

Perfectionizing Clinical Data Management from Dark Data Identification Using ETL and Star Schema Architecture

Leandra Belva Varissa Perwira
Negara
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
leandra.negara@binus.ac.id

Wahyu A. Barmawi
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
wahyu.barmawi@binus.ac.id

Steven Halbert Widjaja
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
steven.widjaja@binus.ac.id

Nadzla Andrita Intan Ghayatrie
Bina Nusantara University
Jakarta, Indonesia
nadzla.ghayatrie@binus.ac.id

Abstract—In the ecosystem of hospital data analytics, the term of "Dark Data" such as inactive patient data, duplicates, and other irrelevant attributes often meet storage capacity and make processing becomes slow. The goal of this study is to identify and separate dark data from clinical datasets to produce a structured data ready for visualization. The methodology used includes an Extract, Transform, Load (ETL) process using Pentaho Data Integration, the cleaned data storing in MySQL, and final visualization using Power BI. By applying the Star Schema storage model, the raw data was successfully divided into efficient Fact Tables and Dimension Tables. So, the results show that removing irrelevant attributes and separating patient statuses significantly maximizes analysis efficiency and improves reporting dashboard performance.

Keywords—Dark Data, Data Engineering, ETL, Star Schema, Data Cleansing, Business Intelligence, Clinical Data

I. INTRODUCTION (HEADING 1)

The primary role of Data Engineering emphasizes data integration, resource optimization, and data governance. This process aims to connect raw data sources to a ready-to-use analytics ecosystem while ensuring that the infrastructure only stores high-value data. One of the biggest challenges in data governance is handling dark data which data is stored within a system without reused for analysis or decision-making. In the context of hospital systems, there are various kinds of dark data especially inactive patient data, duplicate data, irrelevant data, and empty or unused fields. The dark data in this context has significant negative impacts on systems including overloading storage, slowing processing, and reducing the efficiency of data analysis. Hospital datasets often suffer from issue such as mixing data between living and deceased patients as well as data structures become inefficient. The impact of this issue is slow dashboard loading and increased system resource usage. Therefore, the goal of this research was to identifying dark data, separating living and deceased patient data, cleansing

unnecessary fields, and providing a visualization as the final result.

II. METHODOLOGY

To solve clinical data problem in this case, we used several tools which are Pentaho Data Integration is used to the ETL process, Microsoft Excel is used to sample data output, and Power BI is used to visualization.

A. ETL and Data Ingestion Architecture

The primary approach is used was an ETL (Extract, Transform, Load) architecture because the transformation is performed in Pentaho to ensure that only clean and optimal data is stored. The initial stage begins with Data Ingestion, which involves importing a CSV file with a CSV File Input containing 1,000 rows of data. At this stage, data type formats are adjusted, such as converting the Date of Death column to String format.

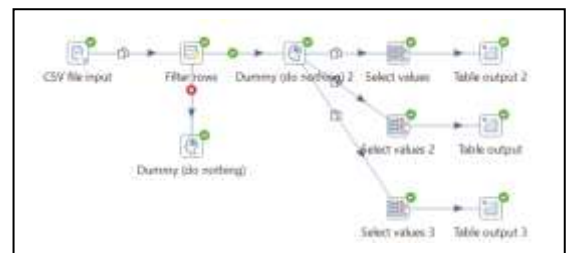


Fig. 1. ETL Architecture workflow and data lineage execution inside Pentaho Data Integration.

As shown in **Fig. 1**, the pipeline provides a clear separation of data stream paths right after ingestion. The transformation flow separates corrupted or uninformative rows in the beginning and prevents unnecessary downstream memory consumption during complex join operations.

B. Filtering dan Data Cleansing

The next step is filtering and cleansing. First step is Dark Data Separation where data is filtered using the Filter Rows component based on patient status (Alive or Dead). Patients with deceased status are not removed from the system, but are separated into dark data paths using a Dummy Step to reduce the burden on the main analysis. Second step is Column Cleaning using the Select Values step, fields that are not important for the main measurement (such as Date of Death, Dead or Alive, and Event) are removed.

C. Star Schema Planning

For storage efficiency and ease of analysis in Power BI, a single dataset is split into two separate tables using the Star Schema approach that is Fact Table and Dimension Table. Fact Table is the central clinical observation metric containing patient IDs, numeric measures such as Time (survival duration), and Events. This table is designed to be very small in size for fast processing by the database. Dimension Table contains static, context-providing attributes that rarely change (such as Sex, Race, and Stage). This method prevents repetition of long text data when a patient has multiple medical records.



Fig. 2. Clinical Star Schema architecture

According picture from **Fig. 2**, the architecture prevents redundant data anomalies. The specific table attributes and storage structures are resulted from this distribution which detailed in Table I.

TABLE I. SCHEMA DEFINITION FOR FACT AND DIMENSION TABLES

Table Name	Column name	Data type	Key type
fact_clinical	PatientID	VARCHAR(50)	FK
	Specimen_Date	DATE	
	Date_Last_Follow Up	DATE	
	Time	INT	
dim_patients	PatientID	VARCHAR(50)	PK
	Sex	VARCHAR(10)	
	Race	VARCHAR(30)	
	Stage	VARCHAR(15)	

The Fact Table (fact_clinical) is the primary clinical observation metric that contains numeric measures such as Time (survival duration) and transactional foreign references. This table is attempted to be make very small table for fast

computational processing by the database engine. The Dimension Table (dim_patients) contains rarely change attributes because the nature of attributes is static and descriptive such as Gender, Race, and Stage. This relational normalization prevents repeated placement of long text when a single patient triggers multiple clinical entries over time.

D. Automation (Job Streaming)

Job Step	Comment	Result	Reason	Status	Log Date
Job Step	Start of job execution	start		2020/05/12	
Start	Start of job execution	start		2020/05/12	
Start	Job execution finished	Success		2020/05/12	
Transformation	Start of job execution	Followed execution...	C:\Users\... \Documents...	2020/05/12	
Transformation	Job execution finished	Success	C:\Users\... \Documents...	2020/05/12	
Start	Start of job execution	start		2020/05/12	
Start	Job execution finished	Success		2020/05/12	
Transformation	Start of job execution	Followed execution...	C:\Users\... \Documents...	2020/05/12	
Transformation	Job execution finished	Success	C:\Users\... \Documents...	2020/05/12	
Start	Start of job execution	start		2020/05/12	

Fig. 3. Continuous automation streaming loop is scheduled in Pentaho Job Integration.

The scheduling routine which is provided in **Fig. 3** applies a structured infinite loop orchestration which has interval of 30 seconds with repeatable. This automated pipeline ensures that incoming Dark Data cleansing and analytical extraction occur continuously. Therefore, the advantage of this automated pipeline is manual database administration intervention is prevented and operational processing lag is mitigated.

III. RESULTS AND DISCUSSION

A. Business Intelligence Storage Ecosystem

MySQL is used as a permanent storage layer to provide data which data will done through ETL process. The implementation of Star Schema in MySQL enables the system to connect a connection through DirectQuery which it is integrated with Power BI and provide high availability capabilities which makes dashboards can load in near-real time. The final dataset contains only active patients with the following columns: PatientID, Specimen date, Date of Last Follow-Up, sex, race, Stage, and Time.

B. Data Quality

Throughout the entire process, data quality is maintained through three aspects: data type integrity through 'Select Values' metadata, the use of Audit Trail through 'Dummy Step' to verify discarded data, and encoding standardization (UTF-8) to prevent data corruption when accessed externally.

C. Power BI Visualization Dashboard

Visualizations are designed exclusively to display active patient analysis:

- Key Metric Indicators: Card Visualization directly displays the total number of surviving patients who passed the screening stage.
- Demographic Distribution: Pie Chart is used to compare the number of patients who survived by sex. Besides that,

horizontal bar chart is used to compare the number of patients which type of specific race is more according to sex between male and female.

- **Specific Composition:** Bar chart is used to know number of patients which of each gender from male or female has the highest number from stage.
- **Temporal Analysis:** Bar chart is used to compare which of both sex have longer the duration of life.
- **Interactivity:** Interactive Slicer component is added to enable users to filter specific data based on special conditions (Stage, Race, Sex).



Fig. 4. Dashboard patient which condition is alive

IV. CONCLUSION

Based on the implementation, Pentaho Data Integration successfully fulfilled its function of cleansing and optimizing clinical data by separating active patient data from dark data. The Filter Rows step successfully made the dataset abstraction more focused on the observation object (live), while Select Values scrubbed the table to be lighter and more efficient. Finally, visualization based on the dataset with Power BI successfully facilitated the need for a more organized monitoring dashboard. The use of ETL and Star Schema design

methodology has been proven to optimize data processing, boost system scalability performance, and support comprehensive analytical needs for hospitals.

ACKNOWLEDGMENT

The authors express their gratitude to God Almighty for the smooth completion of this research. They express their deepest appreciation and gratitude to Bina Nusantara University, especially the School of Computer Science, for providing a supportive academic environment. The authors also thank the lecturers in charge of the Data Engineering course for their valuable guidance, direction, and advice throughout the development of this project. Finally, the authors extend their deepest appreciation to all group members who have demonstrated dedication, hard work, and solid collaboration in completing all stages of ETL, data modeling, and final visualization..

REFERENCES

- [1] R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3rd ed. New York, NY: John Wiley & Sons, 2013.
- [2] J. Heidrich, M. Bauer, and P. Kroll, "Handling Dark Data in Medical Information Systems: Challenges and Opportunities," *IEEE Transactions on Big Data*, vol. 5, no. 2, pp. 142–155, June 2019.
- [3] A. Bouman and C. S. Hoogendoorn, "Optimizing Open-Source ETL Processes for Healthcare Analytics Using Pentaho Data Integration," *International Journal of Medical Informatics*, vol. 124, pp. 45–53, April 2021.
- [4] T. White, *Data Engineering and Architecture for Enterprise Systems*. San Francisco, CA: Morgan Kaufmann, 2020.
- [5] S. R. Lwakatare and J. Sauvola, "Continuous Data Ingestion and Cleansing Pipelines in Business Intelligence Ecosystems," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2022, pp. 889–894.
- [6] M. Ahmad, H. Malik, and F. Khan, "Evaluating DirectQuery and In-Memory Performance in Power BI Visualizations for Large Datasets," *IEEE Access*, vol. 10, pp. 31200–31212, 2022.
- [7] E. Awad and M. Al-Ghamdi, "A Framework for Eliminating Data Redundancy in Clinical Fact Tables using Star Schema Design," *Journal of Healthcare Engineering*, vol. 2023, pp. 1–11, March 2023.