COVID-19 Medical Data Integration Approach

VIOLETA TODOROVA[1], VESKA GANCHEVA[2], VALERI MLADENOV[1]
[1]Department of Fundamentals of Electrical Engineering
[2]Department of Programming and Computer Technologies
Technical University of Sofia
Kliment Ohridski boul. 8, Sofia
BULGARIA

Abstract: - The need to create automated methods for extracting knowledge from data arises from the accumulation of a large amount of data. This paper presents a conceptual model for integrating and processing medical data in three layers, comprising a total of six phases: a model for integrating, filtering, sorting and aggregating Covid-19 data. A medical data integration workflow was designed, including steps of data integration, filtering and sorting. The workflow for Covid-19 medical data from clinical records of 20400 potential patients was employed.

Key-Words: - Clinical Records, COVID-19, Data Analytics, Data Integration

Received: May 15, 2022. Revised: May 28, 2022. Accepted: June 19, 2022. Published: Juluy 18, 2022.

1 Introduction

With the development of health care, science and high technology, the amount of generated information is growing at a tremendous speed and volume [1]. As a result, multiple heterogeneous data emergece, different in terms of types, storage files, sources of data generation. The process of managing different data from different sources is called data integration. This is a typical process in fields such as medicine, biology, bioinformatics, etc.

Big data in medicine includes biological, biometric and electronic health data records [2]. Medical databases have a high degree of differences in terminologies, features of records, data presentation [3]. This, in turn, is associated with problems when querying multiple databases. Therefore, there is a need to automate database integration to do much more than simple data extraction and modification [4, 5]. Records in different medical databases have different formats. Integration requires the use of formats across databases, but high dimensionality and redundancies make such integration impossible.

During the data integration process, filtering operations are performed to remove duplicate data, data conversion, or manage data. The data integration model can also vary between extract, transform and load (ETL), extract, load and transform (ELT), data transformation, data replication, data virtualization, streaming data integration [6].

This paper presents a conceptual model for integrating and processing medical data in three layers, including a total of six phases: a model for integrating, filtering, sorting and aggregating Covid-19 data implemented in Talend Open Studio [7].

2 Material and Methods

The structure of the proposed medical data integration and processing model is illustrated in Fig. 1. The model is organized into three layers, each of which brings together the tasks to be performed.

Data management consists of three main phases: data preparation for analysis, interpretation and visualization, and the preparation phase includes medical data collection, medical data storage, medical data integration. The "Medical Data Collection" phase is based on the data sources, the technical devices providing visual data, the specifics of the generated data, including data types and data formats, images and features. One of the main sources of medical data includes patient data obtained from patient examinations, symptoms, personal data including age, gender, medical history, etc. Also, sensor data, omics data, electronic health data and health records are collected.

The second phase "Medical data storage" of the proposed model is related to the data storage process. Typically, clinical data is collected and stored in various file formats such as ".xls", ".xlsx", ".csv", ".xlsm", DICOM, etc. However, there are two main

problems with the data. First, the large amount of data of different sizes, data types, and file formats requires storage space and data processing tools. Second, the variety of technical characteristics of the large amount of data leads to heterogeneity.

In the third phase "Medical Data Integration" the process of merging data collected from different sources is carried out. That includes cleaning, ETL/ELT, mapping, transformation steps. In the data integration process, data cleaning, unifying formats, grouping and classification, regression and filtering, copying, downloading, loading data warehouses, data extraction, merging, aggregating, object and server data management and so on. Storing the data in a data warehouse allows users to quickly access data stored from different sources in one place to ensure a short time to retrieve information and store a large amount of data from previous periods. The latter allows users to perform analysis over a given period and make predictions about further trends and events.

The phase "Medical Data Processing" of the second layer "Data Analysis" of the presented model includes the process and methods applied to process the medical data. The data processing process involves manipulating the collected data and performing functions and operations in order to extract meaningful information. Features included include validation, sorting, aggregation, analysis, reporting, classification. Sorting is used to arrange the data according to all submitted requirements. Aggregation is the process of combining multiple pieces of data. Analysis is applied to transform and model the data. Classification performs the separation of data into groups according to requirements.

The Medical Data Classification phase involves the process of arranging data into groups based on predefined criteria. Clustering methods and techniques such as k-Nearest Neighbor (kNN), k-Means, Support Vector machine (SVM), Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Naive Bayes, etc. are applied for data purposes.

The "Decision Making" phase is the last phase and is structured in the "Problem Solving" layer, and it is built upon the phases carried out earlier using different methods.

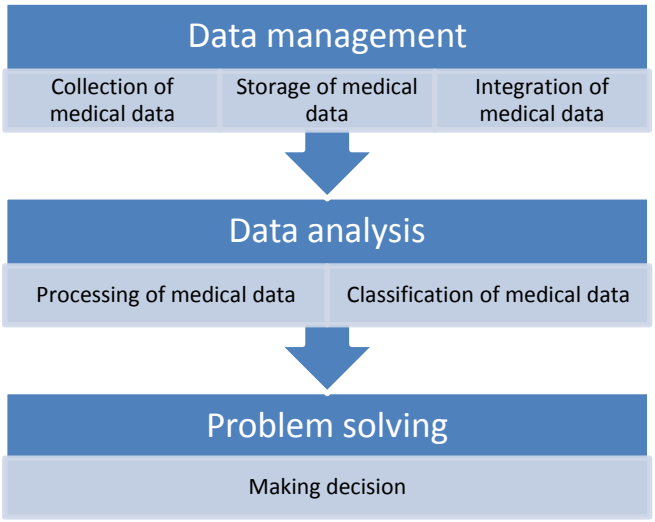

Figure 1: A conceptual model for medical data integration

A workflow was developed for data integration, filtering, sorting and aggregation of Covid-19 data (Fig. 2). The raw medical data consisted of information coming from clinical records covering a period slightly longer than 10 months, from April 2020 to February 2021. The clinical records of 20,400 potential patients were used, of which 10,200 patients were infected with SARS-CoV-2, and the remaining 10,200 hospital patients were not infected. Medical records were linked to 10,232 women and 10,168 men, aged 21 to 85 years.

The data is organized into a structure of 17 variables arranged in the following order: “ID”, “GENDER”, “AGE”, “COVID”, “COVID_SYMPTHOM”, “HEART_DISEASE”, “HYPERTENSION”, “DIASTOLIC”, “SYSTOLIC”, “DIABETS2”, “HBA1C”, “CKD”, “CALCIUM”, ”POTASSIUM”, “PHOSPHORUS”, “CANCER” and with data type integer and text string.

Each field of the proposed data integration and processing model is configured by type, format, and length of data . Designed in Talend Open Studio, the model performs four main tasks: data integration, data filtering, data sorting, and output.

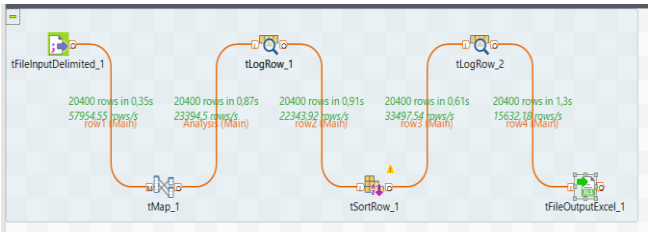

Figure 2: Medical Data Processing Workflow

3 Results and Discussion Problem Solution

The process of developing a data integration workflow begins with the creation of a metadata file based on a test database represented by a .csv file with the data for COVID-19. The generated metadata file is loaded at the input of the data integration workflow via the tFileInputDelimited component. The names of the variables used and the data types required for the integration process are specified. All the variables contained in the .csv file with statistical data are selected for the solution of the given task. Through the tLogRow1 component, an output is defined that shows the result of loading the file with selected attributes. Fig. 3 depicts the loading process and the number of records processed. The resulting output (Fig. 4) shows the availability of the selected variables and records for processing.

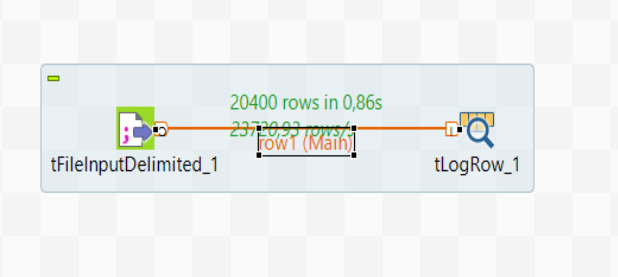

Figure 3: Metadata file loading process

To achieve higher precision in the processing of COVID-19 data, a filter has been added through the tMap component. It contains the variables ID, GENDER, AGE, COVID, COVID_SYMPTHOM, HEARTDISEASE, HYPERTENSION, DIASTOLIC, SYSTOLIC, DIABETS2. The selection of variables is consistent with the medical requirements when registering a COVID-19 illness. According to them, when the presence of a virus is detected, manifested symptoms are reported, patients are defined as high-risk in the presence of heart disease and deviations in blood pressure indicators, as well as diabetes.

For the analysis of the patient's condition, the patient's age was also included (Fig. 5).

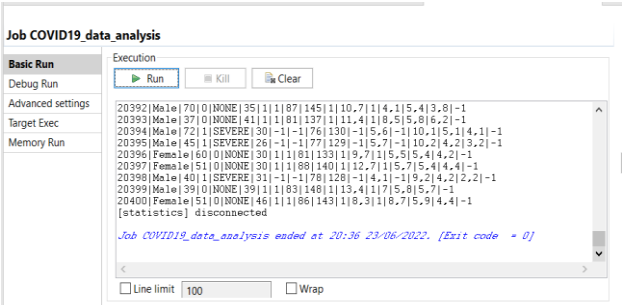

Figure 4: Result of the metadata file loading process

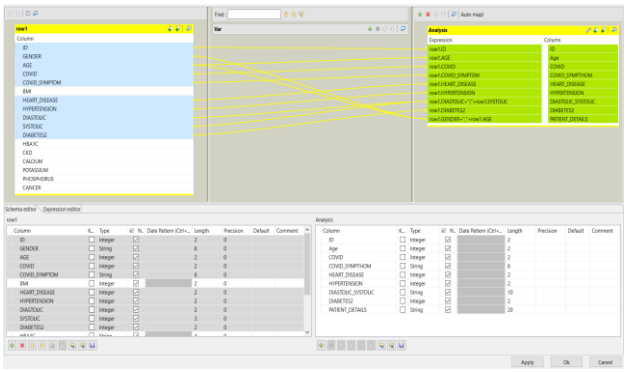

Figure 5: Data processing filter design

A new variable, "PATIENT_DETAILS" , is defined for presenting patient details. It concatenates the records of the "GENDER" and "AGE" variables. " ," is applied as a value separator. The expression is constructed for the implementation:

row1.GENDER+”,”+row1.AGE

Through the developed filter, the diastolic and systolic blood pressures are jointly represented in a common variable "DIASTOLIC_SYSTOLIC", using a delimiter "|" representing the corresponding values. For this purpose, the expression is used:

row1.DIASTOLIC+”|”+row1.SYSTOLIC

Fig. 6 shows the result of the action of the developed filter.

After filtering the data, a sorting process is applied. It is implemented through the tSortRow component, defining a scheme of the variables that can be processed when sorting and those of them, prioritized variables that will form the output result (Fig. 7). The sorting component defines the sorting rules satisfying the given task. In the case under consideration, for the display of all records of patients with a positive COVID-19 test, the variable COVID is defined to be displayed in descending order. Thus, in the result, the cases with a disease (marked with "1"), which require intensive medical assistance, will be displayed as a priority. To

determine the degree of risk for patients with a disease, the variables COVID_SYMPTHOM, HEART_DISEASE, DIABETS2, AGE, ID, configured in descending and ascending order (Fig. 8), have been added. As a result, all cases of COVID-19 disease characterized by severe symptoms and classified by age will be prioritized. The result shows records that indicate the presence or absence of heart disease or diabetes by the values "1" and "-1" respectively. To assess the patient's condition, data on the presence or absence of hypertension, data on systolic and diastolic blood pressure and additional information on the patient's gender and age are included (Fig. 9).

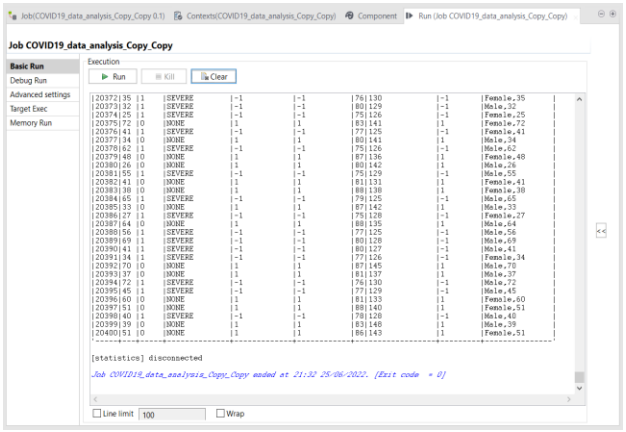

Figure 6: Result of the designed filter action

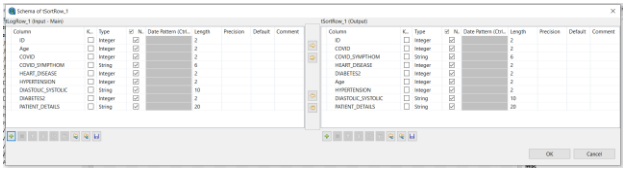

Figure 7: Configuration of the input and output varaibles of the sorting component

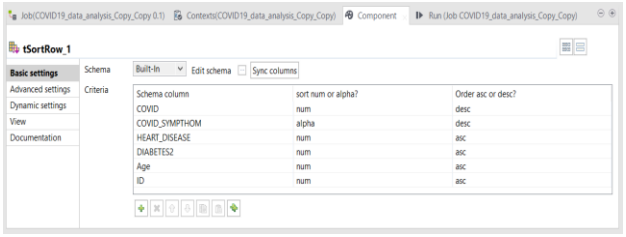

Figure 8: Defining of rules for sorting by attributes

To store the generated result of the data integration process, an .xls file containing the obtained data is generated. For this purpose, the tFileOutputExcel_1 component connected to the output of the tLogRow_2 sorting result component output was used. The component is configured to output and save the result of the integration, with the first line considered as the header.

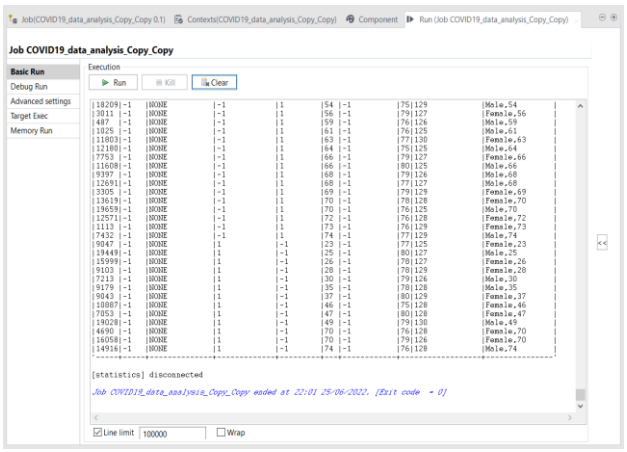

Figure 9: The result of applying a sort process

4 Conclusion

A conceptual model for medical data integration is proposed, consisting of three layers and six phases: data preparation for analysis, interpretation, and visualization, with the preparation phase including medical data collection, medical data storage, and medical data integration. A medical data integration workflow was designed, including steps of data integration, filtering and sorting. The workflow for SARS-CoV-2 medical data from clinical records of 20400 potential patients was employed.

References:

[1] Chen P., Zhang C, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Journal of Information Sciences, 275:314–347, DOI: 10.1016/j.ins.2014.01.015.

[2] Mallappallil M, Sabu J, Gruessner A, Salifu M. A review of big data and medical research. SAGE Open Med. 2020;8:2050312120934839. Published 2020 Jun 25. doi:10.1177/2050312120934839.

[3] Chandra Sekhara Rao, DVLN Somayajulu, Haider Banka, Sawrav Roy, Feature Binding Technique for Integration of Biological Databases with Optimized Search and Retrieve, 2nd International Conference on Communication, Computing & Security [ICCCS-2012], pp.622- 629.

[4] Paton N., etc. (ed.) Data Integration in the Life Sciences: 6th International Workshop, DILS 2009, Manchester, UK, July 20-22, 2009, Proceedings (Lecture Notes in Computer

Science / Lecture Notes in Bioinformatics), Springer, ISBN-10: 3642028780, 2009.
[5] Zhang Zhang, Vladimir B. Bajic, Jun Yu, Kei-Hoi Cheung and Jeffrey P. Townsend, Data Integration in Bioinformatics: Current Efforts and Challenges, Journal Bioinformatics – Trends and Methodologies, November, 2011, pp. 41-56.
[6] Julyeta P.A. RuntuweneIrene, Irene Tangkawarow, C T M Manoppo, Salaki Reynaldo Joshua, A Comparative Analysis of Extract, Transformation and Loading (ETL) Process, IOP Conference Series Materials Science and Engineering 306(1):012066, DOI: 10.1088/1757-899X/306/1/012066.
[7] Talend Open Studio https://www.talend.com/products/talend-open-studio/

Acknowledgments

The presented work was founded by the National Science Fund, Ministry of Education and Science, Republic of Bulgaria under contract KP-06-N37/24, research project “Innovative Platform for Intelligent Management and Analysis of Big Data Streams Supporting Biomedical Scientific Research”.

Creative Commons Attribution License 4.0 (Attribution 4.0 International , CC BY 4.0)