ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Vol 10, Issue 1, January 2023

Data Modeling Best Practices Key to Data Mining and Data Standardization

[1] Pawankumar Sharma, [2] Prasad Chetti, [3] Bibhu Dash, [4] Meraj Farheen Ansari

[1] [3] [4] University of the Cumberlands, KY USA.
[2] Northwest Missouri State University, MO USA.
Corresponding Author Email: [1] spawankumar570@gmail.com, [2] prasad.chetti@gmail.com,
[3] bdash6007@ucumberlands.edu, [4] mansari6529@ucumberlands.edu

Abstract*— The advancement of technology has resulted in prompt action in identifying the various methods of collecting the data generated for later extraction. In the ETL and ELT world, data extraction, commonly known as data mining and AI, helps companies decide as they learn and predict customer behaviors and economic patterns. The onset of technology has highlighted the various data models through which the organization information flow within an organization is hence deciding the data processing and extraction processes such as network, hierarchy, E-R, and the relational model. Data mining occurs through a series of steps to guarantee the generation of valuable data, as exemplified by data cleaning, integration, reduction, transformation, mining, evaluation, and representation. However, data extraction faces various challenges: privacy and confidentiality, data collection, and preparation. This paper highlights the process, and best practices for data modeling as business needs change.

Index Terms*—ETL, ELT, Network model, E-R model, Hierarchical model, Relational model, Data mining, AI.

I. INTRODUCTION

Data modeling revolutionized how the industry stores, analyzes and accesses data. In the ETL(Extract, Load, and Transform) era, whether an organization is creating a new data warehouse or re-engineering the existing one, it follows the guidelines and best practices to store data effectively for the success of the project and efficient retrieval for better machine learning(ML) modeling and dashboarding. Improved data storage techniques facilitate better data mining, and data mining comprises the extraction of the data through the various patterns defining the anomalies and patterns alongside the correlation within the extensive data sets, helping predict future outcomes. The advancement of technology has led to the improvisation of various techniques supporting data extraction as organizations seek to expand their market and monetize their products across online platforms. Organizations marketing across the online account for some of the benefits organizations gain from data extraction hence the extensive research on the data mining process [1-3]. As a result, several research findings have prompted the development of various techniques which help various organizations increase revenues and economies of the costs incurred while enhancing the customer relationship and mitigating the risks exposure of the business. Basic scientific disciplines for data mining include statistics (data interrelation studies) and artificial intelligence (human-like intelligence in software). In addition, they also include machine learning (data algorithms for future predictions) [3]. The data extraction has led to various models through the multiple steps improvised, although it leads to different challenges in achieving standard data as outlined below.

This paper offers and discusses ideas for structuring corporate data strategy and achieving company-wide alignment, guaranteeing that the architecture of the data warehouse business will fulfill both current and future business requirements. All business stakeholders will probably benefit more from the data warehouse, which will also give the blueprint for a data store that can develop and adapt as business requirements change in response to new technological trends.

II. DATA MODELS

As AI, Blockchain, and data analytics increase in organizational decision-making, the first critical stage in developing a data warehouse that fits all needs is establishing a data model. This abstract representation organizes data elements and describes how they relate to one another and aspects of their real-world entities. A data model establishes a shared understanding and description of the information critical to the firm and the organization's larger data environment [4]. A data model can be used to document the data sets included in the data warehouse, their relationships, and the data warehouse's business requirements. Data models entail the foundation structure for an effective database. The data model comprises entities, attributes, constraints, and interrelationships. Therefore, they help in data representation and the database's storage modification within the database management system [4, 5]. Data scientists have classified data models into four; the hierarchical model, the network model, the entity-relationship model, and the relational model.

A. Hierarchical Model

The hierarchical data model entails the various data tree-like structure organization with one primary root providing linkages to other data (see Fig 1). As employed within this structure, the main hierarchy originates from the

ISSN (Online) 2456 -1304
International Journal of Science, Engineering and Management (IJSEM)
Vol 10, Issue 1, January 2023

root with expansion to the tree along the various child nodes as the tree expands. Likewise, to the standard tree structure, the child data node bears a single parent node while the parent acting as a linkage to the child node comprises the link to the multiple child nodes [6].

Data storage in a tree-lie structure facilitates traversing the entire tree from the root node in any case of data extraction and retrieval. The model has, therefore, the various data interconnected based on one-to-many related data. Accordingly, this model's storage obeys the associated data's linkage mode. This data model correlates with the organization's employee data storage model in which the data table contains the employee's name, code, department, and last name, which facilitates data retrieval [6]. All the employees on whom the organization has issued computers have central storage for every information processed as related to the organization. Hence, the data center storage acts as the parent node distributing the information to the various employees based on their departments and codes.

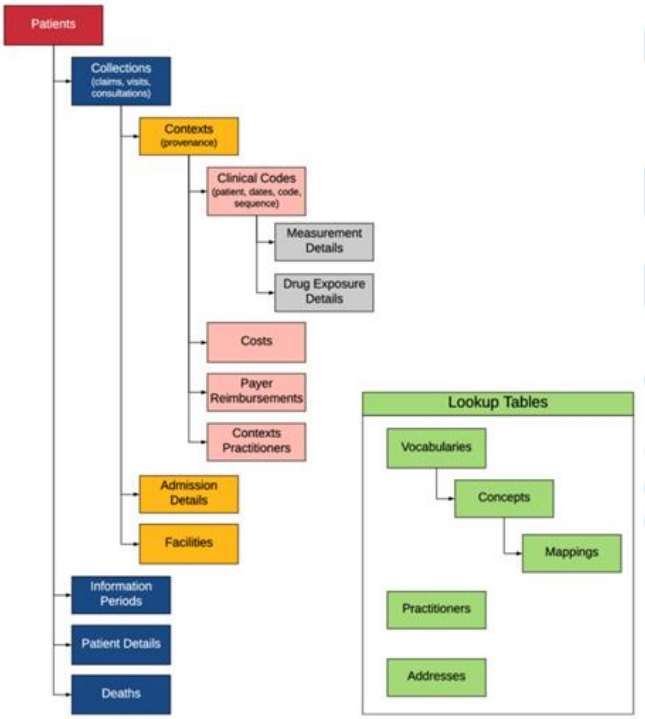

Fig 1. Hierarchical Data Model. Adapted from [4]

B. Network Model

The network data model has a flexible data approach within the database to facilitate the data representation for the various objects and the interrelationship with the multiple objects. The schema employed in the data structuring bears a unique structure of the graphical representation through the utilization of edges as the nodes represent the various objects [7]. Compared to the graphical data representation, the data representation in the hierarchical data form explains the unique structure defining the differences between the hierarchical data model and the network model. The network model presents the ultimate benefit in the essential connections as facilitated within this model as it comprises various interrelationships between the data types comprised in the model [7, 8]. Hence, the data has easy accessibility compared to the other data model exemplified by the hierarchical data model. As defined within this data model, the parent and child nodes remain interconnected due to the interrelationship between the parent-child node. However, the data has no dependency on the other node. The model's disadvantages accrue from the inability to adapt to the various changes. Therefore, any modification of the system will demand a tedious and time-consuming action change the system [8]. The data maintenance has multiple challenges due to the requirement for the record connection through the different pointers, hence the system's complexity.

C. ER Model

This model has a unique feature from the database structure utilization of the entity-relationship diagram. The model, similar to the database structure, helps implement the database. Individuals who form their data embedded within this structure enjoy the ability to illustrate the interrelationship between the data entity set [9, 10]. In addition, the entity set comprises a similar entity composed of various attributes. The data structure includes the unique character in which the components account for the interrelationship between the entity set and features. E-R diagram representation accounts for the particular data set as demonstrated in the two college and student entities with numerous interrelationships due to the high number of students [10]. The double rectangle represents a weak entity whose unique character accounts for the attributes uniquely confined in an entity that requires interrelationships. For instance, a bank fails to recognize an account to link the bank name to the respective bank account [8, 10]. The E-R diagram further illustrates the various attributes demonstrated by the critical, composite, multivalued, and data attributes (see Fig 2).

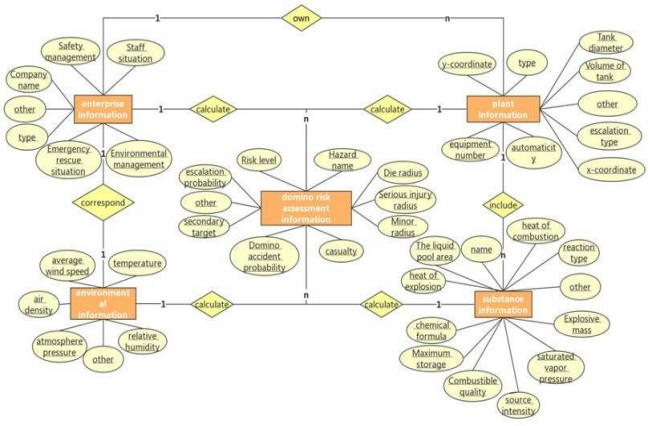

Fig 2. ER Data model. Adapted from[13]

D. Relational Model

The relational model uniquely uses data tables to collect various elements within the group into their respective

ISSN (Online) 2456 -1304
International Journal of Science, Engineering and Management (IJSEM)
Vol 10, Issue 1, January 2023

relations. The interrelated tables are the interrelationships and data [8]. Within the tables, multiple rows and columns represent entity attributes and records, respectively. The various primary keys distinguish the documents in the table; hence the primary essential forms the fundamental tool in history. For instance, Structured Query Language (SQL) helps retrieve the various data elements [8]. The various data entries within the data set have unique features, enabling data retrieval. However, the data table should omit all data inconsistencies as it may culminate in various challenges during the data retrieval. In addition, data duplicity, incomplete data, and various inappropriate links help link the various data.

A well-defined data model positively impacts even after the data warehouse is used. For all things in the data warehouse, for instance, a data model outlines the data lineage, making it simple to add new data objects or onboard new team members when business requirements change.

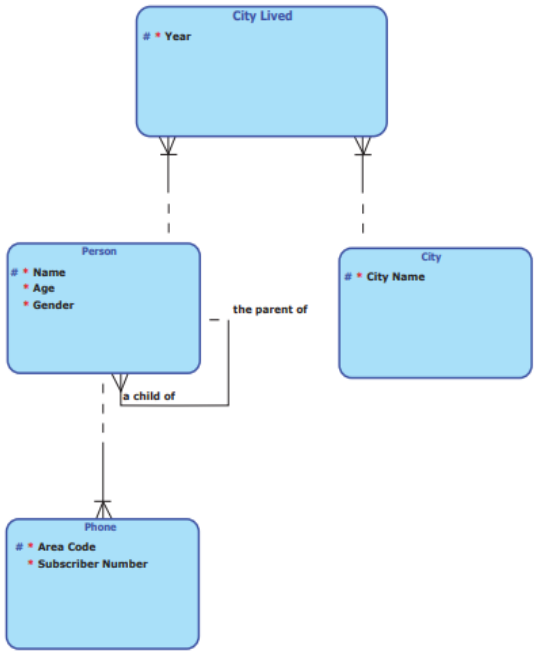

Fig 3. Relational data model.

III. DATA MODELING KEY FOR IMPROVED DATA MINING

The data model provides in-depth documentation of the content, context, and sources in addition to providing information gathered from several historical and contemporary systems by the data warehouse. Because of this, it is simpler to audit or follow new data standards thanks to the GDPR (the EU's General Data Protection Regulation framework that specifies guidelines for collecting and processing personal data) [11]. The future avoidance of misunderstandings and costly re-engineering is another benefit of a robust data model. Always a smart notion is a source-neutral integration layer that enables analysis across many data sets based on the shared properties of the data sets.

Better data modeling facilitates better data understanding, mining, and insight extraction using AI algorithms. Data mining occurs in two steps: data preprocessing and data mining. Data preprocessing entails data cleaning, integration, reduction, and transformation, while data mining comprises data mining, pattern evaluation, and data knowledge representation [10, 12]. Data preprocessing depends on various factors in determining data viability, accuracy within the database, completeness of the data table, and timeliness. The data preprocessing includes various steps, data cleaning, integration, reduction, transformation, mining, evaluation, and knowledge representation. Data cleaning entails the first step of data processing. It expounds on the importance of dirty data upon its usage directly in the mining process, leading to confusion within the procedures and culminating in inaccurate results. This step removes the various incomplete data from the system using various methods. The unique steps carry the routine work: filling in the missing data, ignoring the tuple, and manually filling in the missing values. In addition, it utilizes measures of central tendencies, such as the median, in filling the most probable value [10]. The data cleaning process also entails removing the noisy data as a representative of the random error. The binning method helps remove the noisy data by sorting the various values into buckets or bins. The consultation of the different neighboring values helps smoothen the data. The bin replacement by the mean helps with smoothening, although the median and boundaries can also help to smoothen the data [10]. Boundary smoothening allows setting the minimum and maximum values within the bin, and the closest boundary value replaces the bin value.

Data integration entails multiple heterogeneous data sources analysis, as exemplified by databases, data cubes, and files. Various databases allow for different naming conventions of variables through the redundancy caused by the databases. Additional data cleaning will eliminate redundancies from the integrated data without interferences of the data reliability [10]. Data migration occurs through tools exemplified by oracle data service integrator and Microsoft SQL allow data integration.

Data reduction entails the collection of available data for analysis. The representation size must remain smaller in volume while upholding the data integrity. Therefore, various processes facilitate data reduction, including Naïve Bayes, Decision Trees, and the neural network [1, 13]. The various data reduction strategies utilized in this step include reducing the number of attributes within a dataset, commonly known as dimensionality reduction. Further, they include numerosity reduction, which entails the original data volume replacement by the various more minor forms of data representation. In addition, the data compression utilization in the data reduction facilitates the actual data representation in compressed data.

The data transformation step in the data mining process allows for data transformation into a more suitable form. In

ISSN (Online) 2456 -1304
International Journal of Science, Engineering and Management (IJSEM)
Vol 10, Issue 1, January 2023

contrast, data consolidation within this step allows for efficient data mining and an easier understanding of the patterns. The change of data entails data mapping and code generation. The strategies employed within this step include smoothing, aggregation, normalization, and discretization. The smoothing process strategy facilitates the removal of noise data through clustering and regression techniques [1]. The aggregation facilitates the summary operation applied to the data, while normalization enables data scaling into the smaller ranges. However, the interval data replacement through the raw numerical data values, such as age, occurs through the discretization strategy.

The data mining step entails identifying various exciting patterns and accrued knowledge from the extensive data volumes. Intelligent patterns application allows the extraction of data patterns with a representation of the data occurring in practices and model structuring using classification and clustering techniques. Pattern evaluation involves identifying target patterns with corresponding information per target data [12, 14]. The process facilitates user data understandability through data summarization and visualization methods. The final step involves data representation using visualization and knowledge representation tools for the mined data. The visualization of data occurs through tables and reports.

IV. CHALLENGES INVOLVED IN MAINTAINING BETTER DATA STANDARDS

The definitions and semantic structures used in a highly successful data model should be set by the company, not by the particular definitions of any one source system. Data standardization faces various challenges, from data extraction to the ultimate data representation. For instance, data collection presents a challenge in standardization and leads to challenges in accessing and collecting the necessary data volumes. Challenges in obtaining relevant data still exist, directly impacting the building robust machine-learning models necessary for various data mining [1]. Besides, organizations collect vast amounts of data without effective utilization of the data hence missing various crucial benefits likely to realize from the data. The sheer abundance of data sources challenges the accessibility of data. Various organizations collect data from the various employees, sales, and customers using numerous tools and software whose sheer data volume firing attributed to the various challenges during the data consolidation and management. Organizations will face challenges in consolidating data from disparate and semi-structured sources, making the data extraction and mining complex [9, 15]. Organizations can overcome this challenge by organizing data into more meaningful data sources.

Data mining faces challenges in the accessibility of the data due to the various imposed securities and privacy measures. The advancement of technology has increased privacy and compliance concerns hence the challenges in the data set access [16]. Besides, cloud storage has increased the vulnerability of the data to various cyberattacks; hence the demand for more security strengthened, and regulatory requirements inconveniencing data security [2]. The organizations, which provide the interested parties access to their datasets, also face challenges in ensuring continued safety and data protection adherence as the breach of data securities culminates in severe financial penalties and costly regulatory audits.

Data preparation presents extensive challenges to data miners. For instance, data miners need help to obtain suitable datasets and gain access [17]. The various accessible data have a messed database which implicates the data scientists and machine learning consuming more time in the data processing and preparation due to the data inconsistency and structured data analysis [3, 18]. Machine learning models depend on the demand for accurate data mining, whose limitation accrues to the extensive nature of the data preparation [9]. Besides, managing such chunks of data attracts ample time and robust infrastructure to support the data mining software and perform various data analytics.

V. STANDARD DATA MODEL APPROACH

Data modeling needs to be loosely connected to any specific system as firms undergo transitions, mergers, and acquisitions. Picking a consistent and future-centric strategy for the data model is crucial. The following are the primary categories of data modeling standards applied in data warehouse design:

A. 3NF

The architectural standard 3NF, or "third normal form," was created to reduce data duplication and guarantee the database's referential integrity [19, 20].

B. Star or snowflake Schema

The most basic and widely used architecture for establishing data warehouses and dimensional data marts is the star schema, which consists of one or more fact tables that refer to any number of dimension tables [19]. The complex structure of data warehousing design consisting of many star-schemas is snowflake schema (see Fig 4).

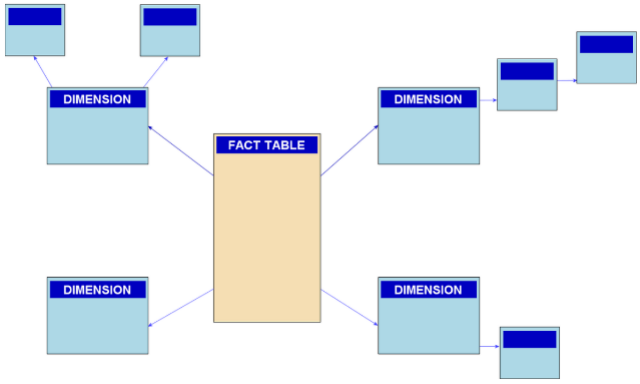

Fig 4. Snowflake Schema [19]

C. Data Vault (DV)

DV modeling was established as a granular, non-volatile, auditable, readily extendable, historical store of corporate data to solve the agility, flexibility, and scalability concerns in existing techniques. It is highly normalized and incorporates 3NF and star model components [20].

It's important to highlight that there is no such thing as a one-size-fits-all model. Each architecture offers benefits, but choosing one will be decided by the organization's business requirements and future needs.

VI. BEST DATA MODELLING BEST PRACTICES

These data warehouse and data modeling development best practices will increase the possibility that all business stakeholders will gain more excellent value from the data warehouse and establish a structure that can grow and adapt as the current world of business demands change endlessly. The best practices are outlined below.

A. Adopt an agile approach

As the traditional waterfall paradigm fails, agile development approaches are increasingly being adopted in software development projects. That is no longer the norm in today's world, as many businesses opt for a more flexible and iterative design process known as Agile [17, 21]. Companies must be able to adapt to change as business requirements change faster than ever before, and new data sources emerge more often. explain these concepts quickly and succinctly. This requires learning about incremental and Agile data warehousing and analytical solution development.

Data warehouse architects are using the Agile methodology, which has its roots in the software development industry, to achieve this goal. In the Agile approach, requirements and solutions develop via the cooperative effort of self-organizing, cross-functional teams and clients. This approach is the best fit for designing modern data warehousing and data solutions. Scrum, Kanban, and BEAM (Business Event Analysis and Modelling) are very popular in IT companies for faster and better data model development [22].

B. Favor ELT over ETL

Previously, the extract-transform-load (ETL) process was employed in data warehouse creation. This strategy entailed taking data from the source systems, cleaning it up or applying business rules to it on an external server, and feeding it into the destination data warehouse. As a result of enhanced data warehousing computer power and capabilities (ELT), a new preferred technique, extract-load-transform, has evolved.

The ELT approach has two primary benefits: reduced costs and enhanced traceability [22]. ELT lowers expenses by enabling businesses to transform data using the data warehouse's capabilities rather than an external server. Transferring data directly to the warehouse is quicker and less expensive since cloud computing capability is frequently much less expensive than carrying out transformations and data processing on an external server. The ELT approach also makes it simpler to audit and follow the data in the future since it immediately displays a snapshot of the original source data inside the data warehouse. This method involves storing raw data in the data warehouse, which has come to be known as a "data lake.

C. Adopt data warehouse tool for automation

Data is intended to be quickly activated and sent by the data warehouse to inform business decisions and create extra value. One tactic to speed up delivery is to adopt the Agile approach. Another choice is to develop and deploy code more quickly using automation tools [23]. The code required for data loading and structuring is usually repeated due to the pattern-based nature of many data warehouse systems, allowing automation. An increasing number of technologies available can automate some or perhaps all of the design and building processes.

In the industry 4.0 era, with modern automation tools, code generation, understanding relations, following standards, and deployment are easy without or with fewer syntax errors. Managing code with automation tools means adding more value, changing the old template, and quick testing with minimal errors or time constraints.

D. Train staff on new approaches

Moving to automated code development or the Agile approach requires a shift in mindset and skill sets. Training and instruction are needed to ensure that a data warehouse team effectively implements these new approaches and technology [18]. This might include bringing in outside professionals to teach groups about Scrum best practices or training teams about the benefits, rules, and best practices for whatever standard architecture the organization has selected for its data warehouse.

As with any new process or cultural shift, organizations should manage the adoption curve to ensure a smooth and successful transition to the new approach in day-to-day operations. Finding proof-of-concept projects will ensure practitioners learn and perfect the procedures in secure yet practical settings, boosting competence and ability in these new skills [21, 23].

VII. CONCLUSION

Finally, effective data modeling techniques and the data mining process significantly benefit the various data miners utilized by the various companies in predicting future consumer behaviors and the economy, hence helping in decision-making and self-service analytics. However, the data mining process passes through the multiple models supporting mining, such as hierarchical, E-R, network, and relational models, whose unique features help identify the correct flow chart for extraction. Data mining involves data cleaning, integration, reduction, transformation, mining,

evaluation, and the ultimate knowledge presentation. Although the data extraction process might seem more straightforward, it attracts various challenges exemplified by the data preparation, collection, and privacy concerns amid the increased confidentiality issues during the cloud storage era.

REFERENCES

[1] Andreeva, P. (2006). Data modelling and specific rule generation via data mining techniques. In International Conference on Computer Systems and Technologies-CompSysTech.

[2] Otter, A., Murphy, J., Pakrashi, V., Robertson, A., & Desmond, C. (2022). A review of modelling techniques for floating offshore wind turbines. Wind Energy, 25(5), 831-857.

[3] Accorsi, R., & Lebherz, J. (2022). A practitioner views process mining adoption, event log engineering, and data challenges. *Lecture Notes in Business Information Processing*, 212–240. https://doi.org/10.1007/978-3-031-08848-3_7

[4] Li, G., Zhou, X., & Cao, L. (2021, June). AI meets database: AI4DB and DB4AI. In *Proceedings of the 2021 International Conference on Management of Data* (pp. 2859-2866).

[5] Baazouzi, W., Kachroudi, M., & Faiz, S. (2022). Towards an efficient fairification approach of tabular data with knowledge graph models. *Procedia Computer Science*, *207*, 2727–2736. https://doi.org/10.1016/j.procs.2022.09.331

[6] Danese, Mark & Halperin, Marc & Duryea, Jennifer & Duryea, Ryan. (2019). The Generalized Data Model for clinical research. BMC Medical Informatics and Decision Making. 19. https://doi.org/10.1186/s12911-019-0837-5.

[7] Dash, B., Ansari, M. F., Sharma, P., & Ali, A. (2022). Threats and opportunities with AI-based Cyber Security Intrusion Detection: A Review. *International Journal of Software Engineering & Applications*, *13*(5), 13–21. https://doi.org/10.5121/ijsea.2022.13502.

[8] Elayam, M. M., Ray, C., & Claramunt, C. (2022). A hierarchical graph-based model for mobility data representation and analysis. *Data & Knowledge Engineering*, *141*, 102054. https://doi.org/10.1016/j.datak.2022.102054

[9] Fraczek, Konrad & Plechawska-Wojcik, Malgorzata. (2017). Comparative Analysis of Relational and Non-relational Databases in the Context of Performance in Web Applications. 153-164. https://doi.org/10.1007/978-3-319-58274-0_13.

[10] Harrington, J. L. (2016). The relational data model. *Relational Database Design and Implementation*, 89–105. https://doi.org/10.1016/b978-0-12-804399-8.00005-3

[11] Mohindru, G., Mondal, K., Dutta, P., & Banka, H. (2022). Mining challenges in large-scale IoT data framework – A machine learning perspective. *Advanced-Data Mining Tools and Methods for Social Computing*, 239–259. https://doi.org/10.1016/b978-0-32-385708-6.00019-9

[12] Seidl, T. (2020). Keynote: Data Mining on Process Data. *2020 2nd International Conference on Process Mining (ICPM)*. https://doi.org/10.1109/icpm49681.2020.00011

[13] Sen, P., Jain, R., Bhatnagar, V., & Illiyas, S. (2022). Big Data and ML: Interaction & Challenges. *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*. https://doi.org/10.1109/iciccs53718.2022.9788257

[14] Sharma, P., Dash, B., & Ansari, M. F. (2022). Anti-phishing techniques – a review of Cyber Defense Mechanisms. *IJARCCE*, *11*(7). https://doi.org/10.17148/ijarcce.2022.11728

[15] Yang, Y & Qian, Y. (2019). DOMIRISK: A User-Friendly Domino Effect Decision Support System. IOP Conference Series: Earth and Environmental Science. 401. 012017. https://doi.org/10.1088/1755-1315/401/1/012017.

[16] Yang, X., Yu, M., & Liu, F. (2022). Construction of power network operation and maintenance cost prediction model based on Data Information Mining. *2022 International Conference on Big Data, Information and Computer Network (BDICN)*. https://doi.org/10.1109/bdicn55575.2022.00031

[17] Dash, B., & Ansari, M. F. (2022). Self-service analytics for data-driven decision making during COVID-19 pandemic: An organization's best defense. *Academia Letters*, 2.

[18] Yuan, C., Fang, F., & Ni, L. (2022). Mallows model averaging with effective model size in fragmentary data prediction. *Computational Statistics & Data Analysis*, *173*, 107497. https://doi.org/10.1016/j.csda.2022.107497

[19] Jyothi, B. S., & Jyothi, S. (2015). A study on big data modelling techniques. *International Journal of Computer Networking, Wireless and Mobile Communications (IJCNWMC)*, *5*(6), 19-26.

[20] Chetti, P. (2020). Assigning Risk Ranks to Civil Infrastructures using Big Data Analytics and Correlation Network Models.

[21] Brdjanin, D., & Maric, S. (2013). Model-driven techniques for data model synthesis. *Electronics*, *17*(2), 130-136.

[22] Borrego, G., Morán, A. L., Palacio, R. R., Vizcaíno, A., & García, F. O. (2019). Towards a reduction in architectural knowledge vaporization during agile global software development. *Information and Software Technology*, *112*, 68-82.

[23] Hannila, H., Silvola, R., Harkonen, J., & Haapasalo, H. (2022). Data-driven begins with DATA; potential of data assets. *Journal of Computer Information Systems*, *62*(1), 29-38.