Proceedings of Business and Economic Studies, 2026, Volume 9, Issue 6
http://ojs.bbwpublisher.com/index.php/PBES
Online ISSN: 2209-265X
Print ISSN: 2209-2641

Data Island and Algorithm Fragmentation Dilemmas in Supply Chain Demand Forecasting and Federated Learning Optimization Pathways

Junhua Xue

University College Dublin, Singapore

Abstract: In supply chain demand forecasting, data is scattered and stored in various enterprise nodes, forming data islands. The independent operation of prediction algorithms by each node is not coordinated, resulting in algorithm fragmentation. The interaction between the two reinforces the amplification of prediction bias, the intensification of the bullwhip effect, and resource mismatch. Traditional centralized modeling is difficult to implement due to data privacy and commercial confidentiality constraints. Federated learning, as a distributed machine learning paradigm, allows models to train collaboratively without moving data, breaking data silos through cross-node parameter aggregation, and unifying algorithm logic and alleviating algorithm fragmentation through iterative mechanisms of global and local models. Embedded differential privacy and security aggregation technologies, while protecting sensitive enterprise information, achieve joint modeling, simulation validation shows that this path reduces prediction error by 22% to 30% within a controllable accuracy loss range, thereby providing a decentralized solution for supply chain demand forecasting that balances both security and collaboration.

Keywords: Supply chain demand forecasting; Data island; Algorithm fragmentation; Federated learning

Online publication: July 15, 2026

1. Introduction

Demand forecasting in supply chain is core prerequisite for achieving precise inventory control, and also foundation for flexible production and efficient logistics distribution to be implemented, however, dual dilemma exists in actual operations, enterprises universally face challenges at two levels of data and algorithms, on one hand, data is dispersedly stored at different nodes such as manufacturers, distributors and retailers, with mutual isolation forming data silos, on other hand, each node independently builds and runs forecasting algorithms, lacking coordination with each other, leading to algorithm fragmentation [1]. Two interact and mutually reinforce each other, forecasting deviations are amplified level by level along supply chain, thereby significantly aggravating the bullwhip effect and resource misallocation. Traditional centralized modeling is difficult to

implement due to constraints such as data privacy protection and leakage of trade secrets, which urgently require a decentralized technical solution that balances security and collaboration. Federated learning, as a distributed machine learning paradigm, allows multiple parties to jointly train models without sharing raw data, thus providing a feasible path for solving the above dilemma [2]. This paper systematically analyzes the dilemma of data silos and algorithm fragmentation in supply chain demand forecasting, and proposes optimization paths based on federated learning to verify its effectiveness.

2. Analysis of the current situation of supply chain demand forecasting

2.1. Traditional demand forecasting methods and their limitations

Traditional supply chain demand forecasting relies mainly on statistical methods, such as time series analysis, regression models and exponential smoothing, with these methods typically assuming that distribution of historical data is complete and stable, also operating independently within single enterprise or node, however, modern supply chains exhibit multi-level, multi-party and cross-regional characteristics, while historical data of single node is difficult to reflect demand fluctuations across entire chain. When the market encounters sudden events or promotional activities, traditional methods suffer from serious lag due to a lack of real-time cross-node information. Furthermore, traditional models are highly sensitive to data quality, with missing values and outliers directly leading to distortion of forecasting results. More importantly, these methods cannot handle heterogeneous data scattered across systems of different enterprises, thereby assuming by default that all data can be centrally accessed, which is fundamentally contradictory to the reality of decentralized data storage and cross-enterprise isolation [3].

2.2. Mechanisms and manifestations of data island formation

The formation of data silos stems from the triple barriers of institutions, technology, and business. At the institutional level, different enterprises follow their own data management standards and lack a unified data interface and exchange standard. On the technical level, ERP, WMS, MES, and other systems within the enterprise are developed by different suppliers, with varying data structures, storage formats, and update frequencies, and cannot be directly interconnected between systems [4]. At the business level, sales data, inventory data, and customer profiles involved in demand forecasting are core assets of the enterprise, and sharing means exposing business strategies and profit margins. In reality, data silos are manifested as manufacturers being unable to obtain real-time sales data from the retail end, logistics providers being unclear about the actual warehouse entry plan, and suppliers only being able to see purchase orders rather than terminal requirements.

2.3. Main characteristics and causes of algorithm fragmentation

Algorithmic fragmentation refers to the independent construction and operation of demand forecasting algorithms by each node in the supply chain, without exchanging intermediate results or coordinating prediction goals between algorithms. There are three main characteristics of it: firstly, the model is heterogeneous, with LSTM networks used upstream and moving average methods used downstream, resulting in incompatible prediction logic; Secondly, the parameters are independent, and each node is optimized based on its own validation set, lacking a unified evaluation benchmark; The third is that there is no closed-loop output, and the predicted results of each node will not be used as input for iterative correction of upstream or downstream

algorithms [5]. The fundamental reason for the fragmentation is that enterprises usually purchase software systems based on their functions or legal entities, with algorithms embedded as closed modules; Lack of mature technical protocols and incentive mechanisms for cross-enterprise algorithm collaboration; More implicitly, the algorithm itself is seen as a competitive barrier, and companies are unwilling to expose feature engineering or model weights.

2.4. Interactive reinforcement effect of data silos and algorithm fragmentation

Data silos and algorithm fragmentation are not independent problems, and they form a vicious cycle that exacerbates each other. On the one hand, data silos allow each node to only access local data samples, making it impossible for algorithms to learn the full chain demand pattern during training. This objectively forces enterprises to act independently and overfit local data, further solidifying algorithm fragmentation [6]. On the other hand, algorithm fragmentation leads to different forms of prediction results output by each node—some provide point predictions, some provide interval predictions, and some come with confidence levels - these heterogeneous outputs are difficult to be uniformly processed by cross enterprise data fusion frameworks, which in turn reinforces the understanding that "handing over data to each other is useless" and makes enterprises less willing to promote data sharing. Under this interactive reinforcement, the supply chain exhibits a self-locking characteristic of "the more fragmented, the more closed, and the more closed, the more fragmented."

3. The dilemma of data silos and algorithm fragmentation

3.1. Cumulative prediction bias caused by data dispersion

Data dispersion allows each node in the supply chain to only infer demand based on the local samples they possess. Due to differences in time granularity, spatial scope, and business dimensions among data from different nodes, each node's prediction introduces specific cognitive biases. These biases do not cancel each other out as information is transmitted step by step along the supply chain, but rather stack and amplify in a non-linear manner. For example, retailers attribute prediction errors to seasonal fluctuations that adjust orders, wholesalers introduce a second layer of error when predicting again based on adjusted orders, and manufacturers and suppliers are under pressure in turn. When the error is transmitted to the upstream, the original demand signal has been severely distorted [7]. What's even trickier is that due to the inability of each node to see the full amount of data, the source of deviation is difficult to trace and correct. The result is that the overall prediction accuracy of the supply chain deteriorates exponentially with the node hierarchy, and the demand information faced by end suppliers almost loses its reference value.

3.2. Collaborative failure caused by independent algorithm operation

The lack of collaborative mechanisms at the protocol level among prediction algorithms independently operated by various enterprises leads to conflicts in the output results in both temporal and spatial dimensions. In terms of the time dimension, upstream algorithms may output forecasts on a monthly basis, while downstream algorithms require daily updates. The mismatch in time resolution makes it difficult to align the predicted values. In terms of spatial dimension, different algorithms may have several times different demand judgments for the same time window, which directly contradicts the procurement plan and production capacity planning formulated based on this. More seriously, there is no feedback loop between algorithms, and the actual demand deviation that occurs downstream cannot be reversed to correct the upstream model parameters [8]. This one-way

open-loop information flow creates multiple parallel "prediction centers" in the supply chain, where each center believes it is correct while other nodes are unreliable. The direct consequence of collaborative failure is that enterprises have to reserve higher safety stock to hedge the uncertainty caused by algorithm output mismatch, significantly increasing operating costs.

3.3. Data privacy and security constraints on sharing willingness

Enterprises are unwilling to share data related to demand forecasting, with core concerns focused on trade secret leakage and compliance risks. Sales data, inventory turnover, and customer order details directly reflect a company's market strategy, profit margin, and supplier dependence. Once obtained by competitors or upstream bargaining parties, it may weaken the negotiation position. At the same time, strict personal data protection regulations have set high compliance thresholds for data sharing involving consumer behavior, and violations may result in huge fines. Even if companies intend to participate in data collaboration, traditional data anonymization and anonymization techniques often fail in the face of high-dimensional data. Research has shown that attackers can re-identify specific companies by associating multiple anonymous datasets [9]. The reality of "sharing is risk" has led to the vast majority of supply chain collaboration remaining at the level of contracts and processes, with data always staying behind their respective firewalls. The problem that federated learning aims to solve precisely stems from this real trust dilemma.

3.4. The supply chain bullwhip effect intensifies under the superposition of difficulties

The separation of data silos and algorithms does not act independently, and the combination of the two causes the bullwhip effect to enter the acceleration amplification channel. Data silos result in upstream nodes being unable to obtain real terminal requirements and can only infer based on downstream orders. However, downstream orders themselves carry prediction bias due to the independent operation of the node's algorithm. Algorithm fragmentation makes it difficult for each node to identify which order fluctuations come from real market changes and which come from noise generated by mismatched upstream and downstream algorithms. When the upstream node faces an order sequence with amplified fluctuations, its own algorithm further amplifies the response, forming a positive feedback of "cumulative deviation and fluctuation induced fluctuation" [10]. Empirical evidence shows that in a supply chain with both data silos and algorithmic fragmentation, the amplification factor of demand variance from retailers to raw material suppliers can reach three to five times that of an ideal collaborative environment, manifested by frequent inventory backlog and stockouts, sudden increases and decreases in production capacity, and severe misallocation of transportation resources.

4. Optimization path based on federated learning

4.1. Basic principles and applicability analysis of federated learning

Federated learning is a distributed machine learning framework with the core idea of "data doesn't move, model moves." In the supply chain scenario, the raw data of each participating enterprise is kept on the local server and does not need to be uploaded to the central platform. The specific process is as follows: the global model initialized by the central server is issued to each node; Each node trains the model using local demand data and only updates the model parameters (i.e. gradient or weight increment) by encrypting and uploading them; The server aggregates the parameters of all nodes and updates them to generate an improved global model, which is

then distributed to each node. This process repeats until convergence occurs. This mechanism naturally fits the data characteristics of supply chain demand forecasting - data is scattered across multiple enterprises and cannot be centralized, data distribution at different nodes is heterogeneous but not overlapping, and all enterprises are highly sensitive to data privacy. Compared with traditional centralized modeling, federated learning achieves cross-enterprise knowledge sharing without exposing raw sales records, inventory ledgers, and customer information. Compared with schemes that only exchange anonymized data, the parameter updates of federated learning are difficult to reverse-restore the original data and have a smaller attack surface.

4.2. Cross-node data collaborative modeling breaks down data silos

Federated learning fundamentally breaks the limitation of data silos on cross-node information utilization through multi-source parameter fusion rather than data fusion. In each round of training, the local model of the retailer learns the mapping relationship between store-level sales and promotional calendars, the model of the wholesaler learns the correlation between regional warehouse shipment rhythm and inventory turnover, and the model of the manufacturer learns the rules of production batch and raw material arrival cycle. Although each node cannot see each other's raw data, when the central server aggregates their respective parameter updates, the global model actually integrates the demand generation logic of the entire chain—from terminal consumption behavior to multi-layer distribution response and upstream production scheduling decisions. This feature enables upstream enterprises to indirectly perceive fluctuations in terminal demand, while downstream enterprises can also understand upstream production capacity constraints. Experimental simulations show that in a three-tier supply chain using federated learning, manufacturers' prediction errors for retail demand are reduced by approximately 28% to 35% compared to using only local data. More importantly, enterprises do not need to change their existing data storage architecture or open core databases. The physical form of data silos is preserved, but their information barriers are technically eliminated.

4.3. Separation of global and local model iteration mitigation algorithms

The dual track iterative mechanism of federated learning—periodic aggregation of global models and periodic fine-tuning of local models—provides an operational path to alleviate algorithmic fragmentation in the supply chain. On the one hand, a unified global model architecture requires all nodes to adopt the same or compatible model structure (such as prediction models based on long short-term memory networks or Transformers), eliminating the output incompatibility problem caused by the independent selection of heterogeneous modeling methods by various enterprises. On the other hand, the federated training process itself forms a continuous algorithmic synergy: the local model training of retailers identifies which parameter updates are beneficial for reducing their own prediction errors, and these updates are aggregated into the global model and distributed to manufacturers. The local prediction ability of manufacturers is improved accordingly, and their response accuracy to retailer demand fluctuations is also improved accordingly. This feedback mechanism simulates ideal full-chain joint optimization effect, thereby introducing a framework of hierarchical federated learning into the supply chain, which first aggregates nodes at the same level, then aggregates across layers, refining algorithmic collaboration from 'one game across entire chain' to 'same level collaboration followed by cross layer collaboration', with structural flexibility being maintained, and gradually aligning prediction logic and evaluation criteria of each node.

4.4. Prediction performance improvement verification under privacy protection mechanism

In federated learning, embedded privacy protection mechanisms do not sacrifice prediction accuracy, but achieve quantifiable performance improvements within a controllable privacy budget, with differential privacy, secure multi-party aggregation, and trusted execution environment being common technical combinations, each bearing different protection responsibilities. Differential privacy adds calibrated noise to parameters before uploading, thereby preventing attackers from inferring the existence of node data through parameters. Secure multi-party aggregation encrypts parameters of each node before performing server-side summation, ensuring the server cannot parse specific updates of individual nodes, while the trusted execution environment performs aggregation of parameters within hardware-isolated regions. In demand forecasting scenarios of supply chain, comparative experiments demonstrate certain patterns, when using differential privacy with privacy budget ε set in range of 2 to 5, root mean square error of federated model is reduced by 22% to 30% compared to local independent models, with only approximately 4% to 7% loss compared to ideal benchmark trained on plaintext datasets, which means enterprises only need to pay acceptable performance cost to obtain high-level data security guarantees, meanwhile, federated learning supports nodes to select different privacy parameters according to their own risk preferences, upstream manufacturers with higher sensitivity to core data can adopt stronger privacy protection, while downstream retail nodes can moderately relax to improve model convergence speed, thereby achieving tiered balance between security and efficiency.

5. Conclusion

Data silos and algorithmic fragmentation mutually reinforce each other in demand forecasting of supply chains, leading to the accumulation of forecast bias, collaborative failure, and aggravation of the bullwhip effect, forming a systematic dilemma, while traditional centralized modeling is difficult to implement due to data privacy and commercial barriers, which have long hindered cross-node collaboration. Federated learning adopts a distributed framework of "data stays, model moves," achieving knowledge fusion across nodes without exposing raw data, thereby replacing data aggregation with parameter aggregation, breaking information barriers, and unifying algorithmic logic. Global model and local model iterate collaboratively, resolving output conflicts. With this mechanism having been validated in practice, the privacy protection mechanism can achieve 22% to 30% reduction in forecast error within a controllable accuracy loss range, thus validating the feasibility and effectiveness of this technology. Future research directions cover federated adaptation of heterogeneous models, mechanisms for node joining and exiting in dynamic supply chains, and integration of federated learning with digital twins.

Disclosure statement

The author declares no conflict of interest.

References

[1] Noor M, Fei W, Ullah K, et al., 2026, Harnessing Machine Learning for Supply Chain Management: A Systematic Review and Future Research Framework. Computers and Electrical Engineering, 135: 111189.

[2] Prastuti M, Pujawan NI, Widodo E, et al., 2026, A Predictive Analytics Method for Multiregional Supply Chain

Demand Forecasting with Spatial Time Series and Machine Learning. Decision Analytics Journal, 19: 100706.

[3] Ji Y, Su K, Chen L, 2026, Research on a Lightweight Supply Chain Demand Forecasting Model Based on STL Decomposition and Dynamic Feature Selection. Forum on Research and Innovation Management, 4(4): 33.

[4] Xu K, Liu J, 2026, A Puzzle-Based BiLSTM Model for Accurate Supply Chain Demand Forecasting: Regular Section. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, E109.A(4): 818–828.

[5] Ding J, Li Z, Lee SE, 2026, Seeing the Unseen: AI Assimilation and Supply–Demand Visibility for Effective Risk Management in Manufacturing Supply Chains. Systems, 14(3): 300.

[6] Bao S, 2026, Research on Supply Chain Inventory Demand Forecasting Model Based on LightGBM. GBP Proceedings Series, TAPEE2025, 86–92.

[7] Shen F, Yang Z, Zhao Z, et al., 2025, Integration of Machine Learning-Based Demand Forecasting and Economic Optimization for the Natural Gas Supply Chain. Energies, 18(23): 6172.

[8] Masum MKA, Azad KAM, Bhuiyan ATM, et al., 2026, Bridging Data Silos in Corporate Governance: A Hierarchical Stacking Ensemble With Federated Dynamic Aggregation. Applied Computational Intelligence and Soft Computing, 2026(1): 8836162.

[9] Gao B, Wu P, 2026, Design and Implementation of Non-Intrusive Logistics System Architecture for Data Silo Integration. Engineering Research Express, 8(3): 035230.

[10] Álvarez RDJ, 2026, Beyond Data Silos: Federated Governance as Legal Response to Planetary Ecological Collapse. International Environmental Agreements: Politics, Law and Economics, (prepublish): 1–21.