Big Data Integration and Processing Model

STELLA VETOVA
Department Informatics, Technical University of Sofia
Sofia, BULGARIA

Abstract— The presented paper deals with data integration and sorting of Covid-19 data. The data file contains fifteen data fiels and for the design of integration and sorting model each of them is configured in data type, format and field length. For the data integration and sorting model design Talend Open Studio is used. The model concerns the performance of four main tasks: data integration, data sorting, result display, and output in .xls file format. For the sorting process two rules are assigned in accordance with the medical and biomedical requirements, namely to sort report date descending order and the Country Name field in alphabetical one.

Keywords—biomedicine, bioinformatics, big data workflow, 3D model, workflow analysis, biomedical data analysis and visualization

Received: December 1, 2021. Revised: May 4, 2021. Accepted: May 20, 2021. Published: May 31, 2021.

1. Introduction

With the development of areas such as high technology, science and business, the amount of generated data needed for future processing and application increase. They are characterized by complexity, variety, volume and specificity of application and are united under the name big data. For the purposes of processing and application of big data, it is necessary to formulate and implement models for data extraction, storage, integration, management and analysis. Being a multidisciplinary problem area [1], [2], Big Data requires engineering means for large-scale data management and semantics for meaningful data matching and combining. Big data integration is a part of such areas as medicine, biology [3], bioinformatics, where the information characterizes with heterogeneity [4], [5] volume, requiring powerful computing resources for data analysis, transformation and storage [1].

Integration is a term defining the process of data capture from heterogeneous sources, combining various types of data, data filtering to eliminate duplicates, format conversion and management on the base of established rules thus, creating a dataset which is accessible by users. The data integration tools vary according to the model used as a base as follows:

- **Extract, Transform and Load (ETL) [6]**: first, copies of datasets from various sources are collected; second, data are converted according rules; third, data are loaded into a storage model; typically used technique for batch processing and for data format transformation to the warehouse requirements;
- **Extract, Load and Transform (ELT)** a model similar to ETL and appropriate for big data;
- **Change Data Capture**: a model for real time data changes capturing and application to the storage models used;
- **Data Replication**: a model for data replication from the source to another database to keep information safe, available and accessible from different locations;
- **Data Virtualization**: a model for creating a virtual view by means of virtually combining data from different sources;
- **Streaming Data Integration**: a model for different data streams integration, analysis and storage in real time process [7].

To perform traditional data integration or big data integration, data is to be downloaded from its original data source, to be formatted according to the new warehouse requirements and to be transferred to it [8]. According to the data integration architecture [9], [10], the three main steps to be performed are schema alignment, record linkage and data fusion.

Schema alignment concerns the problem of semantic ambiguity and determines the matching in attributes meaning. It produces three kinds of outcomes: a mediated schema which is typically manually created and is responsible for providing a unified view of separate data resources; an attributed matching used for matching attributes in source schemes according to the ones in the mediated schema; a schema mapping based on the source schema and the mediated one used for determining the semantic relationships between the contents of the two types of schemes.

Schema mapping includes three different schema mapping models: global–as-view (GAV), local-as-view (LAV), global-local-as-view (GLAV). The GAV task is responsible for obtaining data in the mediated schema on the base of queries used in the source schema. LAV assists in adding new data source with its individual schema and is responsible for presenting source data as a view of mediated data. GLAV produces views of data of a virtual schema on the base of mediated data and the local one.

For the scenarios of uncertainty in domain model design caused by the large number of data sources which leads to the uncertainty in mediation schema design. Furthermore, the variety of attributes and the problem of their matching result in another aspect of uncertainty. The constant process of data schemas update prevents precise mapping schemas design and maintenance causing third uncertainty. To solve this problem probabilistic schema alignment is used including the means of probabilistic mediated schema and probabilistic schema mappings. The former is a set of mediated schemas each specified with probability to indicate the degree of likelihood between the schema and the domain it describes. The task of the latter is to define and capture the uncertainty of correct mappings.

Record linkage [11], [12] is used for partitioning of a set of records in order to achieve compliance between a partition and its identified records which refer to a distinct entity. It performs in three sequent phases: Blocking, Pairwise

matching, Clustering. Blocking is a strategy for scaling record linkage to large data sets [13], [14], [12]. Its main concept is partitioning of the input records into multiple blocks of small sizes to quarantee data integrity and to prevent it from pairwise matching to the records of a given block. Blocking has the advantage of reducing the pairwise comparisons thus, achieving efficiency and performance. On the other hand, its disadvantage is the availability of false negatives which can be compensated by multiple blocking functions [15].

Pairwise matching is applicable for determining the group affiliation of the records. The proposed methods for performing this mathching varies in a wide range including rule-based methods, classification-based methods, distance-based ones. Clustering is used for detemining the rules to be followed to perform records partitioning keeping the record affiliation to a distinct entity.

The third step of the architecture is data fusion. Its tasks are related to determining the true values for data items for the cases of data errors including mis-typing, incorrect calculations, out-of-date information, etc. [9], [10].

In addition, concerning the hot issue for information security big data is protected against threats in the cloud [16] and the network [17], [18] on the base of developed encryption algorithms.

The following paper presents a proposed model for big data integration and processing in three sections including six phases in total and presents a model for Covid-19 data integration and sorting performed in Talend Open Studio. It is organized as follows: Section 2 is focused on the big data integration and processing model for biomedicine described in details and presented graphically. Section 3 presents the realization and experimental results of Covid-19 data integration and sorting model. Section 4 concludes the paper.

2. Big Data Integration and Processing Model for Biomedicine

The proposed model of big data integration and processing model for biomedicine is structured into three main sections: Information Organization, Information Processing, and Problem Solving. The first section comprises the first three phases of seven in total as graphically illustrated in Fig. 1. Phase 1 “Collecting Biomedical Data” concerns the methods for data obtaining in four orientations. For the purpose of clinical database structuring patients’ data obtained from patients’ exams, symptoms, personal data including age, sex, disease history, date, etc. The types of the data can vary comprising integer numbers, floating-point numbers, characters, string, date, graphical data. In addition, for visual data obtaining apparatuses are used. One of the most often applied apparatuses for imaging are X-Ray, Ultrasound, Magnetic Resonance Imaging (MRI) and Computed Tomography (CT). X-ray is suitable for the cases when areas of concern are available but no symptoms appear. Based on sound waves to produce visual representation of the tissue, the ultrasound [19] reveals the tissue composition and blood flow providing information for the level of suspicion. On the contrary, for the cases of high-risk and advanced stage of disease MRI [20], [21] is the appropriate apparatus. Based on the three dimensional image representation, MRI provides imaging of high quality allowing adequate assessment, treatment, response evaluation and pre-surgery planning. In addition, CT [22], [23], [24], [25], [26], [27] generates medical images for the cases of early detection when no symptoms are available. Unlike the other apparatuses discussed above, this technique can detect very small areas of disease which makes it a powerful tool for the experts.

The produced image data using imaging apparatuses differs in format, resolution, color space as shown in Table. 1. Published in 1993, the Digital Imaging and Communication in Medicine (DICOM) format is recognized as a standard for medical imaging transmission, storing retrieval, printing, processing and displaying medical imaging information. DICOM is implemented in a number of medical areas such as radiology, cardiology, radiotherapy, ophthalmology and dentistry. Being compatible with TCP/IP, DICOM is an application protocol used for communication between systems. It can be exchanged between a sender and receiver for the purposes of image and patient data transmission including metadata: patient name, reference number, study number, dates, reports [28]. Furthermore, DICOM standard enables work with variety of software products and apparatuses produced by different manufacturers. In addition, DICOM guarantees data security using the DICOM TLS protocol for encryption [29].

TABLE I. FEATURES OF IMAGING APPARATUSES

Medical Apparatuses	Image Features
X-ray	2D, DICOM [29] format, greyscale color space, resolution: up to 600px;
Ultrasound	3D image, DICOM format, greyscale color space, resolution:512×512×8 bits or 640×480×8 bits; RGB color space: the upper matrix, 24 bits;
Magnetic Resonance Imaging (MRI)	3D image, DICOM (Digital Imaging and Communications in Medicine) format, greyscale color space, resolution: 64×64, 64×128, 128×128, 128×192, 256×512, 512×512, 512×1024, ...);
Computed Tomography (CT)	3D image, DICOM format, resolution: 2048x2048px;

The second phase “Storing Biomedical Data” refers the process of data storing. Usually, clinical data is collected and stored in file formats “.xls”, “.xlsx”, “.csv”, “.xlsm”. In the third phase “Biomedical Data Integration” data integration process is performed on the submitted clinical data files. In short, integration is a process of uniting data collected from different sources. It includes the steps of cleansing, ETL/ELT, mapping, transformation. The tools for data integration such as Integrator.io, Oracle Data Integrator (ODI), Apache Spark, SQL Server Integration Service (SSIS), Zapier, etc. have thefunctionality of data analysis. The approaches for data sharing vary in type such as web-based and cloud-based approaches. Web-based approach includes platforms which enable users to perform analysis using a web-browser and Internet infrastructure. Thus, web-based technology demonstrates efficiency saving the user a great deal of time providing access to a great number of supercomputing resources and experts’ experience shared in web.

Furthermore, it gives the option for distributed resources access, often using dedicated web-service solutions. Web-based representative is Nora platform. On the other hand, cloud-based approach is a technology often defined as a model for enabling network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction [30], [31]. Representatives of this type of sharing data are Siemens Helthineers and IBM Watson Health.

The process of collecting and managing data from heterogeneous sources is united under the term data warehousing. Its main task is to unite the collected data, to analyze it and to produce meaningful information. The data warehouse is a central data repository where the information is submitted using relational databases. The data itself can be (semi)structured or unstructured. The user access is possible after the data processing and transformation using tools, SQL clients, etc. The three types of data warehouses divide into three as follows: enterprise data warehouse (a kind of centralized warehouse used for decision support), operational data store (used for data storing only), data mart (used for the business purposes).

In the second section "Information Processing" of the presented model phase 4 and phase 5 are included. Phase 4 "Data and Image Processing" concerns the process and methods applied for both data formats processing. The process of data processing includes the manipulation of collected data and performing of functions and operations in order to extract meaningful information. The functions involved include validation, sorting, aggregation, analysis, reporting, classification. Validation is needed to provide guarantee that the information is accurate and relative. Sorting is used for the cases to arrange data in accordance to any submitted requirements. Aggregation is the process of combining multiple data. Analysis is applied for the purpose of cleansing, transforming and modeling data. The process of reporting is producing data summary. Classification performs data separation into groups in accordance with requirements. In terms of image processing, operations for cleaning and improving the quality of images are performed. Filtering is the approach of modifying or enhancing the processed image. Image filtering is performed using frequency components to enhance an image in spatial or geometric aspect in order to suppress the effects caused by the intensity of light. For the purpose of smoothing an image low-pass filtering is suitable to be applied. It is a typical practice to be combined with a basic convolutional operator. On the other hand, median filtering is used to eliminate noise spikes in the two-dimensional image. To enhance an image removing or suppressing noise bend-pass filtering is useful to be applied. In addition, high-pass filtering enhances the images in terms of edges. All the described filters can be used individually to arise the quality of the image or as a part of the preprocessing when the images are processed to be used for subsequent analysis. In some cases, in the preprocessing resizing is also possible. It requires the change in the size of the image to perform another function which requires a certain change in the matrix to satisfy a requirement. For the biomedical image analysis segmentation is the main method. It requires the isolation of a certain pattern for further analysis and useful information extraction for diagnosing. In the end, to describe an image its features based on color, texture or shape are to be extracted using a mathematical setting, thus, generating image feature vectors. The latter are stored in a database and used for further operations to perform certain tasks.

The second phase of section "Information Processing" is phase 5 "Biomedical Data Classification". Data classification is the process of data arrangement in groups based on preset criteria. In this phase, data and image classification are discussed. For the purpose of data such cluster methods as k-Means, k-Nearest Neighbor (KNN), Mean Shift Clustering are applied. When it comes to image classification, artificial neural networks (ANN), convolution neural networks (CNN), similarity computation through Euclidean distance, Huffman, Manhattan, Mahalanobis distance are widely used.

Phase 6 "Decision Making" is the last one phase and it is structured in the section "Problem Solving". It is based on the phases performed earlier using a variety of methods including Machine Learning (ML), Decision Support Systems, Optimization, Computational Intelligence, Heuristics.

The described model is graphically illustrated in Fig. 1.

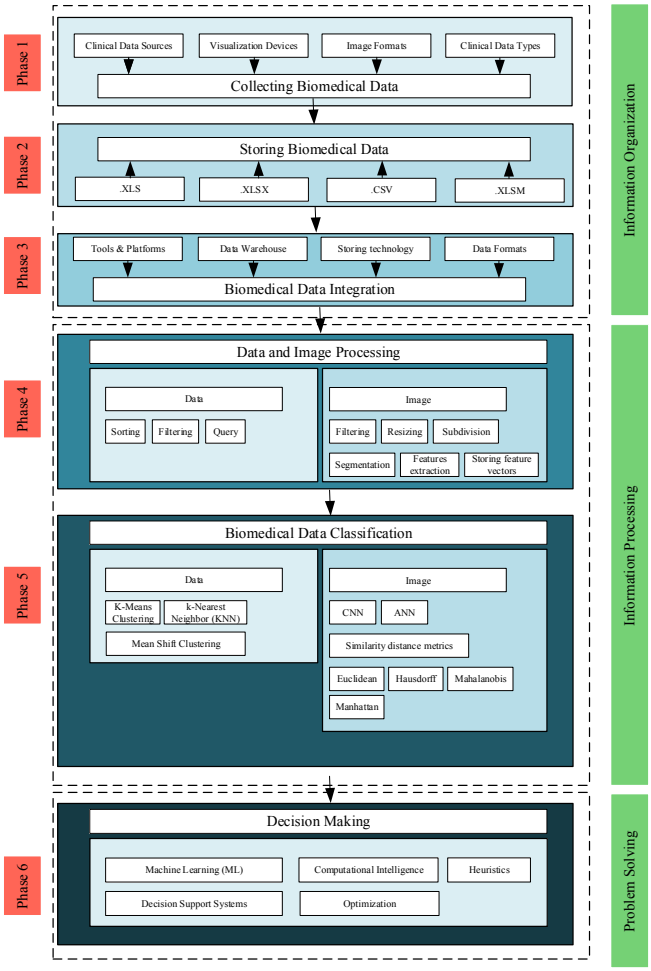


Fig. 1. Big data integration and processing model for biomedicine

3. Data Integration Model and Experimental Results

The purpose of the presented work is to propose a model for Covid-19 data integration and sorting according to rules satisfying medical and biomedical requirements and to present the experimental results. To this end, Talend Open Studio is used. It is an open source ETL software tool for data

integration and big data based on graphical user interface. It enables work with such data sources as RDBMS, Excel, SaaS Big Data ecosystem, SAP CRM Dropbox technologies.

The used statistics file concerns the number of Covid-19 confirmed cases, recoveries, and deaths due to Covid-19 in Europe, by date, country and region. It includes such types of data as date, integer, float, string, Boolean. The used file records are 9699 rows with information up to 2021.03.08. The statistics file format is CSV and comprises the following fifteen fields: date, iso3, country name, region, iat, ion, cumulative positive, cumulative deceased, cumulative recovered, currently positive, hospitalized, intensive care, EUCountry, EUCPMCountry, NUTS.

For the integration process the schema is to be configured according to the data content in the loaded statistics file. To this end, for each of the data field data type, length and format is assigned as shown in Fig. 2.

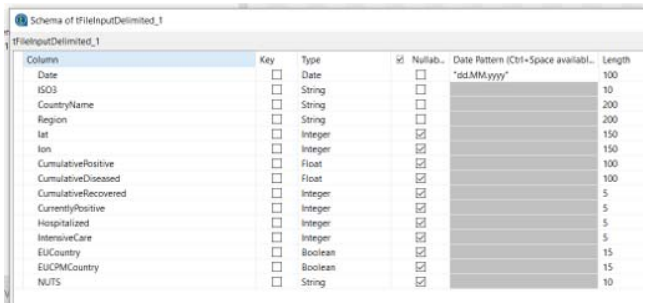


Fig. 2. Covid-19 data fields configuration

The used model for data integration includes the component for data load and configuration and the component for result display. Fig. 3 graphically illustrates the model for Covid-19 CSV file data integration and Fig. 4 integration result display as follows:

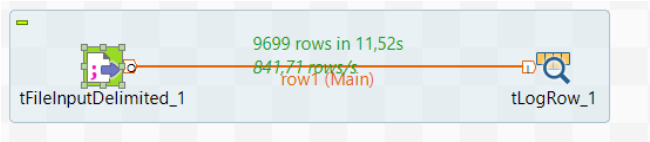


Fig. 3 . Covid-19 data integration model

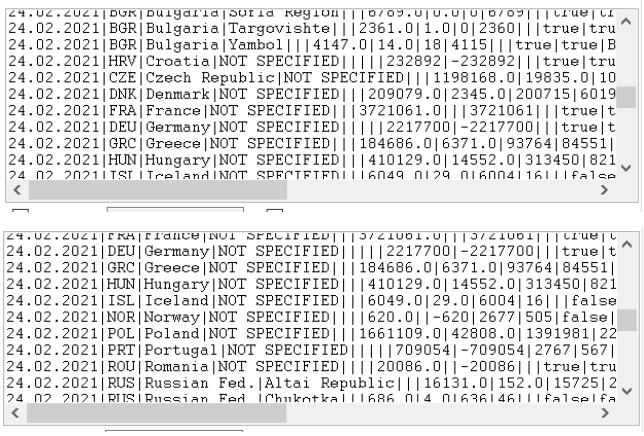


Fig. 4. Covid-19 data integration result

For the purposes of medical and biomedical analysis sorting a file experiment is performed satisfying two sorting rules:

1. Sort date in descending order to display the most recent Covid cases in first positions.
2. Sort the country name in alphabetical order.

For this experiment the schema remains the same as it is described earlier and is shown in Fig. 2. To perform the sorting one more component is added in the model. It inherits the configuration schema of the first component (Fig. 3), thus, no configuration is needed. Fig. 5 presents the configuration for the data load component and the other performing the process of sorting.

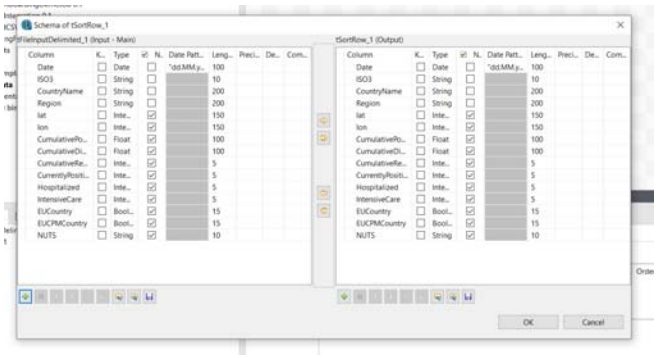


Fig. 5. Configuration schema for data sorting

The sorting rules are assigned to the component performing sorting as follows:

1. For Date the data type is alpha and the sorting order is ascending.
2. For Country Name the data type is alpha and the sorting order is ascending.

In Fig. 6 and Fig. 7 the designed rules are presented and the sorting result, respectively, where the date descending order and the alphabetical order of Country Name for each data is clearly seen.

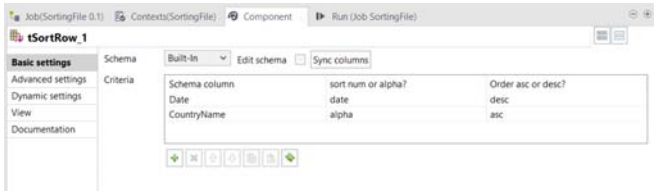
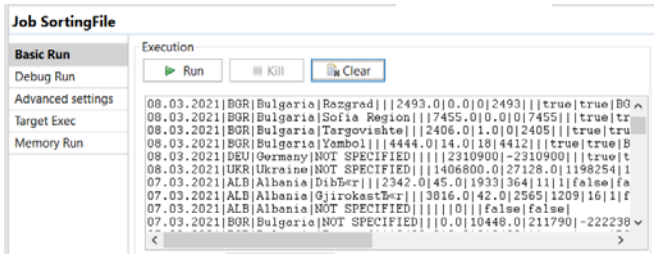


Fig. 6. Sorting rules configuration



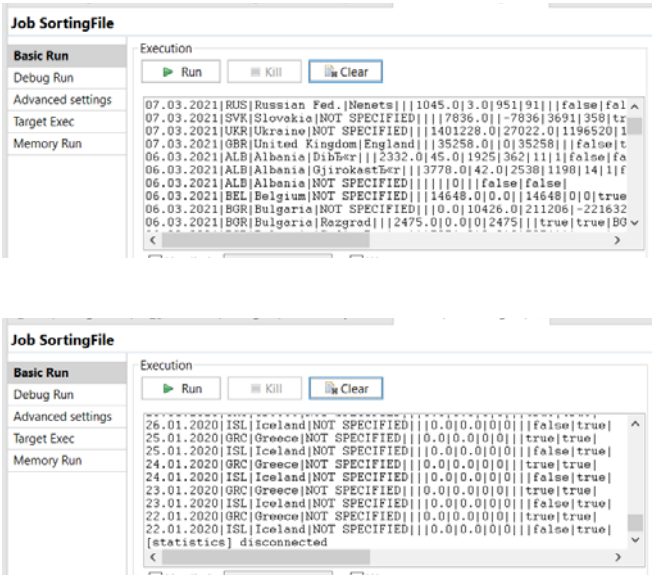


Fig. 7. Sorting result of Covid-19 data

The last step of the experiment includes the sorting data output record in ".xls" format. To this end, a fourth component is added to perform data output. Being linked to the sequence of components for data loading, data sorting according to assigned rules and a component for result display, it generates the ".xls" file with the whole set of sorting data as shown in Fig. 8.

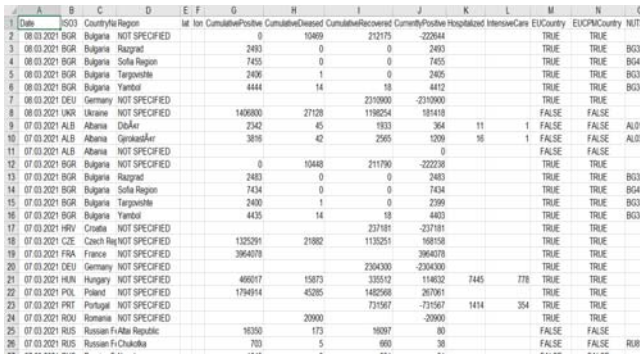


Fig. 8. Sorting data result in .xls format

The final components model for data sorting is illustrated in Fig. 9.

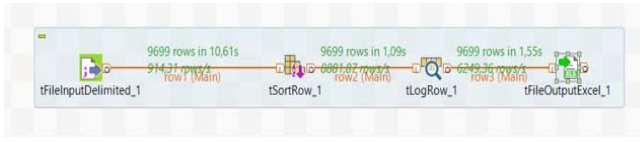


Fig. 9. Sorting data model

4. Conclusion

The focus of the presented report is data integration and sorting of Covid-19 data. To design an integration and sorting model, Talend Open Studio is used. The model performs four tasks: data integration, data sorting, result display, and output in .xls file format. For data integration for each of the fifteen fields data type, format and field length is assigned. To perform sorting two rules are designed and performed according to the medical and biomedical requirements. They concern sorting report date in descending order and the Country Name field in alphabetical order.

Acknowledgment

Primary funding for the presented work was provided 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".

References

[1] C. Bizer, P. Boncz, M. L. Brodie, O. Erling, "The Meaningful Use of Big Data:Four Perspectives–Four Challenges," ACM SIGMOD Record 40(4):56-60, 2011.

[2] M.L. Brodie, M. Greaves, J.A. Hendler, "Databases and AI: The Twain Just Met," STI Semantic Summit, Riga, Latvia, July 6-8, 2011.

[3] V. Gancheva, "SOA based multi-agent approach for biological data searching and integration," International Journal of Biology and Biomedical Engineering, ISSN: 1998-4510, Vol. 13, 2019, pp. 32-37.

[4] V. Gligorijevi´c, N. Malod-Dognin, N. Prž ulj, "Integrative Methods for Analysing Big Data in Precision Medicine," Proteomics,16(5):741-58, 2016, doi: 10.1002/pmic.201500396.

[5] H. Abbes, F. Gargouri, "Big Data Integration: a MongoDB Database and Modular Ontologies based Approach," 20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2016, 2016, pp. 446 – 455.

[6] J. Runtuwene, I. Tangkawarow, C. Manoppo. R. Salaki, "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, 2018.

[7] https://www.omnisci.com/technical-glossary/data-integration

[8] S. Janković, S. Mladenović, D. Mladenović, S. Vesković D. Glavić, "Schema on read modeling approach as a basis of big data analytics integration in EIS," Enterprise Information Systems, 2018.

[9] X. Dong, D. Srivastava, "Big Data Integration," Morgan & Claypool Publishers, 2015.

[10] X. Dong, D. Srivastava, "Big Data Integration," Proceedings of the VLDB Endowment, Vol. 6, No. 11, 2013, pp. 1188-1189.

[11] R. Abd El-Ghafar, M. Gheith, A. El-Bastawissy, E. Nasr "Record Linkage Approaches in Big Data: A State Of Art Study," 13th International Computer Engineering Conference (ICENCO), 2017.

[12] J. Nin, V. Muntes-Mulero, N. Martinez-Bazan, J. Larriba-Pey, "On the Use of Semantic Blocking Techniques for Data Cleansing and Integration," 11th International Database Engineering and Applications Symposium (IDEAS 2007), Banff, Alta., 2007, pp. 190-198, doi: 10.1109/IDEAS.2007.4318104.

[13] D. Bitton, D. J. DeWitt, "Duplicate record elimination in large data files," ACM Transactions on Database Systems, vol. 8, No. 2, 1983, pp. 255-265.

[14] S. Kadochnikov, V. Papoyan, "Blocking Strategies To Accelerate Recordmatching For Big Data Integration," Proceedings of the 27th International Symposium Nuclear Electronics And Computing (Nec'2019) Budva, Becici, Montenegro, September 30 –October 4, 2019, Pp. 219-224.

[15] M. Hernandez, S. Stolfo, "Real-world data is dirty: Data cleansing and the merge/purge problem," Data Mining and Knowledge Discovery, 2, pp. 9–37, 1998.

[16] Ivanov, I., "Basic Cloud Security Threats," Proceedings of Annual University Science Conerence, vol. 6, Veliko Tarnovo, Bulgaria, May 2020, pp. 143 – 147.

[17] Ivanov, I., "Entry Points for Cyberatacs," International Science Conference "Wide Security", vol. 2, New Bulgarian University, Sofia, March, 2020m pp. 336 – 341, ISBN 978-619-7383-19-5.

[18] Ivanov, I., "Analysis of vulnerabilities in web applications," Proceeding of Science Conference "Current Security Issues," Veliko Tarnovo, vol. 6, 2020, pp. стр. 233 – 236. ISSN 2367-7465.

[19] R. Devi and G.S. Anandhamala, "Recent Trends in Medical Imaging Modalities and Challenges For Diagnosing Breast Cancer," Biomedical & Pharmacology Journal, vol. 11(3), p. 1649-1658, September 2018.

[20] SG. Orel, MD. Schnall, CM. Powell, MG. Hochman, LJ. Solin, BL. Fowble, MH. Torosian, EF. Rosato, "Staging of suspected breast cancer: effect of MR imaging and MR-guided biopsy," Radiology, 196(1), pp. 115–22 (1995).

[21] F. S. Azar, D. N. Metaxas, and M. D. Schnall, "A deformable finite element model of the breast for predicting mechanical deformations under external perturbations," Acad. Radiol., 8(10), pp. 965–975 (2001).

[22] Ж. Василева, В. Хаджидеков, В. Тодоров, "Физиката в Биологията и Медицината," Физиката в Образната Диагностика, XXXIV Национална Конференция по Въпросите на Обучението по Физика, Ямбол, 6-9 април 2006 г.

[23] S. Sasada, N. Masumoto, N. Goda, K. Kajitani, A. Emi, T. Kadoya, M. Okada, "Which type of breast cancers is undetectable on ring-type dedicated breast PET?," Clinical Imaging, vol. 51, no. February, pp. 186–191 (2018).

[24] L. Lebron-Zapata, M. S. Jochelson, "Overview of Breast Cancer Screening and Diagnosis," PET Clin., vol. 13, no. 3, pp. 301–323, 2018.

[25] Y. Yamamoto, Y. Tasaki, Y. Kuwada, Y. Ozawa, T. Inoue, "A preliminary report of breast cancer screening by positron emission mammography," Ann. Nucl. Med., 30(2): pp. 130–137 (2016).

[26] D. Narayanan, W. A. Berg, "Dedicated Breast Gamma Camera Imaging and Breast PET: Current Status and Future Directions," PET Clin., 13(3): pp. 363–381 (2018).

[27] H. B. Pan, "The Role of Breast Ultrasound in Early Cancer Detection," J. Med. Ultrasound, 24(4): pp. 138–141 (2016).

[28] R. Bibb, "Export data format and media," Medical Modelling, 2006.

[29] M. Novaes, "Telecare within different specialities," Fundamentals of Telemedicine and Telehealth, 2020.

[30] M. Hogan, F. Liu, A. Sokol, J. Tong, "NIST Cloud Computing Standards Roadmap," National Institute of Standards and Technology, Special Publication, 500-291, 2011.

[31] P. Mell, T. Grance, "The NIST Definition of Cloud Computing," Recommendations of the National Institute of Standards and Technology, NIST Special Publication, 800-145, 2011.