Design and Implementation of an End-to-End Data Engineering Pipeline for Heart Disease Prediction Using Pentaho Data Integration

[1st] Brian Filbert
Department of Computer Science
Bina Nusantara University
Jakarta, Indonesia
brian.filbert@binus.ac.id

[2nd] Farrel Gionovan Surbakti
Department of Computer Science
Bina Nusantara University
Jakarta, Indonesia
farrel.surbakti001@binus.ac.id

[3rd] Viky Stevenlay
Department of Computer Science
Bina Nusantara University
Jakarta, Indonesia
viky.stevenlay@binus.ac.id

[4th] Nadzla Andrita Intan Ghayatrie
Department of Computer Science
Bina Nusantara University
Jakarta, Indonesia
nadzla.ghayatrie@binus.ac.id

***Abstract*— The health datasets are generally messy, consisting of inconsistent and poorly labeled information. The problem is solved by cleaning the information beforehand. Due to inconsistency and unreliability of the data collected from health records, the appropriate data management approach is needed to handle the situation. A proposed solution includes a fully fledged process designed for the Personal Key Indicators of Heart Disease (BRFSS 2022) dataset based on Pentaho Data Integration. Starting from loading the files and ending with automating the processes, all the phases are linked together and consist of the following steps: input, cleansing, transforming, structuring according to the star schema pattern. Once executed, this chain produces a well-structured output: the data is reorganized into a star schema and structured dimensions become ready for analysis. With proper structuring, the dataset becomes well-prepared for machine learning models. The resulting dataset is structured and consistent, making it suitable as input for forecasting and decision making systems. The results show that the suggested approach provides more organized and consistent data. The proposed approach significantly simplifies all the processes, especially preprocessing of health records for further analysis.**

Keywords—Data Engineering, ETL, Pentaho, Machine Learning, Heart Disease, Data Pipeline, Star Schema

I. INTRODUCTION

Though heart disease still ranks high among global causes of death, it demands ongoing attention in public health efforts. Age plays a role, yet so do choices like smoking, being inactive, or carrying extra weight - each adding strain over time. Diabetes stands alongside them, quietly increasing danger without clear warning signs. As medical records grow faster than ever, institutions turn to algorithms that learn patterns, aiming to spot risks earlier. These systems work better when fed trustworthy details, since shaky inputs lead directly to weak conclusions [1], [3], [4].

Messy health data usually comes with gaps, repeated entries, mixed-up categories, wrong labels here and there, plus files in different shapes. Because of this clutter, it cannot be used straight away for analysis. Earlier work shows cleaning up data boosts its reliability while lifting results in medical predictions. That's why shaping raw inputs needs careful handling - turning chaos into order helps machines learn well. The conversion of raw data into usable data involves a process that includes gathering sources, formatting, validation, filing, and finally preparation for further analysis. In today's systems, this is done through automated machines which reduce manual effort, improve consistency, and increase reliability [5]. For hospitals and medical institutions, there has been an increasing need for structured ways of collecting, formatting, and storing data due to the analysis involved.

The data source for this project is the Personal Key Indicators of Heart Disease (BRFSS 2022) dataset from Kaggle. The data is imported and preprocessed using Pentaho Data Integration from CSV files completely. The first step involves importing the initial data and then transforming it by shaping different categories. The cleaning process is done at the early stage, while duplicate attribute combinations are removed only when building the dimension tables rather than dropping respondent records. This means that not only is the data organized but also a star schema model is created for further analysis. All processes are connected through the use of Pentaho Jobs. As a result, the final product becomes a reliable dataset which can be used to train predictive models. This fact was confirmed in earlier studies, which show that cleaned inputs increase the accuracy of the health forecast system [8].

In the course of this project, all the stages of creating a data engineering pipeline, from acquiring uncleaned information to creating clean datasets for machine learning models, will be covered. Moreover, there is going to be an attempt to evaluate the efficiency of using traditional extraction, transformation, and loading methods in health-related analysis.

II. LITERATURE REVIEW

Reflection on key concepts and previous knowledge is instrumental in shaping the project. Data engineering serves as the starting point of the process, with information being transferred and processed through extraction, transformation, and loading processes. Data cleansing becomes equally important as data structuring while striving for the accuracy

of the findings. Star schema is applied to organize storage space intelligently. The next step involves machine learning, as prediction of heart disease is set as the goal using real data inputs.

A. Data Engineering

Behind any intelligent decision lies an invisible labor – data engineering determines how data travels from sources into meaningful insights. This process not only involves transporting data but creating pipelines that ensure proper conversion of data into useful form through collection, cleaning, and organization. Every step of the process – input, transformation, validation, storage, and scheduled execution – is well-coordinated within such systems. If executed correctly, the output will be clean, consistent, and easy to work with for reporting and modeling. Mbata et al. [5] note that these solutions scale up, iterate often, adjust quickly, and support anything from dashboards to machine learning. Fact-based organizations depend a lot on these processes as the basis for their decision-making.

B. Extract, Transform, and Load (ETL)

Extraction is the first step that is done in the process known as ETL. The extracted data is then processed by cleaning, validation, and integration for later understanding. After processing, the new data is transferred to another system where the users analyze the data. Vassiliadis [11] explains that the use of such a process will ensure that the companies maintain consistency and reliability of their data. In work by Kathiravelu et al. [9], good ETL pipelines allow for easier replication of results and better analytics processes. For this research, the file transfer, transformation, validation, and scheduling tools in Pentaho Data Integration have been used.

C. Data Preprocessing and Data Quality

Data cleansing is an initial step in development of information-based systems as all real-life data is characterized by incompleteness, duplications, misclassification and inconsistency. If ignored, such data errors distort analyses and impair predictions produced by algorithms. Data cleansing including filling in missing values, feature scaling, elimination of duplications makes medical data reliable – Benhar et al. [1] proved that it improves detection of heart diseases. For example, Tseng et al. [3] revealed that better preparation improves results obtained with intelligent models trained on cleansed data.

Quality of data depends on its reliability, consistency, completeness and accuracy throughout usage. Such cleaning techniques as removal of duplications and verification of values increase the quality of data – that is what Rahm and Do [10] noted. Duplicates might influence research findings in negative way – Elmagarmid et al. [12] stated this point very clearly. Various checks and preparations are performed for creation of reliable and organized data sets.

D. Data Warehouse and Star Schema

Being collected in one single place, the data warehouse provides an opportunity to combine structured data from various sources to make analysis and decision making easy. The thing that is commonly found in such systems is the star schema – a design where the numeric facts table is related to several tables providing information on context. It is possible to simplify querying because of the separation between facts and their description which is enabled by this structure. Optimization occurs thanks to improved structure making data easily accessible. Star schema can be useful for hospital/clinic scaling and storing data consistently [6].

E. Machine Learning for Health Disease Prediction

Most times, machines learn better when health info is clear. Still, messy details like gaps in records or repeated entries trip up the process. Instead of flowing right into algorithms, patient facts need shaping first. Without clean formatting, predictions can go off track easily. Mixed-up types - numbers next to notes - add more trouble than expected.

Most messy health information needs reshaping before machines can learn from it. Fixing errors comes first, then filling gaps where details are missing. Duplicates get tossed out, while categories shift into clear patterns. Structure emerges through careful reorganization. Smoother analysis follows when inputs behave better. Ready data works well later on.

One way to start: clean data fits machines right away. Built correctly, it feeds straight into tools that sort or forecast outcomes. For health insights to hold up, what comes before matters just as much. Shaping information wisely makes later steps stronger, even if unseen. Getting this part steady helps everything after stay firm.

III. METHODOLOGY

Here's how the data engineering process was built to get a heart disease dataset ready for machine learning. Starting with raw inputs, each step moves through stages like gathering, checking, reshaping, structuring, then automating flows. Instead of manual handling, PDI takes charge of moving and shaping information. Through this method, outputs become uniform, well-organized, fit for analysis. Each phase links into the next, making sure quality stays high without breaks in flow. Tools are chosen not for popularity but for steady performance across tasks. Results come out clean, predictable, prepared ahead of time. Automation runs behind the scenes so repetition does not slow progress down.

A. Dataset

The BRFSS 2022 Personal Key Indicators of Heart Disease (Kaggle dataset) forms the basis for this work. This dataset contains about 445,000 respondents' records and 40 columns. Among other information, it features data on individuals' background, lifestyles, and symptoms related to heart disease. In addition, the values contained are categorical or numerical; anything in between does not exist. Further models will be trained to predict heart attacks among patients based on the feature labeled HadHeartAttack.

Selecting the particular dataset was justified – the data is real and inconsistent in nature. Even though the processing is hard, the messy dataset may offer valuable patterns which can be used further. Similarly to the majority of healthcare datasets, this particular one offers unregulated categories and,

hence, requires restructuring. The process of cleaning included all the steps possible for all data; nothing was omitted. All the values were utilized, regardless of being obscure at first sight. The objective was to prepare the dataset in such a way that would allow models to make use of hidden patterns in messy and inconsistent data [1], [4].

B. Overall Data Engineering Pipeline

Figure 1 demonstrates the manner in which the pipeline was designed for this study. The process begins with the receipt of the initial stream of health records and subsequent steps that aim to rectify and validate data. The process begins with cleaning followed by validating whether the numbers are logical. Cleaning will enable data transformation to take place to give standard forms to ensure it can be merged with other related data. Data merging will facilitate the connection of data pieces in order to integrate them. There will be formation of a star schema to organize the facts according to topics using automation through Pentaho Jobs.

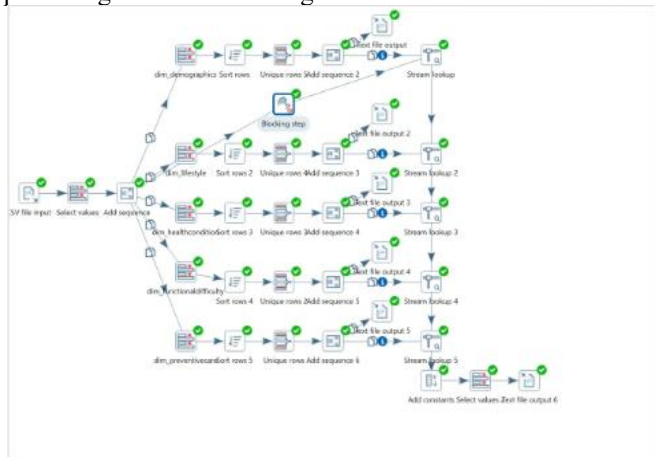

Fig. 1. Pentaho transformation: ingestion, dimension building, and fact assembly.

C. Data Modeling

Consider Figure 2 - this is an example of the star schema used in sorting the cleaned data related to healthcare. This is built around the central fact table and is connected to individual dimensions including age, behavior, health, visits, and everyday activities. Everything here is neatly interconnected, which facilitates pattern recognition. The data is better organized and can be easily accessed whenever required. In the process of model development in the future, all of it will be nicely aligned.

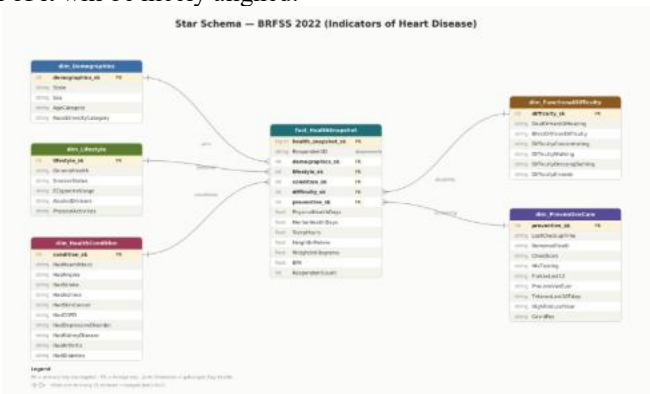

Fig. 2. Star schema data model (one fact table and five dimensions).

D. Workflow Automation

Referring to Figure 3 below, one can understand how workflow is handled by the Pentaho Jobs. This is because of the fact that one task is carried out immediately after another in a very sequential manner. As a result, mistakes are avoided, whereas repeatability increases.

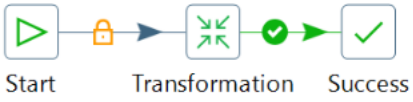

Fig. 3. Pentaho Job orchestration (Start → Transformation → Success).

E. Data Storage

Both raw and processed data are stored in an organized way. Raw source file (heart_2022_with_nans.csv) is stored as it was received in its original form. Five files are generated as processed data: demographics_dim.csv, lifestyle_dim.csv, condition_dim.csv, difficulty_dim.csv, and preventive_dim.csv. These files correspond to dimension tables, and fact_table.csv corresponds to the fact table. Splitting raw data from processed data allows us to rerun the entire pipeline using the original data at any moment in time, while keeping the ready-to-use output separated.

IV. Results and Discussion

Here is what occurred when the data system went live. It is based on Pentaho, and it gathers the data, prepares it, refines it, creates the dimensions, and processes everything automatically. Not separate activities, but an integrated process. Before any changes were introduced, there was no standardization of health records - after, the patterns became very clear. Structure improved as each stage fixed particular issues. Duplicates of attributes combinations were eliminated during dimension creation; modeling helped establish relationships between facts. The warehouse structure is good for analysis while not distorting any patient data. There was referential integrity, as each fact row corresponded to all five dimensions. After comparing both sets of information, it became clear that the post-processed data was consistent with each other in terms of the subject matter. The latter refers to the data set which was intended for future machine learning since it is consistent and complete. Medical information remained accurate and consistent in all phases of the process. Finally, chaos gave way to order, but not by changing the essence, but just by improving the presentation. Information went from scattered sources into consolidated storage through established channels. All the changes were made with practical goals in mind.

TABLE I. Comparison of Dataset Characteristics Before and After

Aspect	Before ETL	After ETL
Dataset Structure	Raw dataset	Structured dataset

Surrogate Keys	None	Generated per dimension
Data Validation	Not validated	Validated through ETL workflow
Data Organization	Flat dataset	Organized using a star schema
Machine Learning Readiness	Not suitable	Machine learning-ready

From a glance at Table I, one can see how good the situation became after adopting the new ETL process. The data was cleaned and transformed into a different structure and validated and eventually dimensioned. This process made it possible to change dirty health data into clean and well-organized rows.

A. Data Quality Improvement

Some of the basic data quality controls that have been conducted through the pipeline include enforcing consistency using star schema by changing the one flat table of 40 columns into one fact table and five dimensions including demographics, lifestyle, health condition, functional difficulty, and preventive care. The surrogate keys have been used for each possible combination of dimension attribute through the Add sequence step to keep the referential integrity of the relationship between the fact table and dimensions without duplication of attributes in the fact table.

The second one was the validation of the referential integrity using the simple control. This was done after the Stream lookup process where the fact rows had to match all five dimensions. Therefore, the five foreign keys (demographics_sk, lifestyle_sk, condition_sk, difficulty_sk, and preventive_sk) had to be populated and not nulls. This has been confirmed through the preview of the final Stream lookup step.

Thirdly, missing values were not imputed but kept intact. The decision to either impute, remove, or treat missing values as a unique class depends on the selected modeling approach and is supposed to be done only using the training data set so that there is no data leakage. This is the main reason why this step will be delayed until the data science/machine learning stage. In this way, the integrity of the entire signal contained in the dataset is maintained.

B. Data Warehouse Implementation

From chaos to order – a star schema was used to organize the healthcare data properly. In the middle there is a central table with numeric information about patients' health events. From it other tables branch out with such information as age, lifestyle, disease diagnoses, examinations and problems associated with daily activities. Everything is structured and therefore does not contain duplicate information if not required.

Normalization of data into the star schema format allows for processing it quickly while creating reports and performing analysis. It separates data from descriptions and stores them in a separate lookup table. Because the descriptions are located in an additional table, which means that the search of data becomes easier than using the single table. The user finds all he needs fast as the data is divided into logical categories of time, place or individual patients. Moreover, new requirements, like including such information as laboratory tests or medical notes can be easily met without changing anything in the overall structure because it allows introducing additional fields due to proper structuring at the very beginning. Methods of analyzing trends and forecasting will be easily accommodated due to this.

C. Machine Learning-Ready Dataset

What emerges after the proposed ETL flow is an impeccably validated data set that can be used for machine learning purposes. After the combination of intake, cleansing, transformation, and organization of dimensions, the data emerges fit for analysis without the loss of essential information from the initial records of the health state. Finally, all this puts into evidence the ability of the design to convert the raw health data into appropriate input for the forecasting algorithms.

The intended machine learning task would be predicting whether the respondent had a heart attack on the basis of the HadHeartAttack variable. When it comes to modeling, the star schema can be denormalized into one table, which includes one row for each respondent. No encoding and scaling operations are performed; this way they can be applied in the stage of machine learning on the training set only to avoid data leakage problem.

This task can be helped by the data in a number of ways. The nature of the target value being just two makes this task a binary classification one, and structured data are very suitable for this kind of problem. The dimensions created from the modeling process have the ability to cluster some known heart disease risk factors into groups: demographic factors (age and sex), behavioral factors (smoking, drinking, physical activities), health factors (diabetes, stroke, etc.), and numerical factors like body mass index and sleep hours. This way, the factors that will predict heart disease are organized in an orderly manner for ease of selection for modeling. One issue that might occur from this target value is that it is imbalanced because those who have suffered heart attack are very few in number.

D. Limitations

There are some restrictions that one should consider. Firstly, because the outputs are written to CSV files, not the database, the surrogate keys are generated using the Add sequence step and, thus, are not persistent and idempotent for different loads. Also, the pipeline only does full loads, no change data capture or incremental loads are implemented. As far as data pre-processing is concerned, it must be noted that there is no encoding and no feature scaling done at the engineering level as these operations are moved to the machine learning stage.

V. CONCLUSION

From beginning to end, one system created an entire path for healthcare data using Pentaho. It not only extracted raw data files, but also cleansed each one along the way. Following extraction is the process of shaping – transforming messy inputs to well-defined formats that can be easily recognized for patterns. Unlike the original format, which contained both details and numbers, the design separated

these two elements. Since facts are now independent from each other, queries will operate much faster on the records. Each process is connected smoothly, so updating requires no additional steps at all. What was once fragmented information becomes accurate, properly structured and ready to be modeled. Due to the added structure, past statistics become clear simply because of the order in which they are presented. There are no unnecessary layers, as the design separates components while connecting them.

The key point is that the proposed approach demonstrates superior results in processing health-related data prior to its further use. The proposed approach does not eliminate any steps and starts with the cleaning of data and organizing the data into structures specifically designed for clarity. Thanks to this approach, the report generation and forecast modeling processes will have improved inputs without additional costs. One process leads to another – automation helps minimize the number of repetitions while maintaining traceability of each step. In the future, there is potential to include different types of medical records, update data incrementally, and test results on various prediction methods for heart disease and alike applications.

ACKNOWLEDGMENT

The authors would like to thank the lecturers of the Data Engineering course at Bina Nusantara University for their continuous guidance, support, and constructive feedback throughout the completion of this project. Appreciation is also extended to Kaggle for providing the publicly available dataset and to the Pentaho community for the development of Pentaho Data Integration, which facilitated the implementation of the proposed ETL workflow.

REFERENCES

[1] H. Benhar, A. Idri, J. L. Fernández-Alemán, and I. Toval, "Data preprocessing for heart disease classification: A systematic literature review," Computer Methods and Programs in Biomedicine, vol. 195, 2020.

[2] N. S. Sulaiman and J. H. Yahaya, "Development of Dashboard Visualization for Cardiovascular Disease Based on Star Scheme," Procedia Technology, vol. 11, pp. 1183-1191, 2013.

[3] C. W. Tseng, L. C. Sun, K. F. Lin, and P. N. Chen, "Enhancing coronary heart disease diagnosis: Comparative analysis of data preprocessing techniques and machine learning models using clinical medical records," Health Informatics Journal, vol. 31, no. 2, 2025.

[4] M. M. Ahsan and Z. Siddique, "Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review," 2021.

[5] A. Mbata, Y. Sripada, and M. Zhong, "A Survey of Pipeline Tools for Data Engineering," 2024.

[6] A. L. Choi, J. H. Lee, J. W. Choi, S. H. Kim, and H. S. Kim, "Building a Healthcare Data Warehouse: Considerations, Opportunities, and Challenges," Journal of Clinical Medicine, vol. 14, no. 2, 2025.

[7] S. M. Winnenburg, P. M. Casano, M. K. Schubart, et al., "Comparative Study of ETL Tools for Transforming Healthcare Data to the OMOP Common Data Model," Studies in Health Technology and Informatics, vol. 327, pp. 474-478, 2025.

[8] R. K. Behera, S. Das, S. R. Sahoo, and B. Majhi, "Data-driven Preprocessing Techniques for Early Diagnosis of Diabetes, Heart and Liver Diseases," in Proc. IEEE International Conference on Computer Communication and Informatics (ICCCI), 2022.

[9] P. Kathiravelu, S. Suhothayan, V. Muras, and C. Zhang, "On-Demand Big Data Integration: A Hybrid ETL Approach for Reproducible Scientific Research," 2018.

[10] E. Rahm and H. H. Do, "Data Cleaning: Problems and Current Approaches," IEEE Data Engineering Bulletin, vol. 23, no. 4, pp. 3-13, 2000.

[11] P. Vassiliadis, "A Survey of Extract-Transform-Load Technology," International Journal of Data Warehousing and Mining, vol. 5, no. 3, pp. 1-27, Jul. 2009.

[12] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate Record Detection: A Survey," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, pp. 1-16, Jan. 2007.