Journal of Information Systems and Informatics
Vol. 6, No. 1, March 2024 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v6i1.665 Published By DRPM-UBD

Empowering Data Transformation: Transforming Raw Data into A Strategic Planning for E-Commerce Success

Angelina Yang[1], Jansen Wiratama[2], Santo Fernandi Wijaya[3]

[1,2,3] Information Systems, Universitas Multimedia Nusantara, Tangerang, Banten, Indonesia
Email: [1]angelina.yang@student.umn.ac.id, [2]jansen.wiratama@umn.ac.id,
[3]santo.fernandi@umn.ac.id

Abstract

The ability to transform data is essential to support strategic decision-making in a company or organization. Data transformation can be done by utilizing data warehouse technology. Therefore, it is necessary to know the description of a data warehouse that uses the Extract, Transform, and Load (ETL) process as its methodology. This research will focus on the context of Datawarehouse implementation at TechTrove, an e-commerce company that uses Pentaho Data Integration (PDI) tools. Star Schema organizes data marts and Online Analytical Processing (OLAP) to optimize data warehouse tasks. Business Intelligence (BI) tools are critical in extracting valuable insights and showcasing the platform's analytical capabilities in customer behaviour analysis, product evaluation, sales monitoring, and inventory management. This research transforms raw data into strategic plans to support decisions in e-commerce companies.

Keywords: E-commerce, Data Transformation, Decision-making, TechTrove, Datawarehouse

1. INTRODUCTION

One of the most important things to reach business success, specifically e-commerce, is the ability to use raw data, transform it, and make it decision-making. The development of a data warehouse is extracting vital information from the scattered data in some information systems into a centralized, integrated storage and supports the need for data history. This integrated data can be utilized for information delivery activities that can be reviewed from various dimensions and set the level of detail [1]. A company data warehouse uses the Extract, Transform, and Load (ETL) process. The first process is extracting the data from various sources. Secondly, it integrates, cleans, and transforms it into a standard form. Lastly, loads it into the data warehouse [2]. The result of the TL process generates data that meets the criteria of a data warehouse, such as historical data, integrated, encapsulated, static, and has a structure designed for business processes [3].

Two standard methods to organize data marts or data warehouses in relational databases are Star Schemas and Snowflake Schemas. These schemas utilize

This work is licensed under a Creative Commons Attribution 4.0 International License.

dimensional tables to describe the data contained in the fact table. The Snowflake Schema, a variation of the star schema model, involves normalizing some dimension tables to decompose the data into additional tables further. The resulting diagram is similar to a snowflake shape. Snowflake is an alternative version of the star schema, wherein the dimensional tables of star schemes are arranged hierarchically through normalization. Moreover, the Star Schema is center fact and change, while the Snowflake Schema is center fact and not change [4].

To facilitate efficient data processing by providing insight into two-dimensional knowledge, software known as OLAP (Online Analytical Processing) is used. OLAP enables interactive analysis of multidimensional data at various granularities [5]. The data warehouse supports online analytical processing (OLAP), the functional and performance requirements of which are quite different from those of the online transaction processing (OLTP) applications traditionally supported by operational databases [6]. OLTP (Online Transaction Processing) is a class of systems that manage and facilitate transaction-oriented applications, emphasizing customer-oriented, and is used for transaction and query processing by clerks, clients, and information technology professionals [7]. OLTP applications are traditionally supported by operational databases [8].

A data warehouse still needs to be a complete solution to the needs of business users. From that, business intelligence systems are required. Business intelligence supports access to all business information, not only the data stored in a data warehouse. A business intelligence system does not negate the need for a data warehouse – a data warehouse is simply one of the data sources a business intelligence system can handle [9][10]. Business intelligence systems combine operational data with analytical tools to present complex and competitive information to planners and decision-makers [11][12].

This research will present the star schema implementation, the following stages of the data warehouse development process, the ETL process, and many more in TechTrove. In this e-commerce, customers can find technology-related products, starting from innovative gadgets to cutting-edge electronics. Also, the store offers a unique and carefully curated selection of tech products. This research provides an innovative approach for Sales Trend Analysis, Customer Analysis, Product Performance Analysis, and Operational Efficiency Analysis using ETL on E-commerce.

2. METHODS

This research uses an approach with direct ETL implementation. The tools used in this research are Pentaho Data Integration (PDI) with a star schema model. A schema is a collection of database objects, including tables, views, indexes, and

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

synonyms [13]. A star schema is a multi-dimensional data model that uses the database to organize data, making it easy to analyze and understand. In this schema, the fact table is placed in the middle, and the relationship is created between each dimension table [14]. Each dimension table is joined to the fact table through a foreign critical relationship. This fact table allows users to query the data in the fact table using attributes from the dimension tables. Here are some purposes of using star schema. Firstly, it will make more straightforward queries. The join logic of the star schema is quite a cinch in comparison to other join logic, which is needed to fetch data from a transactional schema that is highly normalized. Secondly, it also simplified business reporting logic.

Compared to a transactional schema that is highly formalized, the star schema makes more straightforward common business writing logic, such as reporting and period-over-period. Thirdly, star schema is widely used by all OLAP (Online analytical processing) systems to design OLAP cubes efficiently. Major OLAP systems carry a ROLAP (Relational Online Analytical Processing) mode of operation, which can use a star schema as a source without designing a cube structure [15]. Here is the picture of TechTrove's star schema:

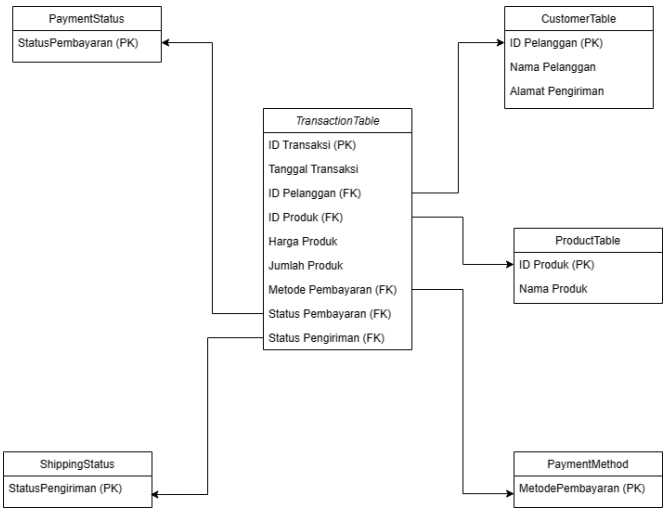

Figure 1. TechTrove Star Schema

In TechTrove star schema, there is a fact table named "Transaction Table" and it contains attributes about each transaction, such as the ID_Transaksi (PK), Tanggal_Transaksi, ID_Customer (FK), ID_Product (FK), Product_Price, Product_Quantity, Payment_Method (FK), Payment_Status (FK), and Shipping_Status (FK). PK means Primary Key, a unique identifier for a record in a database table. Meanwhile, FK means Foreign Key, a column or a set of columns in a database table that refers to the primary key of another table. Moreover,

around the fact table, there are four-dimension tables. Each dimension focus on a specific aspect of the business and is linked to the fact table through foreign keys.

1) **CustomerTable**
   This table contains information about customers, such as their *ID_Pelanggan, Nama_Pelanggan, and Alamat_Pengiriman.*
2) **ProductTable**
   This table stores information about products, such as their *ID_Produk, Nama_Produk, and Harga_Produk.*
3) **PaymentMethod**
   This table contains information about the payment methods in attribute named *Metode_Pembayaran.*
4) **ShippingStatus**
   This table stores information about the shipping statuses in a attribute named *Status_Pengiriman.*

The relationships between the tables are established through foreign keys. For example, the "Transaction Table" has a foreign key called "ID_Pelanggan (FK)" that references the "CustomerTable" by its primary key "ID Pelanggan (PK)." Each transaction record in the "TransactionTable" can be associated with a specific ID_Customer in the "CustomerTable." A star schema is used to optimize query performance for data warehouse tasks, such as online analytical processing (OLAP). By denormalizing the dimension tables and storing frequently used dimension attributes in the fact table, star schemas allow for faster joins and aggregations. The denormalizing is needed because all the data required for analysis is readily available in a single table, minimizing the need for complex joins across multiple tables.

3. RESULTS AND DISCUSSION

Extract, Transform, and Load (ETL) is used as a methodology in this research with Pentaho Data Integration Tools running on the Windows 10 platform. The steps taken are explained in stages according to Figures 3-8, and analysis has been carried out to measure several perspectives, such as sales trend analysis, customer analysis, product performance analysis, and operational efficiency analysis.

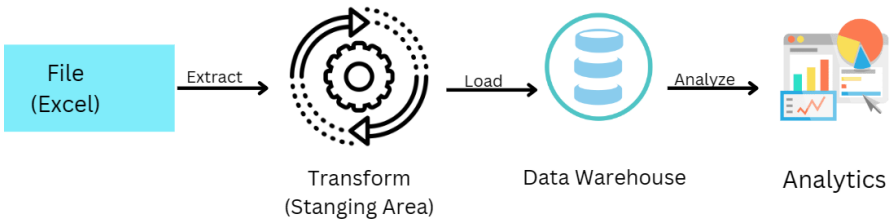

Figure 2. ETL Process on TechTrove

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Data extraction for TechTrove commences with the collection of Excel files containing essential information such as ID numbers and product names. These files are the primary raw data source for the business's analytics endeavors. Following the extraction phase, the data undergoes a meticulous transformation process. During this stage, the data from Excel files is cleaned, validated, and standardized into a consistent format suitable for analysis. Various transformation activities are undertaken, including data cleansing (eliminating duplicates, correcting errors), data enrichment (adding missing information), normalization (ensuring uniform units of measurement), and aggregation (summarizing data for analytical purposes). For TechTrove, this transformation involves tasks like product categorization, customer data management, and overall data accuracy validation. The output of this phase is data that is refined and ready for loading into the Data Warehouse.

The transformed data is in the Data Warehouse, functioning as a centralized repository for all structured and processed information. Loading can occur in batches or in real-time, adapting to the business's specific requirements. The Data Warehouse is optimized to facilitate swift querying and efficient storage of vast data volumes. Data is organized into tables and schemas, ensuring accessibility and usability for analytical purposes. This step culminates in storing data securely, structured, and optimized for comprehensive analytics.

Equipped with securely stored and well-organized data in the Data Warehouse, TechTrove employs various tools, including Business Intelligence (BI) tools and Data Analysis Software, to extract valuable insights. The analytics process encompasses complex querying, report generation, visualization of data trends, and the formulation of data-driven decisions. For TechTrove, this translates to analyzing customer behavior, evaluating product performance, monitoring sales trends, and managing inventory effectively.

The outcome of the analytics phase is a wealth of actionable insights that inform strategic decision-making, enabling TechTrove to enhance its online shopping experience, optimize inventory, and elevate overall business efficiency. Transformations that are made are lists each about who uses several payment methods in TechTrove. Filtering buyers who use 'GoPay,' 'OVO,' 'ShopeePay,' and 'BCA Mobile Banking.' Moreover, filter buyers who have already paid and have not paid the transactions. Here are several examples:

Journal of Information Systems and Informatics
Vol. 6, No. 1, March 2024
p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

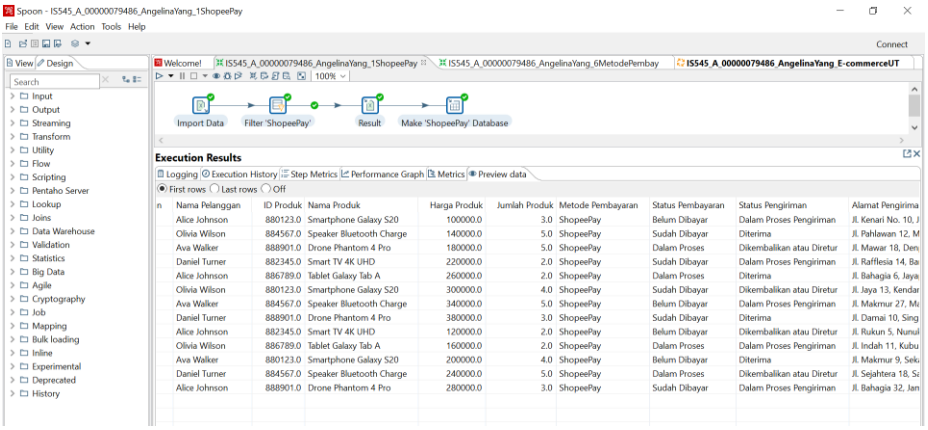

Figure 3. Execution results of TechTrove.

Firstly, use 'Microsoft Excel Input" to take all data from the TechTrove dataset. After that, use "Filter rows" to filter the "Metode Pembayaran" column with "ShopeePay" data. Then, use "Microsoft Excel Output" to display the filtering result from that recent Microsoft Excel. Lastly, "Table Output" to put the filtered data in the database "uts_dwh" like this example:

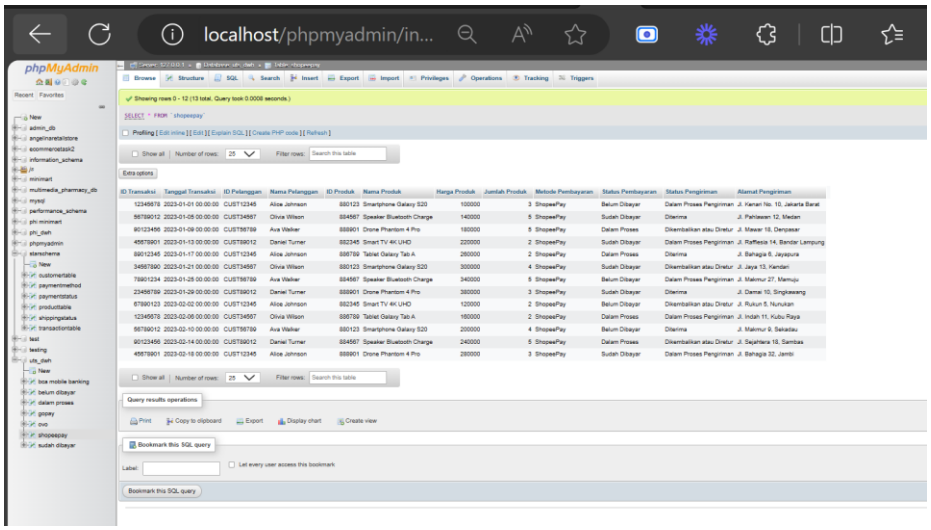

Figure 4. The screenshot of the data stored in the database.

Journal of Information Systems and Informatics
Vol. 6, No. 1, March 2024
p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

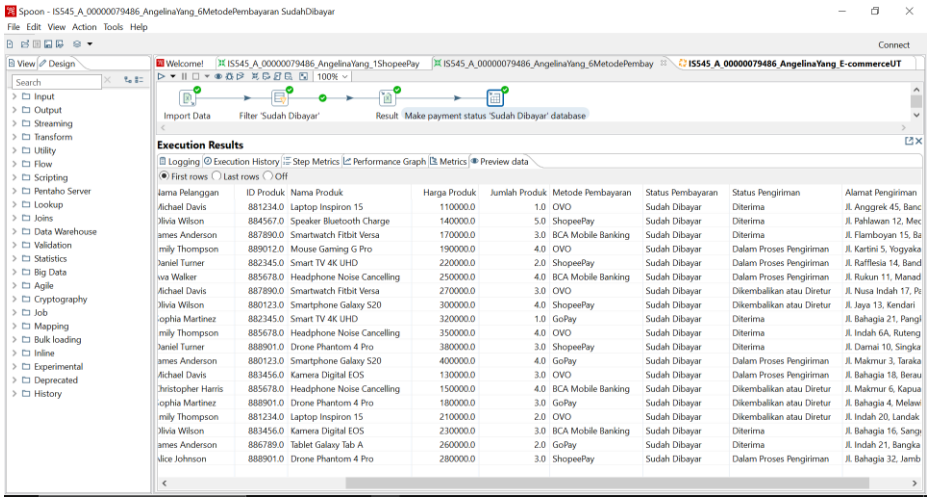

Figure 5. Transformation to filter "Status Pembayaran" which is "Sudah Dibayar"

Figure 5 shows 'Microsoft Excel Input" is used to take all data from the TechTrove dataset. After that, use "Filter rows" to filter the "Status Pembayaran" column with "Sudah Dibayar" data. Then, use "Microsoft Excel Output" to display the filtering result from that recent Microsoft" excel. Lastly, use "Table Output" to put the filtered data in the database "uts_dwh" like this example:

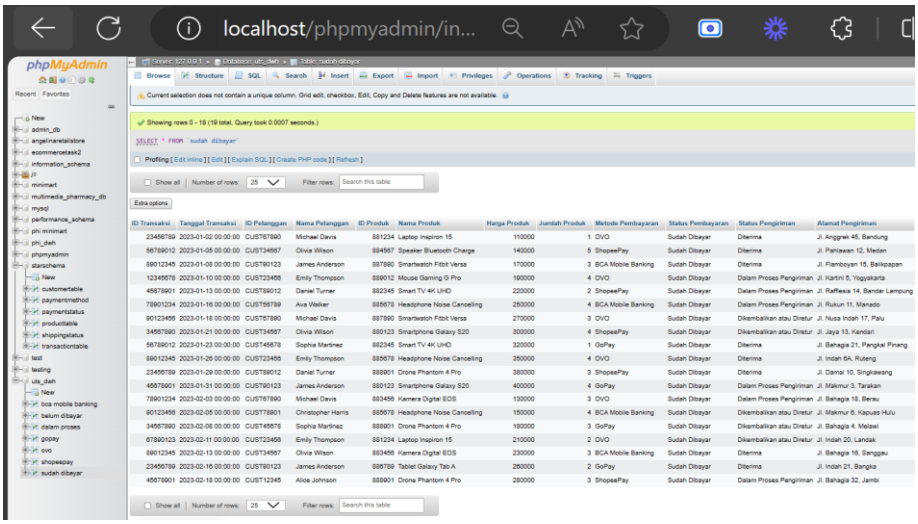

Figure 6. The screenshot of the data stored in the database.

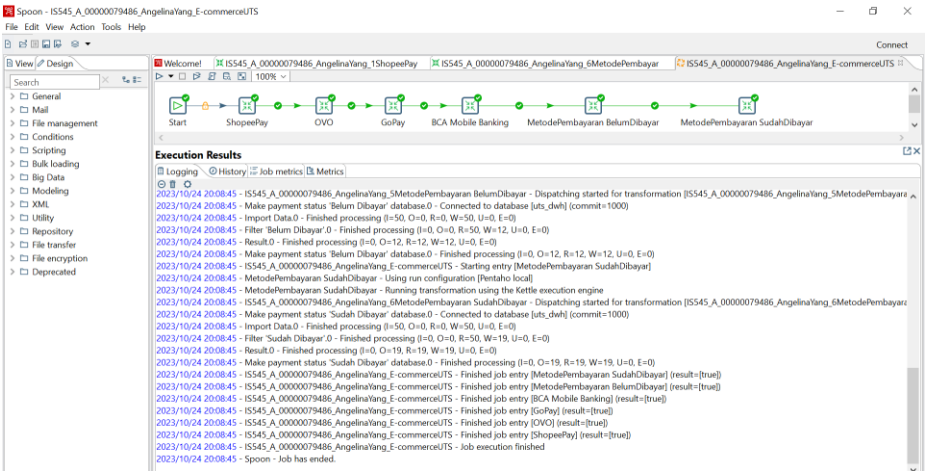

Figure 7. Job Execution results on PDI.

Overall, 6 transformations include the payment methods and payment status. Then, here is the Job. Started from "Start", until PaymentStatus "Sudah Dibayar". "Job" in Pentaho refers to a sequence of steps and transformations orchestrated to achieve specific data integration, ETL (Extract, Transform, Load). Jobs in Pentaho are high-level processes that define and automate data flow and activities within the Pentaho Data Integration (PDI) platform. Once the multidimensional database is populated with data, users can leverage SOA applications to query the cube and analyze the data from various perspectives. Here are some examples:

1) **Sales Trend Analysis,** Analyzing sales trends over time by grouping transactions by day, month, or year and comparing the number of transactions and total sales across different periods. This can help identify seasonal trends or the impact of marketing campaigns.
2) **Customer Analysis,** Segmenting customers by demographics like location or purchase history and comparing their average order value or preferred payment methods. This can help tailor marketing and promotional efforts to specific customer groups.
3) **Product Performance Analysis,** Analyzing the performance of different products by grouping transactions by product and comparing sales figures, profit margins, and return rates. This can help identify best-selling products, underperforming items, and potential areas for improvement.
4) **Operational Efficiency Analysis,** Tracking the efficiency of the shipping process by analyzing the distribution of transactions across different shipping statuses. This can help identify bottlenecks in the fulfillment process and areas for improvement.

4. CONCLUSION

In conclusion, this research underscores the indispensable role of data transformation in empowering strategic decision-making within organizations. Companies can effectively convert raw data into actionable insights by leveraging data warehouse technology and the Extract, Transform, and Load (ETL) process. Moreover, integrating Business Intelligence (BI) tools is crucial for extracting valuable insights across various domains, such as customer behavior analysis, product evaluation, sales monitoring, and inventory management. This research culminates in transforming raw data into a strategic plan tailored to support decision-making processes within E-Commerce. The successful implementation of this multidimensional database, facilitated by efficient ETL processes and data analysis tools, enhances strategic planning, fosters meaningful customer interactions, and boosts overall operational efficiency. Consequently, it positions TechTrove for sustained success in the competitive technology market.

ACKNOWLEDGEMENTS

We extend our heartfelt gratitude to Universitas Multimedia Nusantara for their invaluable support, pivotal in completing this research endeavor. Their substantial contribution was instrumental in achieving our objectives, and we are deeply grateful for their unwavering assistance.

REFERENCES

[1] L. W. Santoso, "Data warehouse with big data technology for higher education," *Procedia Computer Science*, vol. 124, pp. 93-99, 2017.

[2] H. Homayouni, "Testing extract-transform-load process in data warehouse systems," in *2018 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pp. 158-161, Oct. 2018.

[3] W. S. Fana, R. Permana, and M. A. Islam, "Data Warehouse Design With ETL Method (Extract, Transform, And Load) for Company Information Centre," *International Journal of Artificial Intelligence Research*, vol. 5, no. 2, pp. 132-137, 2021.

[4] K. Mohammed, "Data warehouse design and implementation based on star schema vs. snowflake schema," *Int. J. Acad. Res. Bus. Soc. Sci*, vol. 9, 2019.

[5] D. R. SHRIVASTAVA, R. Tiwari, K. Mehta, and S. Bano, "Various Olap Technologies and Their Impact on Decision Making," in *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*, April 2021.

[6] S. Chaudhuri and U. Dayal, "An overview of data warehousing and OLAP technology," *Sigmod Record*, vol. 26, no. 1, pp. 65-74, 1997.

[7] G. S. Reddy, R. Srinivasu, M. P. C. Rao, and S. R. Rikkula, "Data Warehousing, Data Mining, OLAP and OLTP Technologies are essential elements to support decision-making process in industries," *International Journal on Computer Science and Engineering*, vol. 2, no. 9, pp. 2865-2873, 2010.

[8] D. R. SHRIVASTAVA, R. Tiwari, K. Mehta, and S. Bano, "Various Olap Technologies and Their Impact on Decision Making," in *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*, April 2021. (Note: This is a duplicate of reference [5] and should typically be omitted or merged.)

[9] M. S. Almeida, M. Ishikawa, J. Reinschmidt, and T. Roeber, *Getting started with data warehouse and business intelligence*, IBM redbooks, 1999.

[10] J. Wiratama and M. A. Bagioyuwono, "Improving the Data Management: ETL Implementation on Data Warehouse at Indonesian Vehicle Insurance Industry," *International Journal of Science, Technology & Management*, vol. 4, no. 5, pp. 1256-1268, Sep. 2023.

[11] P. Hamm and M. Klesel, "AIS Electronic Library (AISeL)," 2021.

[12] S. F. Wijaya, J. Wiratama, and V. Kuswanto, "An Evaluation of Integrating ERP System to Develop a Strategy Business," in *2023 International Conference on Information Management and Technology (ICIMTech)*, Malang, Indonesia, pp. 1-6, 2023.

[13] E. Gallinucci, M. Golfarelli, and S. Rizzi, "Schema profiling of document-oriented databases," *Information Systems*, vol. 75, pp. 13-25, 2018.

[14] K. S. Harishkumar, "Multidimensional data model for air pollution data analysis," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1684-1689, Sep. 2018.

[15] E. Gallinucci, M. Golfarelli, and S. Rizzi, "Approximate OLAP of document-oriented databases: A variety-aware approach," *Information Systems*, vol. 85, pp. 114-130, 2019.