International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

Vol. 43 No. 2 March 2024, pp. 262-274

An Efficient Supply Chain Data Warehousing Model For Big Data Analytics

[1]Odochi C-E Iheukwumere, [2]Udoka F. Eze, [3]Anthony I. Otuonye, [4]Charles O. Ikerionwu

[1, 2, 3]Department of Information Technology, Federal University of Technology, Owerri, Nigeria.

[4]Department of Software Engineering, Federal University of Technology, Owerri, Nigeria.

E-mail:

[1]chinwedozie1@gmail.com, [2]udoka.eze@futo.edu.ng, [3]anthony.otuonye@futo.edu.ng,
[4]charles.ikerionwu@futo.edu.ng

Abstract - **This research work is aimed at developing a supply chain data warehousing model for big data analytics that will be used for reporting and analysis purposes. Objected-Oriented Design methodology was adopted for the study. A big data supply chain dataset of a retail outlet from a real world business transaction was used for data analysis. Google storage bucket was created in Google BigQuery for storage and analysis of the data. Data was uploaded into Google storage in the cloud, after which the supply chain data table was created using SQL query. Star Schema dimensional model was created for integrating data into the cloud. For descriptive and diagnostic analytics including feature engineering, the integrated datasets, advanced feature engineering techniques were applied to create derived variables that enhanced the model interpretability and predictive power. Google big Query was linked to Google collab for big data analytics, after which a preliminary analysis was conducted in Google collab showing the first row of the dataset. There was then a decomposition of the time series analysis into trend, seasonality, residuals and original. To perform predictive analytics, the processed dataset was split into training and test datasets to prevent over-fitting. To optimize the model performance, the hyperparameters were adjusted. The forecasting model was implemented within the dashboard using ARIMA and Prophet time series forecasting methods in training the models; and Random forest regression machine learning model in order to implement the most important features that drives sales as well as demand. MAPE and RMSE were used as model evaluation metrics for the predictive analytics of the proposed model. After cross validation of the performance metrics, the study revealed that incorporating advanced Prophet, ARIMA and Random Forest models enhanced the predictive capabilities of the proposed system, leading to more precise inventory management. In conclusion, the proposed system offers better improvements with respect to reliability, performance, scalability, and recoverability because it is designed to handle complex, large scale data operations which are very crucial in modern business environments. The proposed supply chain data warehousing model for big data analytics is highly recommended for supply chain management/managers in inventory management, as the model will help in optimizing the inventory levels as well as improving the supply chain business.**

Keywords – Supply Chain, Data Warehousing, Big Data, Analytics.

I. INTRODUCTION

In today's competitive business environment, in which competition is among supply chain networks, businesses are confronted with the need to effectively manage increasingly supply chain activities beyond their boundaries. Big data analytics in supply chain management (SCM) is evolving quickly as a result of recent advancements in machine learning and computing infrastructure; the global big data market size is anticipated to grow from USD 138.9 billion in 2020 to USD 229.4 billion by 2025, at a compound annual growth rate (CAGR) of 10.6% throughout the forecast period. Big data analytics is predicted to expand at the highest CAGR of all the solutions throughout the projection period and to hold the largest market share [1]

The current business climate necessitates quick decisions in the supply chain and shorter turnaround times from raw data to insights and actions. Managing the complexity in supply chain has been of importance to industries that compete in the global market as the complexity in supply chain management is associated with material and information flow between the different supply chain entities considering how information flows in this digital age [2]. Supply chain which can be considered as a significant contributor to big data wherein the diversity of information is large encompasses from raw materials to producer/central organization, wholesalers, retailers, customers, and end users [3]. There are multiple sources of big data across the supply chains with varied trade-offs among volume, velocity, variety, value and veracity attributes [1]. Thus the combination of data sources with their diverse temporal and spatial attributes places a greater emphasis on the use of big data analytics which has been described as a method used to find hidden patterns and shed light on intriguing relationships in contexts by examining, processing, finding, and displaying results [4] in supply chain management.

The term data warehouse is a sizeable, central repository of integrated and arranged data from several sources that is made to assist reporting, analytics, and business intelligence activities. It is the most dependable and frequently used technology for planning, predicting, and managing businesses, as well as the data storage facility that houses a large collection of data [5]. On the other hand data warehousing model is a conceptual framework and organizational structure of a data warehouse that ensures effective data storage, integration and retrieval.

Though supply chain management focuses on effective and efficient creation of value for customers by optimizing the flow of products and services; however, one of the major challenges of data warehousing optimization is the increasing variety, velocity, and volume of data that is generated from varied sources, and in different formats [6]. This has created a need for an efficient supply chain data warehousing model that can handle the complexity and diversity in the process of data integration, cleansing, transformation and loading (ETL) of big data as well as storing the data into a common structure that can be easily accessed and used for reporting and analysis purposes. Hence, the aim of this study is to develop an enhanced supply chain data warehousing model for big data analytics that will identify and integrate data sources from operational databases; use a cloud based data storage; perform advanced analytics for business intelligence and improve insights to enhance forecasting, inventory management and optimization.

II. RELATED WORKS

In the existing literature, a lot of attempts have been made towards solving the supply chain analytic challenges in data warehousing through advanced analytics. [7] in their study assessed the idea of a data warehouse with its design and concluded that data warehouse is generally useful for associations but suggested that future framework should be developed for execution of data warehouse and modern architecture of warehousing based on the cloud. [8] in a study on data warehouse design for big data in academia advocated for a data warehouse design that will involve large data sets and more refined operations.

[9] Talked about the various big data analytics technologies that could be used, including machine learning, data mining, and intelligent analysis, as well as the challenges that come with big data, including data visualization, cloud and distributed computing, and computational difficulties

[10] Concentrated on machine learning, its uses in big data, difficulties, and technological advancements in big data. [11] Developed a data warehousing architecture for big data capable of adapting automatically or semi automatically to user requirements and changes in the underlying data. [12] Proposed a Hybrid Lake Data Warehouse Architecture model that uses Hadoop framework and Apache spark to merge traditional data warehouse strategies. [13] Proposed a multi-layered supply chain big business intelligence model that utilizes cloud-based big data services and tools for data extraction, transformation and

loading, analysis and reporting. [14] Proposed a Decision-Support System (DSS) supported by a Big Data Warehouse (BDW) and a simulation model.

[15] Developed an intelligent forecasting system that is based on the analysis and interpretation of historical data using SVR algorithm, Multi Layer Feed-forward Artificial Neural Network (MFLANN), Exponential Smoothing Model, Holt-trend Method, Holt-Winters Seasonal model, ARIMA, and Moving Average. [16] Focused on comparing ARIMA modeling with SVR on three different and unrelated datasets. [1] In their systematic literature review study suggested that future research should focus on alternative theoretical perspectives and analytical methods on the role of Big Data Analytics in supply chain should be focused on providing integrated frameworks for the selection and the use of big data analytics techniques in supply chain management.

[17] In their work proposed a demand forecasting method based on multi layer networks. Their proposed model considered different combinations of LSTM parameters for a given time series using grid search method. When compared with both statistical and computational intelligence models such as ARIMA, ANN, KNN, SVM, Exponential Smoothing, and single layer LSTM. [18] In their study proposed a hybrid approach that integrates Prophet and SVR models to forecast time series demand in the manufacturing industry with seasonality. [19] Considered the benefits and challenges of cloud and edge computing; and how large organizations can adopt those two computing models in conjunction to create dynamic applications.

[20] Used a customer-responsive time series approach that is is based on algebraic methods of estimation; and suggested that the relevance of the method combined with an inventory control technique and stock optimization should be investigated. [21] Developed a hybrid forecasting model that combined ARIMA and the Holt's Winter Model (ARHOW) on selected therapeutic class. [22] Implemented a cosmetic supply chain as a multi-agent system in which data generated by the simulator were used in order to identify genuine trajectories across the whole supply chain. [23] Proposed new smart business architecture of market prediction to forecast accurate future demand using Prophet and ARIMA hybrid model.

[24] In their study implemented a data warehouse using Kimball Lifecycle method and the PostgreSQl database that will be used to monitor sales performance. [5] In their study analyzed the performance of the data warehouse architectures by comparing many works in the field of data warehouse. Their study identified extensive data warehousing areas that are still active for future research as there are still concerns with managing a huge data by using data warehouse tools that classified them logically. [25] Revealed that visualization, compression, distributed mining, analytics architecture and hidden data are the future important challenges of big data analytics and management. [26] Revealed that while there are some evidences that big data analytics creates value, but that the claim that investments in data analytics can be a source of competitive performance gains requires a deeper analysis.

III. METHODOLOGY

Object-Oriented Design (OOD) was used to create structured and modular systems that effectively represents the entities, relationships, and processes within the supply chain.

3.1 Dataset

The proposed system was implemented and tested on a big data supply chain dataset of a retail outlet from a real world business transaction. The datasets are housed in the Mendeley data repository and is licensed under Creative Commons 4.0 and can be accessed at **(https://data.mendeley.com/datasets/8gx2fvg2k6/5).**

3.2 System Architecture

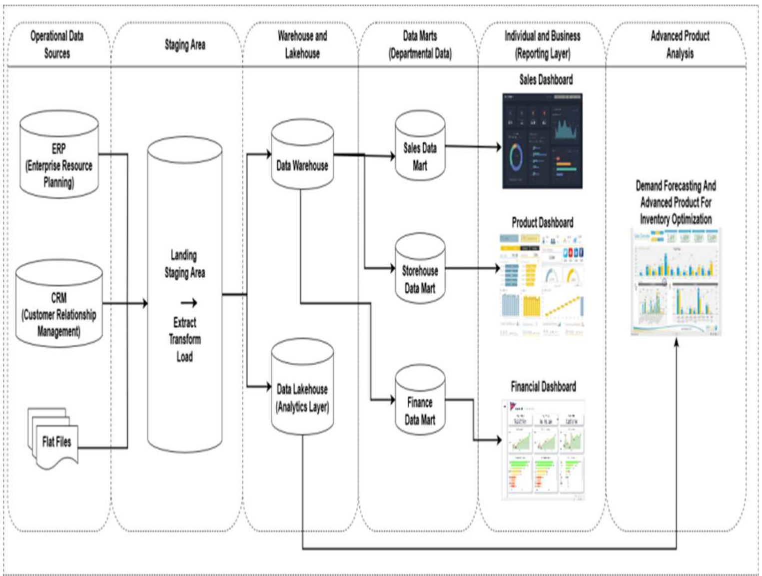

Figure 1: Proposed System Architectural Framework

Figure 1 describes the conceptual overview and model of how the proposed system is arranged.

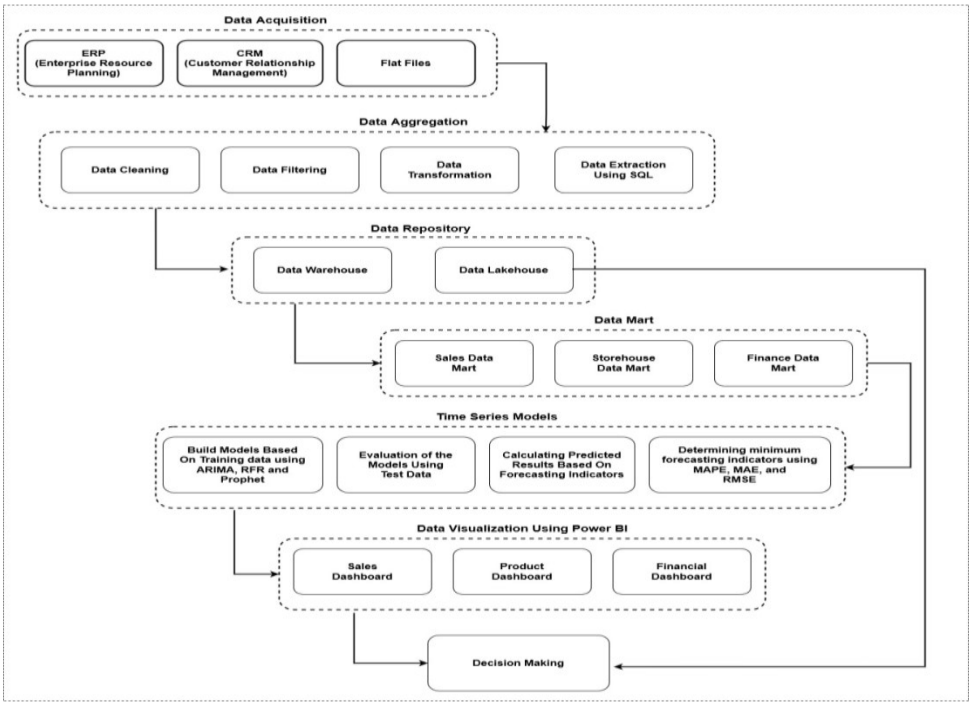

Figure 2: Data Flow Diagram of the proposed system

Figure 2 depicts the flow of data in the proposed model. Data is extracted from the operational databases and are accessed via direct queries or through API endpoints that support Extraction, Transform, and Load (ETL) operations. The data is further aggregated to focus on each department or subject matter, hence data marts are created from the data repository. To perform predictive analytics, time series models will be built based on training data; the models will be evaluated using the test data, while the forecasting assessment will be done based on the forecasting indicators using MAPE, MAE, and RMSE. After which the data will be visualized on individual subject using Power BI to enable fast and easy decision making by stakeholders.

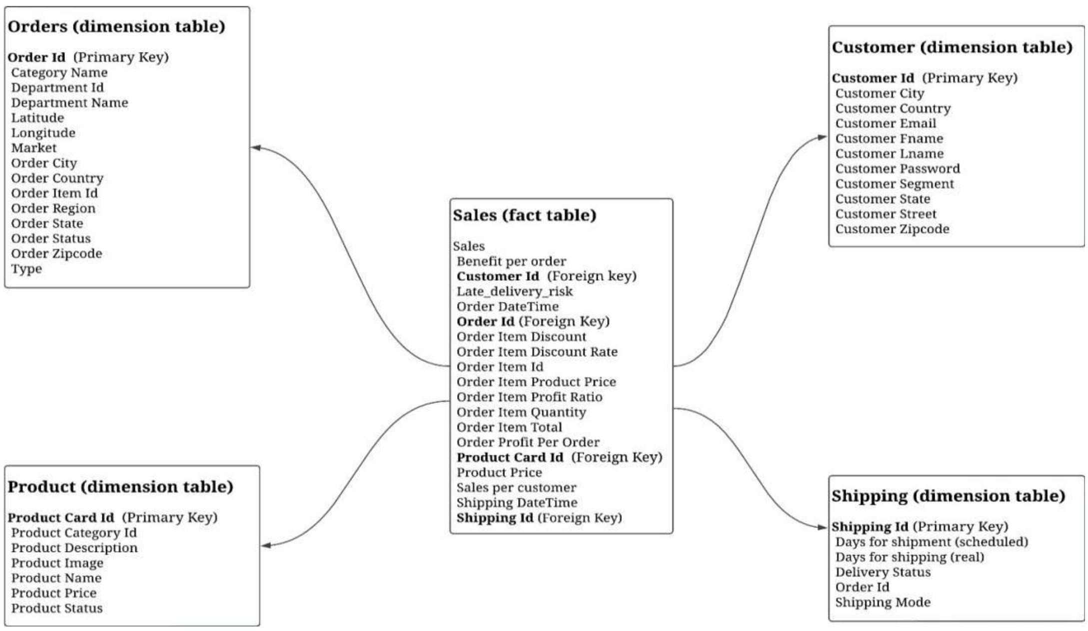

Figure 3: Star Schema Model

3.3 Experimental Setup

This section provides a summary of the performed experiment in this research. Data was downloaded from the Enterprise Resource system and uploaded into Google Big Data warehouse (Google BigQuery). The Google storage bucket was then created in Google BigQuery for storage and analysis of the data. Then data was uploaded into Google storage in the cloud, after which the supply chain data table was created using SQL query. Star Schema dimensional model was created for integrating data into the cloud. For descriptive and diagnostic analytics including feature engineering, the integrated datasets, advanced feature engineering techniques were applied to create derived variables that enhanced the model interpretability and predictive power. Google big Query was linked to Google collab for big data analytics, after which a preliminary analysis was conducted in Google collab showing the first row of the dataset. There was then a decomposition of the time series analysis into trend, seasonality, residuals and original.

To perform predictive analytics, the processed dataset was split into training and test datasets to prevent over-fitting. Considering that the dataset spanned within three years; the first two years was used as the training dataset in order to train the predictive model, the third year was used as testing dataset. To optimize the model performance, the hyperparameters were adjusted. The forecasting model was implemented within the dashboard using ARIMA and Prophet time series forecasting methods in training the models; and Random forest regression machine learning model in order to implement the most important features that drives sales as well as demand. MAPE and RMSE were used as model evaluation metrics for the predictive analytics of the proposed model.

IV. Result and Discussion

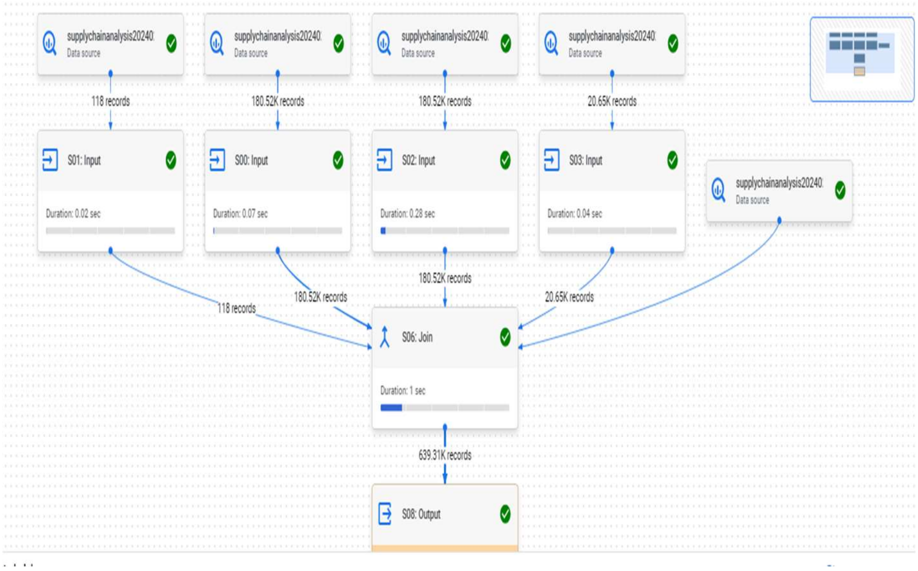

Figure 4: Integration of data sources using Google Big Query data warehouse

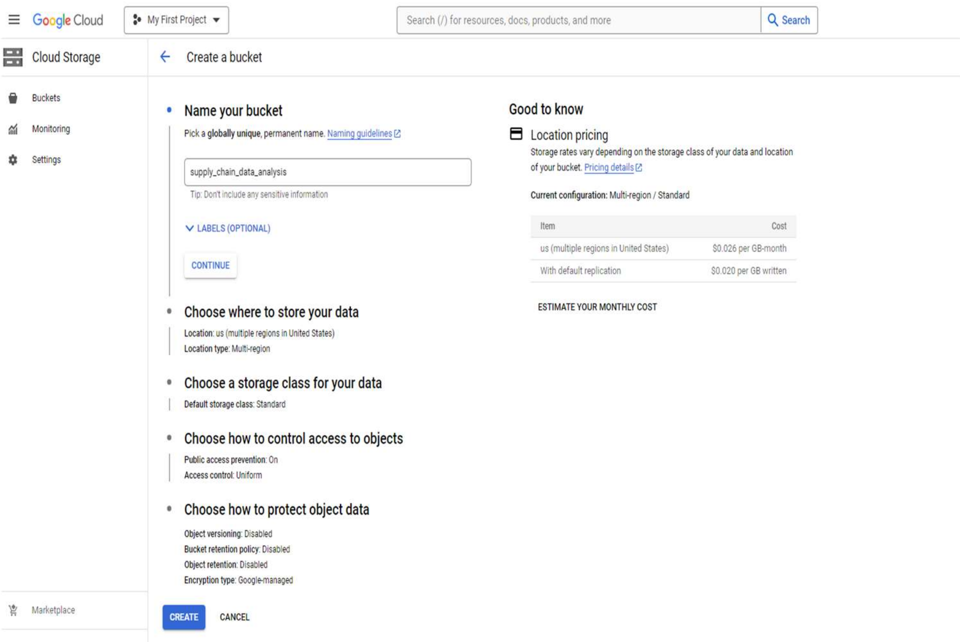

Figure 5: Creating big data storage bucket in Google cloud

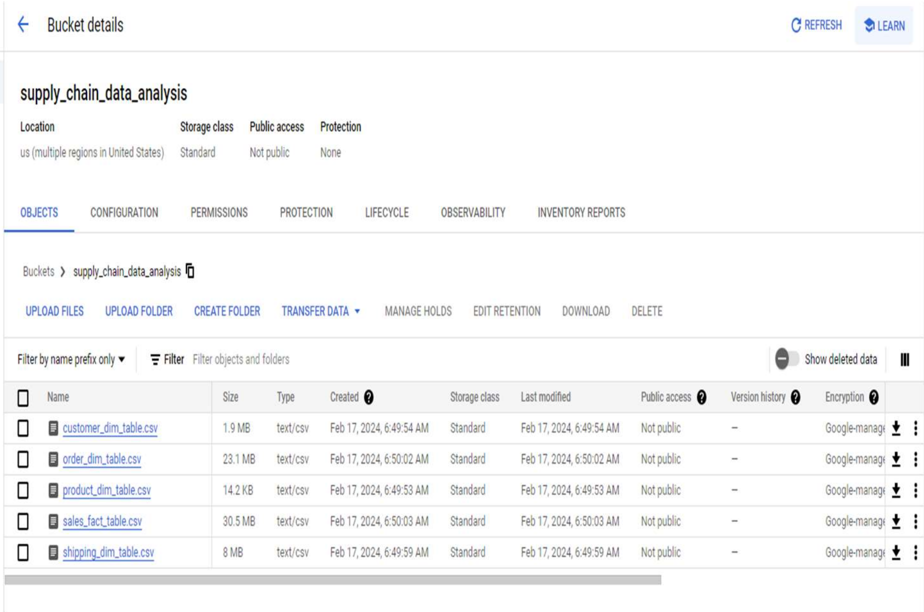

Figure 6: Big data imported into Google Big query data warehouse from Google Big data storage bucket

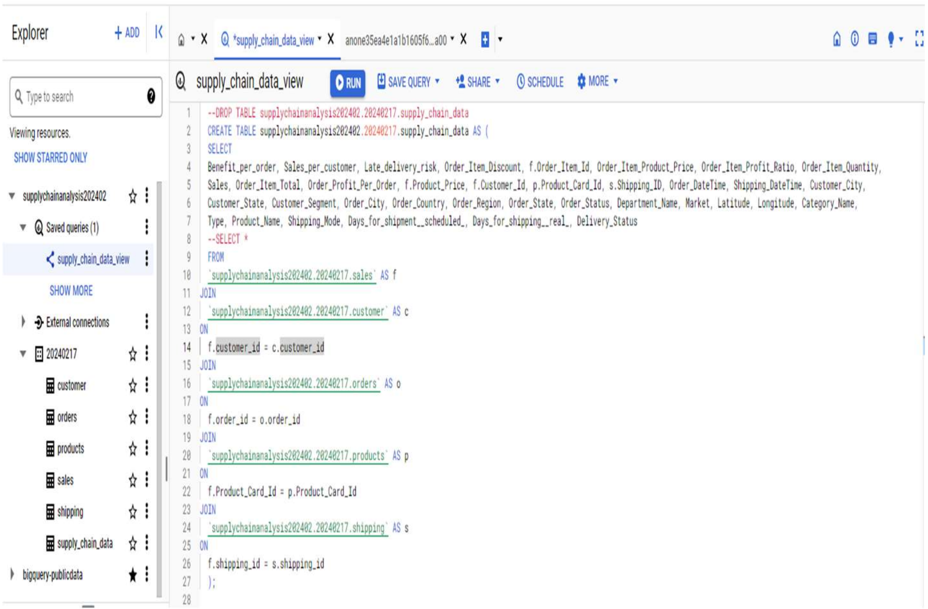

Figure 7: SQL query for creating the required table for analysis within Google Big query data warehouse

4.1 Descriptive and Diagnostic analytics including feature Engineering

Advanced feature engineering techniques are applied to create derived variables that enhance model interpretability and predictive power. These features are crucial for developing robust machine learning models for prediction.

BigQuery bquxjob_6cd8736d_18db5f67803

File Edit View Insert Runtime Tools Help Cannot save changes

+ Code + Text Copy to Drive

Step 1: Install the BigQuery Client Library Install the BigQuery client library in Colab notebook.

```
[ ] #!pip install --upgrade google-cloud-bigquery pandas gbq
```

Step 2: Authenticate Your Session To access BigQuery from Colab, you need to authenticate your session. Colab provic
do this with its built-in authentication module.

```
[ ] from google.colab import auth
    auth.authenticate_user()
    print('Authenticated')

    Authenticated
```

Step 3: Set Your Google Cloud Project ID You need to specify which Google Cloud project you'll be billing queries to.

```
[ ] project_id = 'supplychainanalysis202402'
```

Figure 8: Google Big Data Query is now linked to Google collab for big data analytics.

Step 4: Run a Query on BigQuery Now you're ready to run a query on your BigQuery dataset and load the results into a pandas DataFrame. You can use the google.cloud.bigquery client or the pandas-gbq library.

```
[ ] from google.cloud import bigquery

    client = bigquery.Client(project=project_id)

    # Replace this SQL query with your actual query
    sql = """
    SELECT * FROM `supplychainanalysis202402.20240217.supply_chain_data`
    LIMIT 100
    """

    df = client.query(sql).to_dataframe()
    df.head()
```

	Benefit_per_order	Sales_per_customer	Late_delivery_risk	Order_Item_Discount	Order_Item_Id	Order_Item_Product_Price
0	-192.789993	145.500000	0	4.5	121314	50.000000
1	-192.789993	145.500000	0	4.5	121314	50.000000
2	-192.789993	145.500000	0	4.5	121314	50.000000
3	-100.790001	335.980011	0	64.0	121315	399.980011
4	-100.790001	335.980011	0	64.0	121315	399.980011

5 rows × 36 columns

Figure 9: Preliminary analysis in Google collab showing the first rows of the dataset.

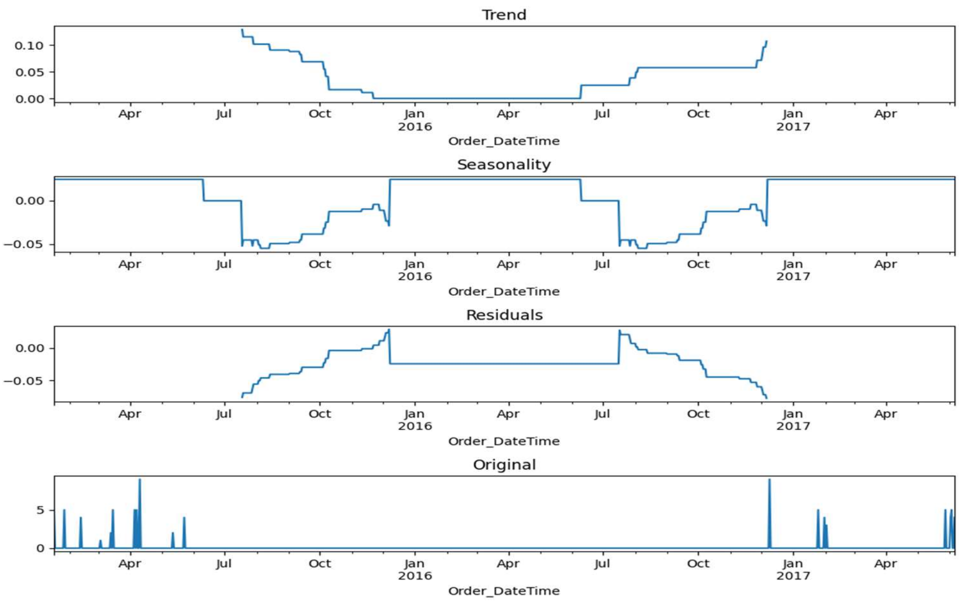

Figure 10: Decomposition of time series analysis

Figure 10 shows the decomposition of the time series analysis:

i. **Trend**: Is the long-term progression of the data when fluctuations due to seasonal effects or irregular components have been removed. There appears to be a slight upward trend over time.

ii. **Seasonality**: Is the repeating patterns or cycles within the time series sales data within a monthly fixed period. The chart does not show any clear seasonality patterns.

iii. **Residuals**: Is what remains in the data after the seasonal components and trend have been removed from the original data. It is essentially the noise that cannot be attributed to trend or seasonality.

iv. **Original**: Is the actual time series data showing the raw data points over time representing sales.

4.2 PREDICTIVE ANALYTICS

4.2.1 Forecasting using Prophet Model

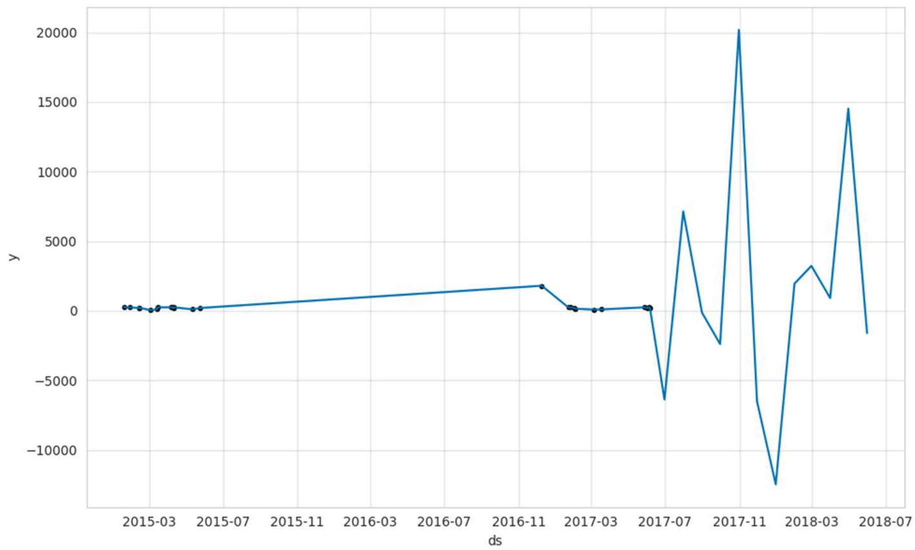

Figure 11: Forecast by using the Prophet model

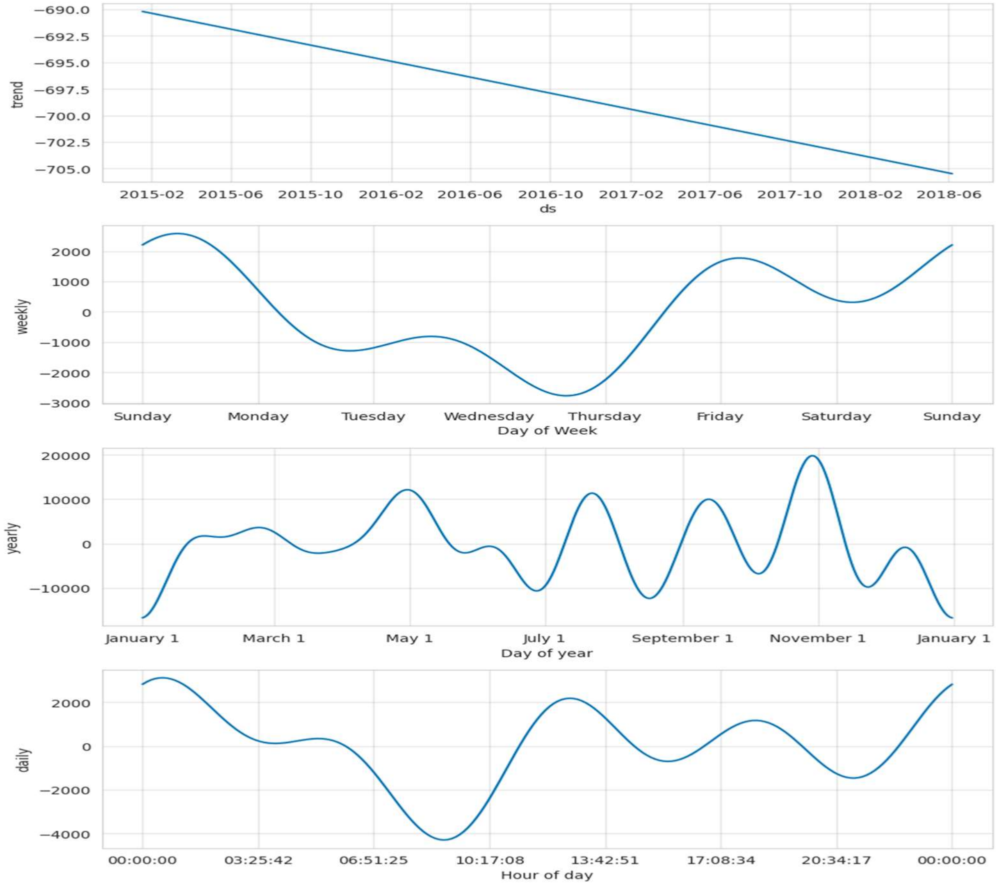

Figure 12: Decomposition of the Prophet Forecast model.

Figure 11 shows the point on the left side representing the historical data for the variable 'y' over time ('ds' on the x-axis) representing the actual data the model has been trained on; while the lines extends to the right beyond the historical data represents the model's forecast for future periods. Initially the data shows a relatively stable trend around the zero mark suggesting consistent 'y' values with some fluctuations, and a period of stability. There are significant spikes and dips in the data. The continued fluctuations in the trend suggests that the model expects sales which is the variable 'y' to continue to have similar variability in the future, as seen in the latter part of the historical data.

Figure 12 illustrates the components of the vanilla model and the tuned models; the trend, weekly seasonality, yearly seasonality and daily seasonality. As seen from figure 14, there is a general downward trend overtime suggesting a period of sales decrease early 2015 to mid-2018. The weekly seasonality indicates lower sales starting from Friday; the yearly seasonality shows some yearly pattern with November accounting for higher sales activity; and lower sales activity in January; while the daily sales seasonality shows a daily pattern of high sales activity around lunch time 12 noon and lowest sales activity in the morning around 9 am.

Table 1: Prophet Model Comparison of Vanilla and the Tuned Model

	Vanilla	**Tuned model**
RMSE	4,899	616
MAPE	20 (20%)	2 (2%)

Table 1 indicates that the vanilla model has an RMSE of 4,899 and MAPE of 20% while the tuned model has RMSE of 616 and MAPE of 2%. This shows that the vanilla model is better than the tuned model due to the lower error (RMSE value) of 616 and lower MAPE of 2%.

4.2.2 Forecasting using ARIMA Model

To check for stationarity Augmented Dickey-Fuller (ADF) test was conducted with an output of: ADF Statistic: -18.017500; p-value: 0.000000, which suggested that the time series data was stationary a p-value of 0.000, which is less than the common significance level threshold of 0.05. Forecasting for the vanilla model using basic p,d,q, values and showing the evaluation metrics RMSE and MAPE. The result showed RMSE of 19940.010891014408, and MAPE of 99.40585184361528. While the tuned model showed RMSE of 27208.655788235134 and MAPE of 6948.298258019693. This indicates that the hyperparameter-tuned model has worse RMSE and MAPE values when compared with the vanilla model, because of its over-fitting to the training data. This means that while it performs exceptionally well on the training dataset, it fails to generalize to the unseen test data. Overfitting possibly occurred because the chosen parameters (p, d, q) for the ARIMA models are too complex, capturing noise in the data as if it were a signal.

	RMSE	MAPE
Vanilla ARIMA model	**19940.010891014408**	**99.40585184361528**
Tuned ARIMA model	**27208.655788235134**	**6948.298258019693**

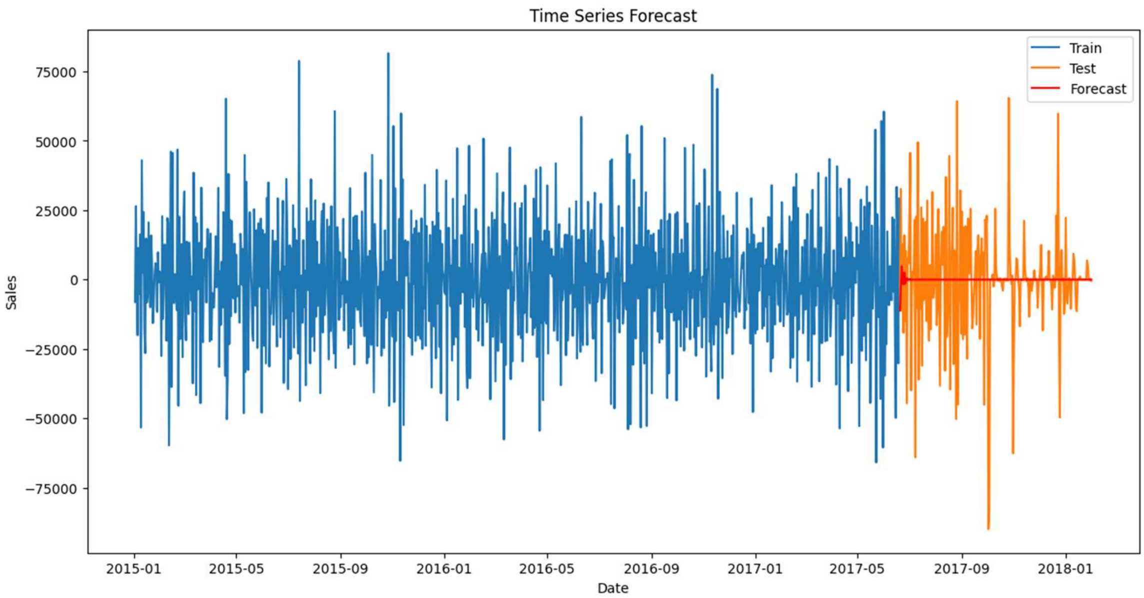

Figure 13: The ARIMA time series data showing the training, test and forecast data.

Figure 13 shows the training data (blue), test data (orange), and forecast data (red). The figure shows that the forecast does not capture the variability of the test data well, which is the reason for the high RMSE and MAPE values observed.

4.2.3 Identifying drivers of Sales using Random Forest Regression.

In order to build the Random forest regression feature (x) and target (y) variables were selected; one-Hot encoding was used to convert the categorical variables into numerical, then a column transformer was created to transform the data as the random forest regression model does not work with missing values. The preprocessed dataset was split into training and testing sets, with 80% training and 20% testing. The Random forest regression vanilla mode demonstrates that with the R-squared at 99.7%, the variance in the sales is captured by the features used in modeling. The RMSE at 6.44 which is high also demonstrates that the difference between the predicted and training data is not very high. The random forest regression hyperparameter tuning and model optimization shows that the RMSE has been reduced significantly from 6.44 (before training) to 1.77 (after training). That is 73% improvement in the model performance.

V. CONCLUSION AND FUTURE WORK

In this research work the enhanced supply chain data warehouse for big data analytics offers better improvements with respect to reliability, performance, scalability, and recoverability because it is designed to handle complex, large scale data operations which are very crucial in modern business environments. Thus, the proposed system's design makes it inherently flexible and scalable as data volumes and business complexity grows. These improvements come at a cost though usability, compatibility, and maintainability are considered. Furthermore, incorporating advanced time series (Prophet, ARIMA) and machine learning (Random Forest) models enhanced the demand forecasting capabilities of the proposed system, leading to more precise inventory management. Future extension of this work may include advanced model enhancement by the application of other forecasting methods, real-time streaming data processing, integrating other sources such as GPS locations of their logistics division. Furthermore, the business will need to do a long-term impact assessment and cost-benefit analysis to ensure that value derived by the new system implementation outweighs the cost.

REFERENCES

[1]. Lee, I., & Mangalaraj, G. (2022). Big data analytics in supply chain management: A systematic Literature Review and Research directions *Big Data Cognitive. Computing,* 6(17)

[2]. Awwad, M. H., Kulkarni, P., & Bapna, R. (2018). Big data analytics in supply chain: a literature review. *Proceedings of the International conference on Industrial Engineering and Operations Management, September,* 27-29.

[3]. Saeid, S. D., Iman, R. V., Farzaneh, M. M. (2020). Big Data Analytics and its Applications in Supply Chain Management. New Trends in the use of Artificial Intelligence for the Industry 4.0. DOI: http://dx.doi.org/10.5772/intechopen.89426

[4]. Vidhya, K. & Shanmugalakshmi, S. (2020). Modified adaptive neuro – fuzzy inference system (M – ANFIS) based multi-disease analysis of healthcare Big. *Journal of Supercomputing,* 012345689: 1-22.

[5]. Merceedi, J. K., Yazdeen, A. A., Ibrahim, A. K., Abdulrazzaq, M. B. & Mahmood, M. R. (2022). Analysis of the performance of data warehouse architecture types. *Journal of Soft Computing and Data Mining, 3(1):* 45-57.

[6]. Ghuman, B. (2023). What are the key challenges and opportunities for data warehouse optimization in the era of AI and ML? *Data Warehouse Architecture* https://www.linkedin.com/advice/0/what-key-challenges-opportunities-data

[7]. Saxena, G. D., Elur, N. R., David, D. S., Kumbhar, S. T. & Premkumar, W. (2022). Picture of data warehouse portrayed with thought for challenges, modernization and improvement of a conventional data warehouse as well as its possible future perspectives. *Journal of Pharmaceutical Negative Results, 13(9),* 8244 -8254.

[8]. Rudniy, A. (2022). Data warehouse design for b ig data in academia. *Computers, Materials & Continua, 71(1),* 979-992

[9]. Acharjaya, D. & Anitha, A. (2017). A comparative study of statistical and rough computing models in predictive data analysis. *Int. Journal Ambient Computing Intelligent, 8(2),* 32-51.

[11]. Agapito, G., Zucco, C. & Cannataro, M. (2020). COVID-warehouse: a data warehouse of Italian COVID-19, pollution, and climate data. *International Journal of Environ. Res. Public Health; 17(15),* 1-22.

[12]. Saddad, E., El-Bastawissy, A., Mokhtar, H. M. O. & Hazman, M. (2020). Lake data warehouse architecture for big data solutions. *Int. Journal, Adv. Comput. Sci. Appl.; 11(8),* 417-424.

[13]. Stefanovic, N. & Milosevic, D. (2017). Model for big data analytics and management. *7th International Conference of Information Society and Technology (ICIST),* 163-168.

[14]. Vieira, A. A.C., Dias, L. M. S., Santos, M. Y, Pereira, G. A. B. & Oliveira, J. A. (2019). Simulation of an automotive supply chain using big data. *Computers & Industrial Engineering, 137:* 106033.

[15]. Kilimci, Z. H., Akyuz, A. O., Uysal, M., Akyokus, S., Uysal, M. O., Bulbul, B. A., & Ekmis, M. A. (2019). An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. *Hindawi Complexity.*

[16]. Amirkhalili, Y. S, Aghsami, A, & Jolai, F. (2020). Comparison of time series ARIMA and support vector regression. *International Journal of Hybrid Information Technology; 13(1):* 7-18.

[17]. Abbasimehr, H., Shabani, M., & Yousefi, M. (2020). An optimized model using LSTM network for demand forecasting. *Computers & Industrial Engineering,* 143: 106435.

[18]. Guo, L., Fang, W., Zhao, Q., & Wang, X. (2021). The hybrid Prophet-SVR approach for forecasting product time series demand with seasonality. *Computers & Industrial Engineering, 161*: 107598.

[19]. Sriram, G. K (2022), Edge computing vs. cloud computing: an overview of big data challenges and opportunities for large enterprise. *International Research Journal of Modernization in Engineering Technology and Science, 4(1):*1331 – 1337.

[20]. Hamiche, K., Abouaissa, H., Goncalves, G., & Hsu, T. (2018). A robust and easy approach for demand forecasting in supply chains. *IFAC- PapersOnline, 51(11):* 1732-1737.

[21]. Siddiqui, R., Azmat, M., Ahmed, S. & Kummar, S. (2022). A hybrid demand forecasting model for greater forecasting accuracy: the case of the pharmaceutical industry. *Supply Chain Forum: An International Journal, 23(2):*124-134.

[22]. Banatia, M. A., Baudry, D., & Louis, A. (2022). Detecting counterfeit products by means of frequent pattern mining. *Journal of Ambient Intelligence and Humanized Computing, 13(1):* 3683-3692

[23]. Rayani, A., Lathigara, A. & Mehta, A. (2021). Usage of time series forecasting model in supply chain sales prediction. *IOP Conference eries: Materials Science and Engineering,* 1042:012022.

[24]. Ramadhani, P.P., Hadi, S. & Rosadi. R. (2021). Implementation of data warehouse in making business intelligence dashboard development using PostreSQL database and Kimball lifecycle. *International Conference on Artificial Intelligence & Big Data Analytics (ICAIBDA);* 88- 92.

[25]. Singh, D. (2014). Big data analytic and mining with machine learning algorithm. *International Journal of Information and Computation Technology.* 4(1), 33-40.

[26]. Sharma, R., Mithas, S. & Kankanhalli (2014). Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organizations. *European Journal of Information systems,* 23, 433-441