



A Privacy-Preserving Pipeline for Frequent Itemset Mining Over Federated Retail Transaction Logs

Rika Fenwick^{1*}, Rahel Wolde², Kofi Wolde³

¹Department of Computer Science, Ebro School of Computer Science, Bangladesh

²Department of Computer Science, Balkan Institute for Advanced Computer Science, Nigeria

³Department of Computer Science, Mekong Graduate Institute of Computer Science, Peru

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijdwdm.2026.v1.i1.pp21-26](https://doi.org/10.62715/ijdwdm.2026.v1.i1.pp21-26)

Received: 27 February 2026

Accepted: 14 April 2026

Published: 25 June 2026

Abstract

Retailers who pool their transaction logs can find purchasing patterns that none of them could find alone, but the logs are commercially sensitive and describe identifiable customers, so they are rarely shared. Two research traditions address the obstacle. Secure multiparty computation lets several parties compute joint itemset supports without revealing their inputs, and differential privacy limits what the published result can disclose about any one customer. Neither is sufficient alone: the first protects the computation and not the output, the second protects the output but presumes a trusted party to hold the data. This article proposes a five-stage pipeline that combines them for a federation of retailers. Each retailer harmonises its item codes and bounds every customer's contribution, mines candidate itemsets locally, and submits candidate supports through a secure aggregation protocol; calibrated noise is added to the aggregated counts before thresholding, and every release is charged against a recorded privacy budget. The article states the threat model, describes each stage with the choices it involves, and discusses how the unit of privacy, the number of participants and the length of itemsets affect the design. The pipeline is a proposal assembled from published techniques. It has not been implemented or evaluated here, and the closing sections say what an evaluation would need to measure.

Keywords: *Frequent itemset mining, Differential privacy, Secure aggregation, Federated analytics, Association rules, Retail transaction data*

© 2026 Rika Fenwick et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

Frequent itemset mining finds sets of products that are bought together in at least a given fraction of transactions. It was introduced for exactly this retail purpose (Agrawal et al., 1993), and the itemsets it returns feed association rules, shelf layouts, promotions and recommendation. The patterns that a single retailer can find are limited by its own assortment and customers. A regional grocer, a pharmacy chain and an online marketplace hold complementary views of the same households, and supports computed over their combined logs would reveal regularities that are too weak to detect in any one of them.

* Corresponding author: Rika Fenwick, Department of Computer Science, Ebro School of Computer Science, Bangladesh. E-mail: rika.fenwick@ebro.edu

© 2026 Rika Fenwick et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Pooling the logs is seldom acceptable. A transaction log exposes prices, volumes and assortment strategy to competitors, and it is personal data wherever baskets are tied to loyalty accounts or payment cards. Removing names does not make such data safe: sparse, high-dimensional records of the kind that purchase histories form can be re-identified from a few known items (Narayanan and Shmatikov, 2008). The General Data Protection Regulation and similar laws accordingly treat pseudonymised baskets as personal data (European Parliament and Council, 2016).

This article proposes a pipeline that lets a federation of retailers compute frequent itemsets over their joint transactions without any party handing over its log, and with a formal bound on what the published itemsets disclose about an individual customer. The pipeline is described as a method. No implementation, dataset or experiment is reported.

2. Background

The design rests on results from three areas: itemset mining in distributed data, secure computation of supports, and differentially private pattern mining.

2.1. Itemset mining and its distributed form

Apriori searches level by level, generating candidate itemsets of one size from the frequent itemsets of the size below and pruning any candidate that has an infrequent subset (Agrawal and Srikant, 1994). FP-growth avoids candidate generation by compressing the transactions into a prefix tree and mining it recursively (Han et al., 2000). When the transactions are split horizontally across sites, a simple property limits the work: an itemset whose relative support reaches the threshold over the union of the sites must reach it at one site or more. Cheung et al. (1996) built a distributed algorithm on this observation, in which each site proposes its locally frequent itemsets and only those are counted globally.

2.2. Protecting the computation

Kantarcioglu and Clifton (2004) made the distributed scheme private with respect to the participating sites. Their protocol forms the union of locally frequent itemsets under commutative encryption, so that no site learns who proposed which candidate, and then uses a secure sum to test whether each candidate's global support exceeds the threshold. Secure summation can be realised in several ways. Additively homomorphic encryption allows ciphertexts to be combined into an encryption of the sum (Paillier, 1999). Secure aggregation, developed for federated learning, has each pair of participants agree on random masks that cancel when all contributions are added, with secret sharing to recover from participants that drop out (Bonawitz et al., 2017). An older alternative perturbs each transaction at the source before it is sent (Evfimievski et al., 2002).

2.3. Protecting the output

Secure computation reveals nothing beyond the result, but the result itself may disclose a good deal. An itemset that barely crosses the threshold, or the difference between two releases, can betray the presence of one customer's baskets. Differential privacy addresses this by requiring that the distribution of the output change by at most a factor governed by a parameter ϵ when one individual's data are added or removed, and the basic mechanism adds noise scaled to the largest change one individual can cause, the sensitivity (Dwork et al., 2006). Privacy losses add up over repeated releases, which makes ϵ a budget to be spent (Dwork and Roth, 2014).

Itemset mining is awkward under this definition because one transaction of m items supports $2^m - 1$ non-empty itemsets, so the sensitivity of the full support vector is very large. Bhaskar et al. (2010) gave the first algorithms for releasing the most frequent patterns privately. Zeng et al. (2012) showed that truncating long transactions reduces sensitivity enough to outweigh the information lost, and PrivBasis projects the data onto a small number of item groups that cover the frequent itemsets (Li et al., 2012). In the local model, where each contributor randomises its own data before sending it, itemset mining is possible but needs far more data for the same accuracy (Wang et al., 2018).

3. Design Goals and Threat Model

The federation consists of a small number of retailers, typically fewer than would take part in a consumer-device deployment of the kind for which federated learning was conceived (McMahan *et al.*, 2017), and a coordinator that schedules rounds and publishes results. Every party is assumed to follow the protocol while trying to learn what it can from the messages it sees, the semi-honest model. The coordinator is not trusted with raw logs or with any single retailer's counts. Retailers are assumed not to collude with the coordinator in numbers large enough to isolate one member's contribution; the protocol parameters fix that number.

Two kinds of secret are protected. The first is commercial: a retailer's supports for candidate itemsets, and which candidates it proposed, should stay hidden from the others. The second is personal: the published itemsets and supports should satisfy differential privacy at the level of the customer, not the single transaction, since a regular shopper contributes many baskets. The design aims to achieve both with accuracy close to that of the central model of differential privacy, in which noise is added once to an aggregate, and not the much noisier local model.

4. The Proposed Pipeline

The pipeline has five stages, run once per release. Table 1 sets out what each stage does, what it protects and what it costs in accuracy or effort. Figure 1 pairs each stage with its safeguard.

Stage	What is done	What it protects	Principal cost
1. Harmonise and bound	Map items to a shared taxonomy; cap baskets and items per customer	Bounds any one customer's influence	Truncation discards part of large histories
2. Mine locally	Each retailer mines its own log at a lowered threshold	Raw transactions never leave the site	Low thresholds lengthen candidate lists
3. Aggregate securely	Candidates merged; supports summed under masking or encryption	Each retailer's counts and proposals	Communication and cryptographic computation
4. Perturb and threshold	Noise scaled to the contribution cap is added to each sum	Customers, in the published output	Noisy supports; rare itemsets lost
5. Release and account	Supports made consistent; privacy loss recorded	Against erosion over repeated releases	Finite number of releases

4.1. Stage 1: Harmonise items and bound contributions

Retailers use different product codes, so the federation first agrees a shared item vocabulary, normally a level of a product category hierarchy and not individual stock-keeping units. Coarser items make joint patterns possible and shorten transactions. Each retailer then enforces two limits within its own log: a maximum number of items per transaction, following the truncation argument of Zeng *et al.* (2012), and a maximum number of transactions per customer in the mining period. Together the limits fix the largest change a single customer can make to any support count, which is the quantity to which the noise in Stage 4 is calibrated. Baskets that cannot be tied to a customer are treated as coming from separate individuals.

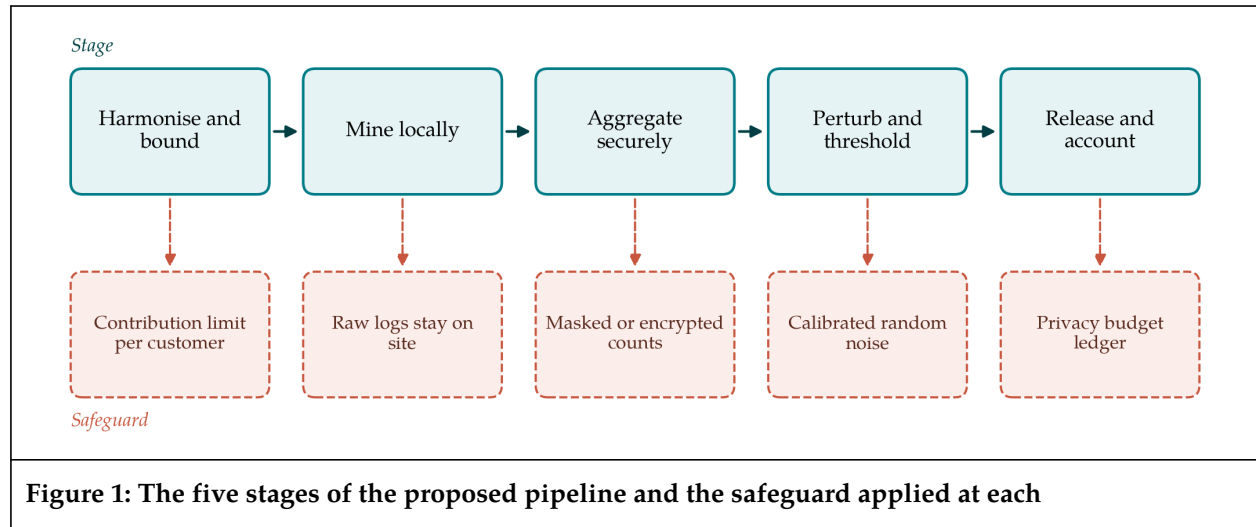
4.2. Stage 2: Mine candidates locally

Each retailer runs an ordinary miner, FP-growth for instance, over its bounded log and keeps the itemsets that are locally frequent. By the property used by Cheung *et al.* (1996), every globally frequent itemset appears in some retailer's list, so the union of the lists is a complete candidate set. In practice the local threshold is set somewhat below the global one, because the noise added later can lift a borderline itemset above the threshold or push it below. The retailer also computes its support for every subset of its

candidates, which will be needed when other retailers propose them.

4.3. Stage 3: Aggregate supports securely

The candidate lists are merged without attribution, using the commutative-encryption union of [Kantarcioglu and Clifton \(2004\)](#) or, where the item vocabulary is small, by fixing the candidate set in advance as all itemsets up to a given length over the most popular items. Each retailer then reports its local count for every candidate, together with its bounded transaction total. The counts are summed under a secure aggregation protocol in the style of [Bonawitz et al. \(2017\)](#), or under additively homomorphic encryption with the decryption key shared among the parties. The coordinator obtains only the totals.



4.4. Stage 4: Perturb and threshold

Before any total is revealed, random noise is added to it. So that no single party has to be trusted to add the noise honestly or to forget its value, each retailer adds a share of the noise to its own masked contribution, with the shares calibrated to give the required total even if a stated number of parties disclose theirs. The scale of the noise follows from the contribution limits of Stage 1, the number of candidates reported and the portion of the privacy budget assigned to the release. Itemsets whose noisy support falls below the global threshold are discarded. Selecting which itemsets to publish is itself a disclosure, and the analysis must account for it; fixing the candidate set in advance, as in the second option of Stage 3, simplifies that accounting.

4.5. Stage 5: Release and account

Noisy supports can violate the rule that an itemset is no more frequent than its subsets. They are adjusted to restore consistency, which costs no privacy because any computation on an already private output is itself private ([Dwork and Roth, 2014](#)). The coordinator then records the parameters of the release in a ledger that all members can inspect. When the cumulative loss reaches the limit the federation has set for the period, no further release is made from the same data.

5. Fitting the Pipeline to Practice

Several choices depend on circumstances. With very few retailers, secure aggregation protects little, since each can subtract its own count from the total and reason about the rest; the noise of Stage 4 then carries most of the protection for commercial secrets as well as personal ones, and the federation should say so in its agreement. Itemset length matters in a different way. The number of candidates grows quickly with length while supports shrink, so the noise soon exceeds the signal; limiting the pipeline to short itemsets over a few hundred categories is a reasonable starting configuration. Where the patterns of interest are ordered sequences and not sets, the interactive, prefix-by-prefix approach developed for federated heavy

hitters is the closer model (Zhu et al., 2020).

6. Limitations and What Remains to Be Shown

The pipeline has not been implemented or tested in this article, and its practical value depends on quantities that only experiment can supply. The most important is utility: how closely the released itemsets and supports match those of non-private mining over the pooled data, as a function of the privacy parameter, the contribution limits and the number and size of the participants. Public basket datasets, partitioned to imitate retailers of unequal size, would allow this to be measured. Cost is the second: the time and communication that secure aggregation requires as the candidate set grows. The third is the sensitivity of results to the item vocabulary, since harmonisation decides which patterns can be seen at all.

Some limits are inherent. The semi-honest assumption does not cover a retailer that submits false counts to distort the result or to probe for a competitor's data, and defending against such behaviour needs verifiable computation or auditing outside the protocol. Customer-level privacy holds only within each retailer, because a person who shops at three members is three customers to the pipeline unless identities are linked, which the design deliberately avoids. And differential privacy bounds what the output reveals about individuals; it gives no assurance about aggregate facts, such as a category's overall popularity at one member, that the federation may also regard as sensitive.

7. Conclusion

Joint itemset mining across retailers is obstructed less by algorithms than by the sensitivity of the logs. The pipeline proposed here keeps raw transactions on site, hides each retailer's counts inside a secure sum, and bounds the influence of any one customer on what is published. Every component is taken from the literature; the contribution lies in their arrangement and in making the unit of privacy and the trust assumptions explicit. Whether the released patterns remain useful at privacy levels a regulator would accept is the question on which adoption turns, and it remains to be answered by evaluation.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Agrawal, R. and Srikant, R. (1994). [Fast algorithms for mining association rules](#). *Proceedings of the 20th International Conference on Very Large Data Bases*, 487-499.
- Agrawal, R., Imieliński, T. and Swami, A. (1993). [Mining association rules between sets of items in large databases](#). *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 207-216. <https://doi.org/10.1145/170035.170072>
- Bhaskar, R., Laxman, S., Smith, A. and Thakurta, A. (2010). [Discovering frequent patterns in sensitive data](#). *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 503-512. <https://doi.org/10.1145/1835804.1835869>
- Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A. and Seth, K. (2017). [Practical secure aggregation for privacy-preserving machine learning](#). *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175-1191. <https://doi.org/10.1145/3133956.3133982>
- Cheung, D.W., Han, J., Ng, V.T., Fu, A.W. and Fu, Y. (1996). [A fast distributed algorithm for mining association rules](#). *Proceedings of the Fourth International Conference on Parallel and Distributed Information Systems*, 31-42.

- Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4): 211-407. <https://doi.org/10.1561/04000000042>
- Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography (TCC 2006), Lecture Notes in Computer Science 3876*, 265-284. https://doi.org/10.1007/11681878_14
- European Parliament and Council (2016). Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Official Journal of the European Union*, L 119: 1-88.
- Evfimievski, A., Srikant, R., Agrawal, R. and Gehrke, J. (2002). Privacy preserving mining of association rules. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 217-228.
- Han, J., Pei, J. and Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 1-12. <https://doi.org/10.1145/342009.335372>
- Kantarcioglu, M. and Clifton, C. (2004). Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge and Data Engineering*, 16(9): 1026-1037. <https://doi.org/10.1109/TKDE.2004.45>
- Li, N., Qardaji, W., Su, D. and Cao, J. (2012). PrivBasis: Frequent itemset mining with differential privacy. *Proceedings of the VLDB Endowment*, 5(11): 1340-1351.
- McMahan, B., Moore, E., Ramage, D., Hampson, S. and Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research*, 54: 1273-1282.
- Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, 111-125. <https://doi.org/10.1109/SP.2008.33>
- Paillier, P. (1999). Public-key cryptosystems based on composite degree residuosity classes. *Advances in Cryptology: EUROCRYPT 1999, Lecture Notes in Computer Science 1592*, 223-238. https://doi.org/10.1007/3-540-48910-X_16
- Wang, T., Li, N. and Jha, S. (2018). Locally differentially private frequent itemset mining. *Proceedings of the 2018 IEEE Symposium on Security and Privacy*, 127-143. <https://doi.org/10.1109/SP.2018.00035>
- Zeng, C., Naughton, J.F. and Cai, J.-Y. (2012). On differentially private frequent itemset mining. *Proceedings of the VLDB Endowment*, 6(1): 25-36.
- Zhu, W., Kairouz, P., McMahan, B., Sun, H. and Li, W. (2020). Federated heavy hitters discovery with differential privacy. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research*, 108: 3837-3847.