Efficient Data Management for Federated Analytics through Optimized Compression Techniques

Donghyun Sohn, donghyun.sohn@u.northwestern.edu

As the volume of data generated directly from end-user devices surges, the imperative for efficient data analysis while concurrently preserving privacy has become increasingly significant. Federated analytics [1] emerges as a compelling solution to this challenge, enabling the analysis of data at the edge, thereby preserving user privacy. Notably, Google's deployment of federated analytics for the Now Playing feature on Pixel phones [3] exemplifies the practical application of this technology. However, a main bottleneck in federated analytics is the communication cost, particularly since end-user connections usually have lower bandwidth compared to inter-datacenter links [2]. While research efforts have focused on reducing communication overhead through various techniques—e.g., gradient compression and model broadcast compression—to my knowledge, there has been no attempt to apply traditional database compression methods, e.g., run-length encoding and delta encoding, specifically within federated analytics contexts. This acknowledgment of the unexplored potential of traditional database compression techniques in federated analytics sets the stage for our proposed investigation. By exploring this untouched area, we aim to contribute novel insights and methodologies to the field, enhancing data management practices in federated analytics environments.

This paper proposes a novel approach by adapting and optimizing traditional database compression techniques to improve data management within federated analytics. We introduce the concept of developing lightweight, optimized compression algorithms specifically designed for the federated analytics environment. By compressing raw data at the source—directly on end-user devices—prior to its transmission, our strategy significantly reduces bandwidth requirements and enhances the efficiency of the data aggregation process. This is curcial for facilitating scalable and privacy-preserving data analytics across distributed networks. The primary motivation behind adopting these compression techniques is to minimize local computation and storage demands on end devices through efficient data handling, which inherently contributes to bandwidth conservation.

Furthermore, our exploration extends to integrating these optimized compression techniques with secure aggregation protocols. This ensures the preservation of privacy guarantees, which are essential in federated analytics. By enabling the aggregation of compressed data while maintaining user privacy, we tackle the critical balance between achieving data compression efficiency and upholding the integrity of secure aggregation. Our goal is to develop a methodology that offers more scalable, efficient, and privacy-conscious federated analytics solutions.

By applying traditional database compression techniques for federated analytics, this paper lays the groundwork for the development of advanced, secure, and scalable data analytics frameworks. Our approach not only addresses the challenges inherent in federated analytics but also opens avenues for future advancements in data management and privacy-preserving analysis. This work highlights the need for technically robust and feasible solutions across devices involved in federated analytics, a significant step towards effective data processing in distributed computing.

References

[1] Ahmed Ramzy Elkordy, Youssef H Ezzeldin, and Shuguang Han. Federated analytics: A survey. *IEEE Transactions on Signal and Information Processing over Networks*, 2023.

[2] Peter Kairouz, H. Brendan McMahan, Brandon Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, and Zachary Charles. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 2021.

[3] Daniel Ramage and Stefano Mazzocchi. Federated analytics: Collaborative data science without data collection. https://blog.research.google/2020/05/federated-analytics-collaborative-data.html, 2020, May 27. Posted by Daniel Ramage, Research Scientist and Stefano Mazzocchi, Software Engineer, Google Research.