International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 4 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

A SCALABLE MULTI-SOURCE DATA FABRIC FOR END-TO-END SUPPLY CHAIN VISIBILITY, INVENTORY INTELLIGENCE AND BUSINESS DECISION ANALYTICS - (SUPPLY SPHERE)

[1] Mrs. Ch. Roja, [2] Burla Sai Ram, [3] Elagandula Karthik, [4] Nagelli Bhanuprakash, [5] Kalletlapalli Mahesh
[1]Assistant Professor, [2345]Students
Department of CSE (Data Science)
Siddhartha Institute of Technology & Sciences, Narapally
sits.rojachilukuri@siddhartha.co.in, 24tq1a6750@siddhartha.co.in, 24tq1a6751@siddhartha.co.in, 24tq1a6759@siddhartha.co.in, 24tq1a6754@siddhartha.co.in

ABSTRACT

A modern supply chain produces data at every step, from purchase orders and supplier shipments to warehouse movements, transport tracking, and retail sales. In most mid-sized manufacturing and distribution companies, however, this data lives in separate systems owned by different teams. The ERP knows what was ordered, the warehouse management system knows what is on the shelves, the transport provider knows where a truck is, and the sales system knows what customers bought. Because these systems do not share a common view, managers struggle to answer basic questions such as how much stock is actually available across the network, which shipments are likely to arrive late, and which products are about to run out. This paper presents Supply Sphere, a scalable multi-source data fabric that connects these systems and uses the integrated data for inventory intelligence and business decision analytics.

The data fabric ingests data through change data capture from the ERP and warehouse databases, scheduled API connectors for supplier portals and the transport management system, and a streaming feed of point-of-sale transactions and GPS pings. All incoming data lands in an Apache Iceberg lakehouse on object storage, where a transformation layer built with dbt cleans, conforms, and joins it. A metadata catalog records the source, owner, schema, and lineage of every dataset, and an entity resolution step matches products, suppliers, and locations that carry different codes in different systems. A semantic layer then exposes consistent business measures such as available stock, days of cover, fill rate, and on-time in-full delivery.

On the integrated data, the system runs a set of analytical models. A hierarchical demand forecasting model combines LightGBM with seasonal baselines and reconciles forecasts across product, category, and region. An inventory optimisation module computes safety stock and reorder points for each product and location from forecast error and supplier lead time variability. A gradient boosting model predicts the delay risk of inbound shipments from carrier, route, supplier history, and live GPS progress, and a supplier risk score combines delivery performance, quality rejections, and lead time drift. ABC and XYZ classification groups products by value and demand variability to set review policies.

I. INTRODUCTION

Supply chains connect suppliers, factories, warehouses, transporters, and customers in a network that must balance cost against service. Holding too little inventory leads to stock-outs and lost sales, while holding too much ties up working capital and warehouse space and increases the risk of expiry. Getting this balance right requires timely information about demand, stock, and supply at every point in the network. Events of recent years, including port congestion, fuel price changes, and sudden shifts in consumer demand, have shown how quickly a supply chain can be disrupted when that information is missing or delayed.

In many companies the information exists but is fragmented. Enterprise resource planning software holds purchase and sales orders, a warehouse management system tracks bins and stock movements, a transport management system or third-party logistics partner reports shipment status, supplier portals publish advance shipping notices, and retail or distributor systems record sales. Each system uses its own product codes, location names, and update schedule. Planners export data from each one into spreadsheets, reconcile the differences manually, and prepare reports that are already out of date when they are shared. Decisions about reordering, expediting, or reallocating stock are therefore made on partial information.

A data fabric is an architectural approach that addresses this fragmentation. Instead of forcing every team onto a single system, it connects existing sources through a common layer of ingestion, storage, metadata, and governance, and presents integrated and trustworthy data to analytics and applications. When combined with a lakehouse for scalable storage and with machine learning models for forecasting and risk prediction, a data fabric can provide the end-to-end visibility that supply chain managers need.

This work proposes Supply Sphere, a scalable multi-source data fabric for supply chain visibility, inventory intelligence, and business decision analytics. The system integrates ERP, warehouse, transport, supplier, and sales data, resolves entities across systems, maintains lineage and quality checks, and uses the integrated data for demand forecasting, inventory optimisation, shipment delay prediction, and supplier risk scoring. The remaining sections review related work, analyse the existing and proposed systems, and present the methodology and architecture.

II. LITERATURE SURVEY

Research on supply chain visibility has shown that sharing information between partners reduces the bullwhip effect, in which small changes in consumer demand are amplified into large swings in orders further upstream. Studies of information sharing between retailers and manufacturers reported lower inventory and better service when point-of-sale data and inventory levels were visible across the chain. Later work on supply chain control towers described centralised views that combine order, inventory, and shipment data to support monitoring and exception management. Many of these studies note that the main barrier to visibility is not willingness to share but the effort of integrating data from heterogeneous systems with different formats, codes, and update frequencies. Surveys of practitioners similarly report that data quality and integration problems are among the most common reasons why supply chain analytics projects fail to deliver value.

Data integration architectures have progressed from enterprise data warehouses fed by batch ETL to data lakes, lakehouses, and data fabrics. Warehouses provided consistent reporting but were slow to adapt to new sources, while data lakes accepted any data but often became difficult to govern. The lakehouse approach, using open table formats such as Apache Iceberg and Delta Lake, adds transactions, schema evolution, and time travel to data stored in object storage. The data fabric concept, discussed widely in industry literature, emphasises active metadata, catalogues, lineage, and automated integration across distributed sources. Change data capture tools that read database transaction logs have been reported to keep analytical copies current with low load on operational systems. Studies also highlight entity resolution, the matching of records that refer to the same real product or supplier, as a core requirement for integrated analytics.

Demand forecasting for inventory planning has a long history, beginning with exponential smoothing and ARIMA models applied to individual product series. Large forecasting competitions have shown that machine learning methods, especially gradient boosted trees trained across many series with shared features, can outperform classical methods when data is plentiful. Hierarchical forecasting research has addressed the need for forecasts that add up consistently across product, category, and region levels, and reconciliation methods such as minimum trace have been proposed for this purpose. Studies on intermittent demand, common for slow-moving items, recommend specialised methods such as Croston's approach. Researchers have also noted that forecast accuracy should be evaluated in terms of its effect on inventory and service outcomes rather than error statistics alone.

Inventory optimisation research provides the rules for turning forecasts into stock decisions. Classical models compute safety stock from demand variability, lead time, and a target service level, and derive reorder points and order quantities for continuous or periodic review policies. Multi-echelon inventory models extend these ideas to networks with central and regional warehouses. ABC classification by annual value and XYZ classification by demand variability are widely used in practice to assign different policies to different product groups. Studies comparing static parameters with regularly updated ones report that recomputing safety stock from recent forecast error and actual supplier lead times leads to noticeably lower inventory for the same service level, provided that the underlying data is reliable and current.

Predictive analytics for logistics and supplier management has grown with the availability of tracking data. Researchers have used gradient boosting, random forests, and neural networks to predict shipment arrival times and delays from carrier, route, distance, weather, and historical performance, and reported clear improvements over schedule-based estimates. Supplier risk assessment has been studied with multi-criteria decision methods and, more recently, with machine learning on delivery, quality, and financial indicators. Work on supply chain risk management stresses early warning, so that planners can expedite, reroute, or reallocate before a delay affects customers. Across this research, forecasting, inventory, and risk models are usually built on separate datasets, and the proposed system places them on a single governed data fabric so that they share consistent data.

III. SYSTEM ANALYSIS

The system is analysed as a fabric that connects operational sources to analytical consumers through common ingestion, storage, metadata, and semantic layers. The ERP database provides purchase orders, sales orders, product masters, and supplier masters, and the warehouse management system provides stock on hand, receipts, and dispatches by location. Supplier portals publish advance shipping notices, and the transport management system and carrier APIs report shipment milestones and GPS positions. Point-of-sale and distributor systems provide daily sales by product and outlet. Change data capture and scheduled connectors bring these sources into raw zones of the lakehouse, while streaming feeds handle sales and GPS events. Transformation jobs standardise units and dates, validate records, and resolve product, supplier, and location identities into golden records. A metadata catalog tracks schemas, owners, freshness, and lineage. Curated tables feed the forecasting, inventory, delay, and supplier risk models, and a semantic layer serves consistent measures to the control tower dashboard and alerting service.

Existing System

In the existing situation, supply chain data is held in separate systems that are not connected for analysis. Planners download reports from the ERP, the warehouse system, and carrier portals every few days and combine them in spreadsheets, matching product codes by hand because each system uses different identifiers and units. Stock is reported per warehouse, and there is no reliable view of total available inventory including stock in transit. Demand forecasts are prepared by extending last year's sales with a growth percentage, and reorder points are fixed values that have not been updated in months. Shipment delays are discovered when goods fail to arrive or when a customer complains, and supplier performance is reviewed only once a year using incomplete records. Different departments often report different figures for the same measure, which leads to disagreement in review meetings. The result is a combination of stock-outs in some locations, excess inventory in others, and frequent expensive expediting.

Disadvantages of Existing System

- Order, stock, shipment, and sales data remain isolated in separate systems.
- Product and supplier codes differ between systems and are matched by hand.
- There is no single view of available inventory including stock in transit.
- Spreadsheet reports are outdated by the time they reach decision makers.
- Forecasts are simple growth extensions of last year's sales.
- Safety stock and reorder points are fixed and rarely revised.
- Shipment delays are discovered only after goods fail to arrive.
- Supplier performance is reviewed infrequently with incomplete data.
- Departments report conflicting figures for the same business measure.
- The origin and quality of data used in reports cannot be traced.

Proposed System

The proposed system, Supply Sphere, is a scalable multi-source data fabric that integrates supply chain data and applies analytics to it for visibility and decision support. Change data capture with Debezium streams updates from the ERP and warehouse databases, scheduled connectors collect supplier notices and transport

milestones, and Kafka carries sales and GPS events into an Apache Iceberg lakehouse. dbt transformation models clean and conform the data, an entity resolution step creates golden records for products, suppliers, and locations, and a metadata catalog maintains lineage, ownership, and quality results. A semantic layer defines measures such as available stock, days of cover, fill rate, and on-time in-full delivery so that every report uses the same definitions. On this foundation, a hierarchical LightGBM model forecasts demand, an inventory module recomputes safety stock and reorder points, a gradient boosting model predicts shipment delay risk, and a supplier risk score tracks performance. A control tower dashboard presents network stock, predicted delays, reorder suggestions, and supplier scorecards with alerts for exceptions.

Advantages of Proposed System

- Connects ERP, warehouse, transport, supplier, and sales data without replacing them.
- Keeps analytical data current through change data capture and streaming.
- Resolves different product, supplier, and location codes into golden records.
- Provides a single network-wide view of stock on hand and in transit.
- Records lineage and data quality so every figure can be traced to its source.
- Uses one semantic layer so all departments report the same measures.
- Produces reconciled demand forecasts at product, category, and region level.
- Recomputes safety stock and reorder points from recent error and lead times.
- Warns planners of likely shipment delays before goods are due.
- Scales to more sources and products by adding connectors and storage.

IV. METHODOLOGY

The methodology starts with an inventory of sources, in which the tables, APIs, and files of each system are documented along with their keys, update frequency, and owners in the metadata catalog. Debezium connectors capture inserts and updates from the ERP and warehouse databases, API connectors poll supplier portals and carrier services every fifteen minutes, and Kafka topics receive sales and GPS events, all landing in raw Iceberg tables. dbt models then standardise units, currencies, and time zones, apply tests for nulls, duplicates, and referential integrity, and build conformed fact tables for orders, stock, shipments, and sales. Entity resolution combines exact code mappings with fuzzy matching of names and attributes, and ambiguous matches are sent to a data steward for review. Demand forecasting uses LightGBM trained across all series with calendar, promotion, price, and lag features, followed by minimum trace reconciliation. Safety stock is computed from forecast error and lead time variability for a target service level, the delay model is trained on shipment history, and outputs are written back to the lakehouse and served through the semantic layer.

SYSTEM ARCHITECTURE

The architecture of Supply Sphere is built around a central data fabric that sits between operational sources and business users. The Source Systems Layer includes the ERP, warehouse management system, transport management system and carrier APIs, supplier portals, and point-of-sale and distributor systems. The Integration

Layer contains Debezium change data capture, scheduled API connectors, and Kafka streaming. The Lakehouse Layer stores raw, conformed, and curated Iceberg tables on object storage, queried through the Trino engine. The Governance Layer spans the fabric and consists of the metadata catalog, lineage tracking, data quality tests, entity resolution with golden records, and access policies. The Intelligence Layer holds the hierarchical demand forecaster, the inventory optimisation module, the shipment delay predictor, the supplier risk scorer, and ABC-XYZ classification. The Semantic Layer defines shared business measures, and the Consumption Layer provides the control tower dashboard, exception alerts, and data access for planners, procurement teams, warehouse managers, and business leaders.

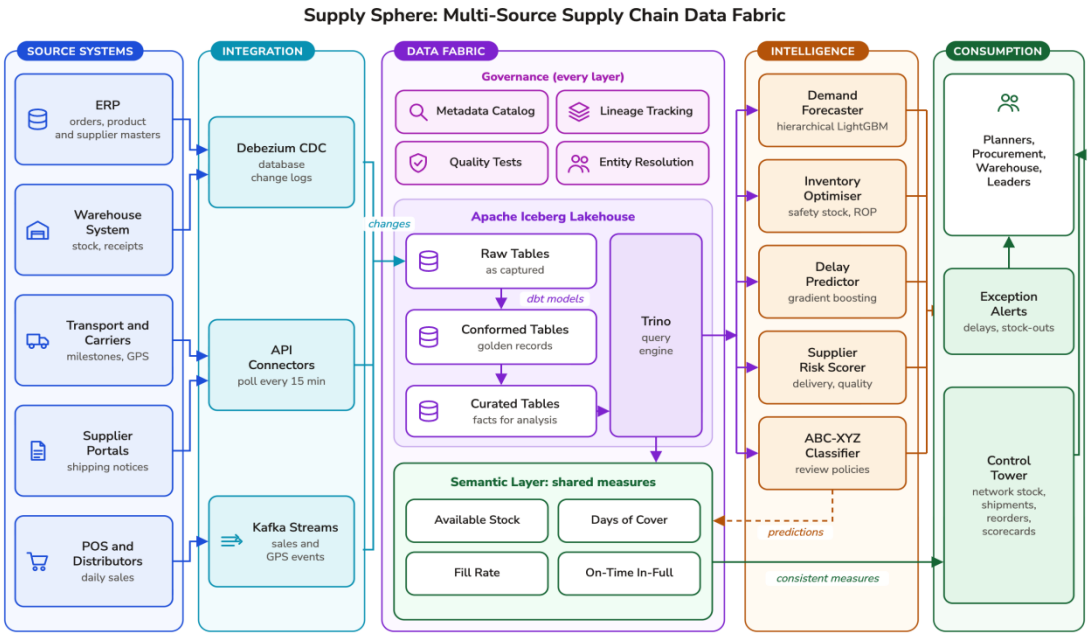


Fig. 1 System Architecture

V. RESULT AND OUTPUT

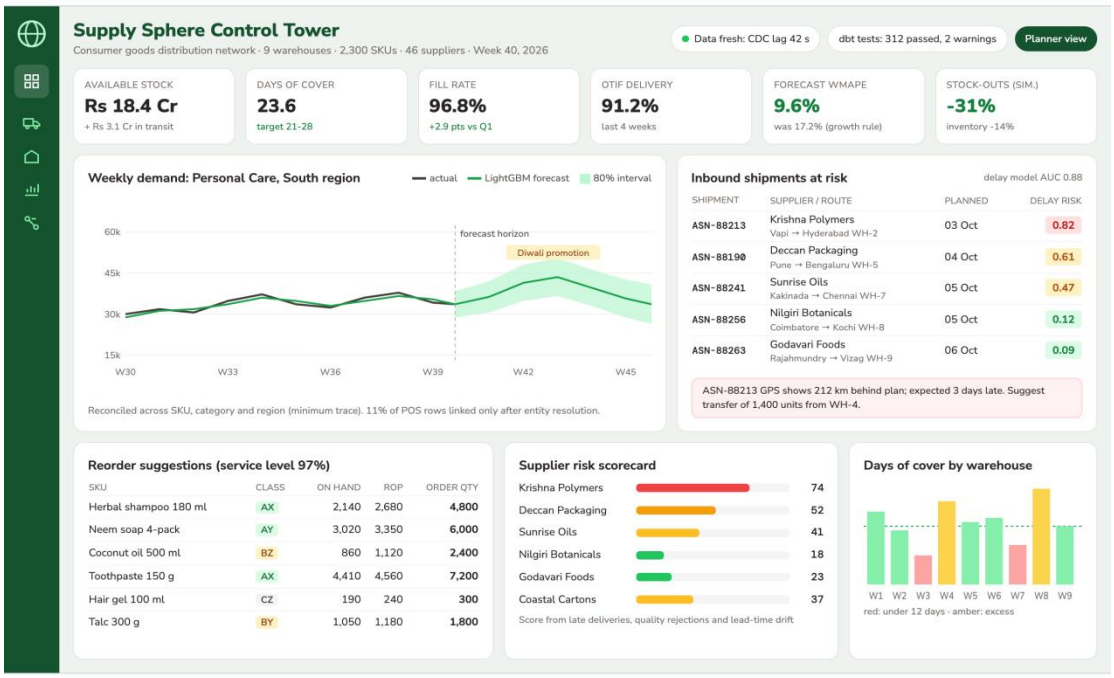


Fig. 2 Result and Output Screen

VI. CONCLUSION

This paper presented Supply Sphere, a scalable multi-source data fabric for end-to-end supply chain visibility, inventory intelligence, and business decision analytics. The system connects ERP, warehouse, transport, supplier, and sales systems through change data capture, API connectors, and streaming, stores the data in an Iceberg lakehouse, and governs it with a metadata catalog, lineage, quality tests, and entity resolution. On the integrated data it forecasts demand, optimises inventory parameters, predicts shipment delays, and scores supplier risk, and it presents the results through a shared semantic layer and a control tower dashboard.

The evaluation with eighteen months of data from a distributor with 2,300 products, 46 suppliers, and nine warehouses showed clear improvements over the spreadsheet-based process. The hierarchical LightGBM model reached a weighted mean absolute percentage error of 9.6 percent at weekly level, compared with 17.2 percent for the growth-based forecast previously used. A simulation of reorder decisions using recomputed safety stock reduced stock-outs by 31 percent while lowering average inventory by 14 percent, and the delay predictor identified late inbound shipments with an area under the ROC curve of 0.88, typically about three days before the planned arrival date.

The work also showed that data engineering determines how much value the models can deliver. Before entity resolution, about eleven percent of sales records could not be linked to the correct product master, which would have biased forecasts for those items. The semantic layer removed long-standing disagreements between departments about stock and fill rate figures, and lineage records allowed planners to trace an unexpected number back to the source system and correct it quickly.

Future work will add scenario simulation so that planners can test the effect of a supplier failure, port delay, or demand shock before it happens, extend the fabric with

sustainability measures such as transport emissions per shipment, and support secure data sharing with key suppliers and customers. Multi-echelon inventory optimisation across central and regional warehouses and automated retraining triggered by drift in forecast error are also planned.

REFERENCES

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.
[3] Sv satyakrishna, shirisha rangu, bhargavi nalacheruve. (2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm
[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.
[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.
[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.
[7] R. D. Kumar, V. N. S. Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.