

HERSCHEL, Melanie; DIESTELKÄMPER, Ralf; LAHMAR, Housseem Ben. A survey on provenance: What for? What form? What from? **The VLDB Journal**, Springer, v. 26, n. 6, p. 881–906, 2017.

JURIĆ, Mario *et al.* **The LSST Data Management System**. [S. l.: s. n.], 2015. arXiv: [1512.07914](https://arxiv.org/abs/1512.07914) [astro-ph.IM]. Disponível em:

<https://arxiv.org/abs/1512.07914>.

KAIROUZ, Peter; MCMAHAN, H. Brendan; AVENT, Brendan *et al.* Advances and Open Problems in Federated Learning. **Foundations and Trends in Machine Learning**, v. 14, n. 1-2, p. 1–210, 2021. DOI: [10.1561/22000000083](https://doi.org/10.1561/22000000083).

KOIVU, Aki *et al.* Evaluation of machine learning algorithms for improved risk assessment for Down’s syndrome. **Comput. Biol. Medicine**, v. 98, p. 1–7, 2018. DOI:

[10.1016/J.COMPBIOMED.2018.05.004](https://doi.org/10.1016/J.COMPBIOMED.2018.05.004). Disponível em:

<https://doi.org/10.1016/j.compbimed.2018.05.004>.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. ImageNet classification with deep convolutional neural networks. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 60, n. 6, p. 84–90, mai. 2017. ISSN 0001-0782. DOI: [10.1145/3065386](https://doi.org/10.1145/3065386). Disponível em:

<https://doi.org/10.1145/3065386>.

LAM, Heidi *et al.* Empirical Studies in Information Visualization: Seven Scenarios. **IEEE Transactions on Visualization and Computer Graphics**, v. 18, n. 9, p. 1520–1536, 2012. DOI: [10.1109/TVCG.2011.279](https://doi.org/10.1109/TVCG.2011.279).

- LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, mai. 2015. ISSN 1476-4687. DOI: [10.1038/nature14539](https://doi.org/10.1038/nature14539). Disponível em: <https://doi.org/10.1038/nature14539>.
- LI, Quan *et al.* Inspecting the Running Process of Horizontal Federated Learning via Visual Analytics. **IEEE Transactions on Visualization and Computer Graphics**, v. 28, p. 4085–4100, 2022. DOI: [10.1109/TVCG.2021.3074010](https://doi.org/10.1109/TVCG.2021.3074010).
- LI, Tian *et al.* **Federated Optimization in Heterogeneous Networks**. [S. l.: s. n.], 2020. arXiv: [1812.06127](https://arxiv.org/abs/1812.06127) [cs.LG]. Disponível em: <https://arxiv.org/abs/1812.06127>.
- LINDSKOG, Jakob *et al.* Artificial intelligence and machine learning in sports medicine: mapping clinical tasks and assessing clinical maturity - a scoping review. **BMC Medical Informatics Decis. Mak.**, v. 26, n. 1, p. 212, 2026. DOI: [10.1186/S12911-026-03615-W](https://doi.org/10.1186/S12911-026-03615-W). Disponível em: <https://doi.org/10.1186/s12911-026-03615-w>.
- LIU, Ji; PACITTI, Esther *et al.* Scientific Workflow Scheduling with Provenance Data in a Multisite Cloud. **Trans. Large Scale Data Knowl. Centered Syst.**, v. 33, p. 80–112, 2017. DOI: [10.1007/978-3-662-55696-2_3](https://doi.org/10.1007/978-3-662-55696-2_3). Disponível em: https://doi.org/10.1007/978-3-662-55696-2_3.
- LIU, Juan; WILSON, Aaron; GUNNING, David. Workflow-based Human-in-the-Loop Data Analytics. In: PROCEEDINGS of the 2014 Workshop on Human Centered Big Data Research. Raleigh, NC, USA: Association for Computing Machinery, 2014. (HCBDR '14), p. 49–52. ISBN 9781450329385. DOI: [10.1145/2609876.2609888](https://doi.org/10.1145/2609876.2609888). Disponível em: <https://doi.org/10.1145/2609876.2609888>.
- LOPES, Camila; FERREIRA, Wesley *et al.* Federated Outlier Detection for Astronomical Data: Performance Analysis on Commercial Clouds. In: ANAIS do XXVI Simpósio em Sistemas Computacionais de Alto Desempenho. Bonito/MS: SBC, 2025. p. 362–373. DOI: [10.5753/sscad.2025.16731](https://doi.org/10.5753/sscad.2025.16731). Disponível em: <https://sol.sbc.org.br/index.php/sscad/article/view/37912>.
- LOPES, Camila; NUNES, Alan L. *et al.* Provenance-Based Dynamic Fine-Tuning of Cross-Silo Federated Learning. In: LATIN American High Performance Computing Conference (CARLA 2023). [S. l.]: Springer, 2023. (Communications in Computer and Information Science). DOI: [10.1007/978-3-031-52186-7_8](https://doi.org/10.1007/978-3-031-52186-7_8).
- MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: disponível em: <https://api.semanticscholar.org/CorpusID:6278891>.

MATTOSO, Marta *et al.* Dynamic steering of HPC scientific workflows: A survey. **Future Generation Computer Systems**, Elsevier, v. 46, p. 100–113, 2015.

MCMAHAN, H. Brendan *et al.* Communication-Efficient Learning of Deep Networks from Decentralized Data. *In: PROCEEDINGS of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. [S. l.: s. n.], 2017. v. 54. (JMLR Workshop and Conference Proceedings). arXiv:1602.05629. DOI: [10.48550/arXiv.1602.05629](https://doi.org/10.48550/arXiv.1602.05629).

MENG, Linhao *et al.* VADAF: Visualization for Abnormal Client Detection and Analysis in Federated Learning. **ACM Transactions on Interactive Intelligent Systems**, v. 11, 26:1–26:23, 2021. DOI: [10.1145/3426866](https://doi.org/10.1145/3426866).

MISSIER, Paolo *et al.* Provenance and data differencing for workflow reproducibility analysis. **Concurrency and Computation: Practice and Experience**, v. 28, n. 4, p. 995–1015, 2016. DOI: <https://doi.org/10.1002/cpe.3035>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.3035>. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.3035>.

MOREAU, Luc; CLIFFORD, Ben *et al.* The Open Provenance Model core specification (v1.1). **Future Generation Computer Systems**, v. 27, n. 6, p. 743–756, 2011. ISSN 0167-739X. DOI: <https://doi.org/10.1016/j.future.2010.07.005>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167739X10001275>.

MOREAU, Luc; GROTH, Paul *et al.* The rationale of PROV. **Journal of Web Semantics**, v. 35, p. 235–257, 2015. ISSN 1570-8268. DOI: <https://doi.org/10.1016/j.websem.2015.04.001>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1570826815000177>.

NEPOMUCENO, João *et al.* Agentic AI Meets Fundamental Financial Analysis: Limits and Possibilities. *In: FREITAS, Rosiane de; SILVA, Diego Furtado (ed.). Intelligent Systems - 35th Brazilian Conference, BRACIS 2025, Fortaleza-CE, Brazil, September 29 - October 2, 2025, Proceedings, Part I*. [S. l.]: Springer, 2025. v. 16179. (Lecture Notes in Computer Science), p. 131–145. DOI: [10.1007/978-3-032-15987-8_9](https://doi.org/10.1007/978-3-032-15987-8_9). Disponível em: https://doi.org/10.1007/978-3-032-15987-8_9.

NVIDIA CORPORATION. **NVIDIA FLARE: Federated Learning Application Runtime Environment**. [S. l.: s. n.], 2025. <https://developer.nvidia.com/flare>. Accessed: 2026-07-20.

OGASAWARA, Eduardo *et al.* An algebraic approach for data-centric scientific workflows. **Proc. VLDB Endow.**, VLDB Endowment, v. 4, n. 12, p. 1328–1339, ago. 2011. ISSN 2150-8097. DOI: [10.14778/3402755.3402766](https://doi.org/10.14778/3402755.3402766). Disponível em: <https://doi.org/10.14778/3402755.3402766>.

OLIVEIRA, Daniel de; COSTA, Flavio *et al.* Debugging Scientific Workflows with Provenance: Achievements and Lessons Learned. *In: MORO, Mirella M. (ed.). XXIX Simpósio Brasileiro de Banco de Dados, SBBD 2014, Curitiba, Paraná, Brasil, October 6-9, 2014.* [S. l.]: SBC, 2014. p. 67–76. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/sbbd/2014/006.pdf>.

OLIVEIRA, Daniel de; OCAÑA, Kary A. C. S. *et al.* A Provenance-based Adaptive Scheduling Heuristic for Parallel Scientific Workflows in Clouds. **J. Grid Comput.**, v. 10, n. 3, p. 521–552, 2012. DOI: [10.1007/s10723-012-9227-2](https://doi.org/10.1007/s10723-012-9227-2). Disponível em: <https://doi.org/10.1007/s10723-012-9227-2>.

OLIVEIRA, Daniel C. M. de; LIU, Ji; PACITTI, Esther. **Data-Intensive Workflow Management: For Clouds and Data-Intensive and Scalable Computing Environments.** [S. l.]: Morgan & Claypool Publishers, 2019. (Synthesis Lectures on Data Management). ISBN 978-3-031-00744-6. DOI: [10.2200/S00915ED1V01Y201904DTM060](https://doi.org/10.2200/S00915ED1V01Y201904DTM060). Disponível em: <https://doi.org/10.2200/S00915ED1V01Y201904DTM060>.

OXFORD UNIVERSITY PRESS. **Provenance: Definition of Provenance.** 2026. Disponível em: <https://www.oxfordlearnersdictionaries.com/definition/english/provenance>. Acesso em: 20 jul. 2026.

PERALTA, Maxime; JANNIN, Pierre; BAXTER, John S. H. Machine learning in deep brain stimulation: A systematic review. **Artif. Intell. Medicine**, v. 122, p. 102198, 2021. DOI: [10.1016/j.artmed.2021.102198](https://doi.org/10.1016/j.artmed.2021.102198). Disponível em: <https://doi.org/10.1016/j.artmed.2021.102198>.

PEREGRINA, José A.; ORTIZ, Guadalupe; ZIRPINS, Christian. Towards a Metadata Management System for Provenance, Reproducibility and Accountability in Federated Machine Learning. *In: ADVANCES in Service-Oriented and Cloud Computing.* Cham: Springer Nature Switzerland, 2022. p. 5–18. ISBN 978-3-031-23298-5.

PINA, Débora B. *et al.* DLProv: a suite of provenance services for deep learning workflow analyses. **PeerJ Comput. Sci.**, v. 11, e2985, 2025. DOI: [10.7717/PEERJ-CS.2985](https://doi.org/10.7717/peerj-cs.2985). Disponível em: <https://doi.org/10.7717/peerj-cs.2985>.

PRESIDÊNCIA DA REPÚBLICA DO BRASIL. **Lei No 13.709, de 14 de Agosto de 2018: Lei Geral de Proteção de Dados Pessoais (LGPD).** [S. l.: s. n.], 2018. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Official text of the Brazilian General Data Protection Law.

RUPPRECHT, Lukas *et al.* Improving Reproducibility of Data Science Pipelines through Transparent Provenance Capture. **Proc. VLDB Endow.**, v. 13, n. 12,

p. 3354–3368, 2020. DOI: [10.14778/3415478.3415556](https://doi.org/10.14778/3415478.3415556). Disponível em: <http://www.vldb.org/pvldb/vol13/p3354-rupprecht.pdf>.

SAMUEL, Sheeba; LÖFFLER, Frank; KÖNIG-RIES, Birgitta. Machine Learning Pipelines: Provenance, Reproducibility and FAIR Data Principles. *In: PROVENANCE Week 2020*. [S. l.: s. n.], 2020. DOI: [10.48550/arXiv.2006.12117](https://doi.org/10.48550/arXiv.2006.12117).

SCHLEGEL, Marius; SATTLER, Kai-Uwe. MLflow2PROV: Extracting Provenance from Machine Learning Experiments. *In: PROCEEDINGS of the Seventh Workshop on Data Management for End-to-End Machine Learning*. Seattle, WA, USA: Association for Computing Machinery, 2023. (DEEM '23). ISBN 9798400702044. DOI: [10.1145/3595360.3595859](https://doi.org/10.1145/3595360.3595859). Disponível em: <https://doi.org/10.1145/3595360.3595859>.

SCHLEGEL, Marius; SCHELIGA, Daniel *et al.* Lowering barriers to federated learning: collaboration management and provenance. *en. J. Big Data*, Springer Science e Business Media LLC, v. 13, n. 1, fev. 2026.

SILVA, Vítor; CAMPOS, Vinícius *et al.* DfAnalyzer: Runtime dataflow analysis tool for Computational Science and Engineering applications. **SoftwareX**, v. 12, p. 100592, 2020. ISSN 2352-7110.

SILVA, Vítor; OLIVEIRA, Daniel de *et al.* Dfanalyzer: runtime dataflow analysis of scientific applications using provenance. **Proc. VLDB Endow.**, VLDB Endowment, v. 11, n. 12, p. 2082–2085, ago. 2018. ISSN 2150-8097. DOI: [10.14778/3229863.3236265](https://doi.org/10.14778/3229863.3236265). Disponível em: <https://doi.org/10.14778/3229863.3236265>.

SOUZA, Carolina Veiga Ferreira de *et al.* ProWis: A Visual Approach for Building, Managing, and Analyzing Weather Simulation Ensembles at Runtime. **IEEE Transactions on Visualization and Computer Graphics**, v. 30, n. 1, 2024. DOI: [10.1109/TVCG.2023.3326514](https://doi.org/10.1109/TVCG.2023.3326514).

SOUZA, Renan *et al.* Workflow provenance in the lifecycle of scientific machine learning. **Concurr. Comput. Pract. Exp.**, v. 34, n. 14, 2022. DOI: [10.1002/cpe.6544](https://doi.org/10.1002/cpe.6544). Disponível em: <https://doi.org/10.1002/cpe.6544>.

STOCKINGER, Murilo B. *et al.* A Provenance-based Execution Strategy for Variant GPU-accelerated Scientific Workflows in Clouds. **J. Grid Comput.**, v. 20, n. 4, p. 36, 2022.

WANG, Xumeng; CHEN, Wei *et al.* HetVis: A Visual Analysis Approach for Identifying Data Heterogeneity in Horizontal Federated Learning. *In: IEEE Visualization Conference (VIS)*. [S. l.: s. n.], 2022. DOI: [10.48550/arXiv.2208.07491](https://doi.org/10.48550/arXiv.2208.07491).

WANG, Yuexuan; YIN, Jie. A reliable scheduling method in equipment grid using provenance information. **Inf. Syst. Frontiers**, v. 15, n. 4, p. 589–598, 2013. DOI: [10.1007/S10796-013-9437-6](https://doi.org/10.1007/S10796-013-9437-6). Disponível em: <https://doi.org/10.1007/s10796-013-9437-6>.

YANG, Qiang *et al.* Federated Machine Learning: Concept and Applications. **ACM Trans. Intell. Syst. Technol.**, Association for Computing Machinery, New York, NY, USA, v. 10, n. 2, jan. 2019. ISSN 2157-6904. DOI: [10.1145/3298981](https://doi.org/10.1145/3298981). Disponível em: <https://doi.org/10.1145/3298981>.

ZHANG, Jiachi; ZHOU, Wenchao; UJCICH, Benjamin E. Provenance-Enabled Explainable AI. **Proc. ACM Manag. Data**, v. 2, n. 6, 250:1–250:27, 2024. DOI: [10.1145/3698826](https://doi.org/10.1145/3698826). Disponível em: <https://doi.org/10.1145/3698826>.