INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 11, Issue 1 - V11I1-1513)
Available online at: https://www.ijariit.com

Federated Learning in Data Warehousing: A Privacy-Preserving Approach for Distributed Analytics

Akash Vijayrao Chaudhari
`chaudhari.avj@googlemail.com
Santander Bank, Flohram Park, NJ USA

Pallavi Ashokrao Charate
charatepallavi07@gmail.com
Santander Bank, Flohram Park, NJ USA

ABSTRACT

With the explosion of data generated across distributed environments, organizations face challenges in extracting insights while maintaining data privacy and regulatory compliance. Federated Learning (FL), a machine learning paradigm where models are trained across decentralized data sources without moving the data, has emerged as a promising solution. This paper explores the integration of FL with modern data warehousing architectures to enable secure, scalable, and privacy-preserving distributed analytics. We outline a federated data warehousing framework, highlight real-world use cases, evaluate system performance, and discuss future research directions.

Keywords: *Federated Learning, Data Warehousing, Privacy-Preserving Analytics, Distributed Machine Learning, Cloud Data Warehouse, Homomorphic Encryption, Differential Privacy, Secure Multi-party Computation, Federated Data Integration, Hybrid Cloud Architecture, Compliance-Aware AI, Real-Time Distributed Analytics*

INTRODUCTION

The demand for analytics on distributed data has intensified with the proliferation of edge devices, IoT, and multi-cloud strategies. Traditional centralized data warehousing involves data aggregation, which introduces privacy risks and data transfer bottlenecks. Federated Learning offers a new approach by enabling collaborative learning without raw data exchange. This paper proposes a novel architecture that combines FL with cloud-based data warehouses, offering both analytical performance and compliance with data sovereignty laws.

Background

Data Warehousing Evolution: Data warehousing has evolved from traditional on-premise systems to distributed and cloud-native platforms such as Snowflake, Amazon Redshift, and Google BigQuery. These modern systems provide scalability, elasticity, and on-demand analytics but still often require centralization of data, which poses challenges for privacy and data movement.

Federated Learning: FL is a decentralized machine learning paradigm introduced by Google in 2017. It enables model training across multiple devices or data centers where data resides, without moving the data itself. Clients train models locally and only share model updates (e.g., gradients or weights) with a central aggregator, preserving data privacy.

Privacy & Compliance: As global data privacy regulations such as the GDPR (EU), HIPAA (USA), and India's DPDP Act become stricter, organizations are under increasing pressure to avoid centralized data storage and sharing. Federated Learning provides a compelling compliance-aligned framework for analytics without breaching jurisdictional data boundaries.

PROPOSED FRAMEWORK

The proposed architecture integrates FL into data warehousing environments to achieve privacy-preserving distributed analytics. It is composed of four key layers working in coordination:

Data Nodes: These are the distributed endpoints where the raw data resides. They include hospitals, banks, mobile devices, industrial IoT sensors, and edge computing systems. Each node independently trains a local model on its private data. These models are periodically updated and sent to the central coordinator without exposing the original data. This approach enables local computation while preserving privacy.

FL Coordinator: The FL coordinator is the backbone of the federated learning infrastructure. It receives the local model updates from all participating data nodes and aggregates them into a unified global model. This is typically done using algorithms like Federated Averaging (FedAvg). The coordinator is responsible for managing training rounds, synchronizing updates, detecting anomalies in model weights, and redistributing the updated model to all clients. In advanced settings, the coordinator also manages differential privacy budgets and verifies compliance with encryption protocols.

Cloud DWH Interface: The cloud-based data warehouse functions as a central hub for storing the aggregated insights and trained models. It provides a unified access layer where analysts, data scientists, and business users can query insights, visualize trends, and feed results into business intelligence (BI) tools. This interface abstracts away the complexity of federated learning, presenting users with clean, queryable datasets and model outputs that can drive decision-making.

Security Layer: This layer underpins the entire architecture, ensuring data security and regulatory compliance. Techniques like homomorphic encryption allow computations on encrypted data, preventing data leakage even during aggregation. Differential privacy techniques are applied to ensure that individual data contributions cannot be reverse-engineered from the aggregated model. Additionally, secure multiparty computation (SMPC) may be employed to enhance confidentiality among untrusted participants.

The architecture is designed to be modular and extensible, enabling organizations to plug in various AI/ML components, cloud services, and compliance tools according to their operational needs. It supports hybrid deployments across on-premise, private, and public cloud environments and is compatible with popular federated learning libraries and data warehouse connectors.

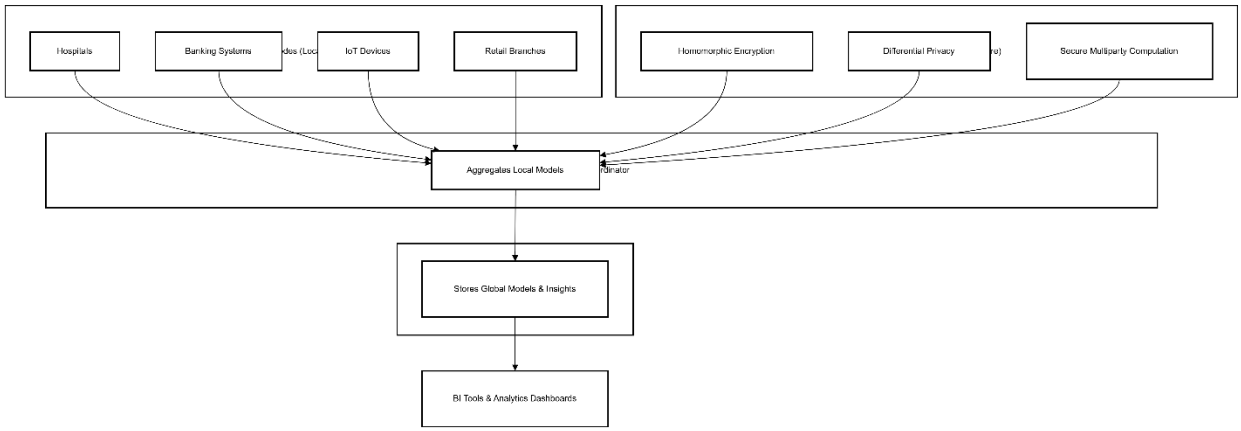

Diagram -1: Architecture

Use Cases

Healthcare: Hospitals and clinics can collaboratively build predictive models for disease detection, outbreak forecasting, or treatment optimization without exchanging patient records. For instance, training a COVID-19 severity prediction model across hospitals in different countries without violating patient confidentiality.

Finance: Banks and credit institutions can jointly train fraud detection models by learning from diverse datasets representing various fraud patterns, while keeping customer financial histories local and private.

Retail: Retailers with geographically dispersed stores can build personalized recommendation engines by leveraging purchase behaviors across locations. This enables tailored marketing without exposing customer identities or purchase histories.

Evaluation Metrics

Model Accuracy and Convergence Rate: Measures how well the federated model performs compared to a centralized model. It includes metrics like F1 score, precision-recall, and the number of communication rounds required to reach acceptable performance.

Data Transfer Overhead and Latency: Evaluates the amount of bandwidth consumed during model updates, and the time delay between rounds of communication. Efficient compression techniques can reduce overhead.

Compliance with Privacy Regulations: Assesses the system's adherence to standards like GDPR, ensuring no identifiable data leaves its source node.

Scalability Across Nodes: Examines how well the framework scales as the number of clients grows, considering both computational load balancing and synchronization delays.

CHALLENGES & LIMITATIONS

Communication Overhead and Straggler Effects: Federated learning requires frequent communication between nodes and the coordinator. Nodes with slower connections (stragglers) can delay the entire training process.

Model Bias Due to Non-IID Data: Data in federated environments is often non-identically distributed across nodes. This heterogeneity can lead to biased models if not properly addressed through techniques like clustering or personalization layers.

Integration Complexity with Existing DWH Ecosystems: Incorporating FL into established data warehousing workflows requires careful coordination of APIs, data schemas, and model versioning to ensure interoperability and maintain data lineage.

FUTURE RESEARCH DIRECTIONS

FL with Federated Analytics: Combining federated query processing with model training, allowing organizations to perform SQL-like analytics and ML collaboratively.

Cross-Silo FL Optimization Techniques: Enhancing FL algorithms to work better across larger, more stable clients like enterprises, as opposed to edge devices, with improvements in performance and personalization.

Automated Orchestration in Hybrid Cloud Environments: Developing orchestration layers that can manage training jobs across hybrid environments, ensuring resilience, autoscaling, and seamless workload migration.

CONCLUSION

Federated Learning holds immense potential for secure, privacy-preserving analytics in distributed data warehousing. By avoiding raw data movement, organizations can unlock insights across silos while maintaining compliance. Future advancements in FL algorithms, encryption techniques, and orchestration platforms will further enhance its applicability across industries.

REFERENCES

[1] Chaudhari, A. V., & Charate, P. A. (2024). Data Warehousing for IoT Analytics. International Research Journal of Engineering and Technology (IRJET), 11(6), 311–320.

[2] Kairouz, P., et al. (2021). Advances and Open Problems in Federated Learning. Foundations and Trends in Machine Learning.

[3] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated Learning: Challenges, Methods, and Future Directions.

[4] Bonawitz, K., et al. (2019). Towards Federated Learning at Scale: System Design. Proceedings of MLSys.

[5] Zhang, C., et al. (2022). FLaaS: Federated Learning as a Service for Data Warehousing Platforms. IEEE Cloud Computing.

[6] GDPR.eu. (2023). General Data Protection Regulation (GDPR) Compliance Guidelines.

[7] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Machine Learning: Concept and Applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.

[8] Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., ... & Eichner, H. (2018). Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604.

[9] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Artificial Intelligence and Statistics (AISTATS).

[10] Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security.

[11] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security.