International Journal of Information Science and Computing
Citation: *IJISC*: **08**(02): 69-78, December 2021
DOI: 10.30954/2348-7437.2.2021.3

INFORMATION SCIENCE AND TECHNOLOGY FOUNDATION

Data Mining and Warehousing-A view & Case Study

Mohmmed Ashraff Shemsudheen and Shanavas Moosafintavida

Azteca University, Mexico, North America

*Corresponding author: shanavas.musafi@gmail.com

Received: 10-08-2021 **Revised:** 23-11-2021 **Accepted:** 05-12-2021

ABSTRACT

The aim of this paper is to show how important data warehouses and data mining are. It also aims to demonstrate the data mining process and how it can assist decision-makers in making informed decisions. A literature review on data mining and data warehousing formed the basis of this paper. The models were created using information gleaned from a literature review and a real-world application. The phases of data mining methods, which are illustrated by the established model, and the relevance of data warehousing and data mining are the most significant findings. It can assist in obtaining clearer responses, allowing both technical and non-technical users to make even more informed decisions. In practice, data warehousing and data mining are extremely beneficial to any company with a large volume of data. Standard (operational) databases may benefit from data warehousing and data mining. Because of the right decisions taken with the aid of data mining, they also help to save millions of dollars and raise profit. This paper explains how data mining works and how it can be used by any company to help users get better answers from massive amounts of data. It demonstrates a different method of querying data. Instead of performing standard database queries, data mining goes a step further by collecting more valuable data.

Keywords: Component, Data Mining, Data Warehousing, Operational Database

Have you ever considered the suggestions you get while shopping online? If you buy a refrigerator online, for example, the website can suggest other items that you should consider purchasing. Have you ever considered the warnings you get from your bank when you use your credit card in a different city? These are data mining instances, which is the process of finding useful trends in a large data set. This massive amount of data is generated by combining current and historical data from various sources and storing it centrally in a Data Warehousing (DW)[1] repository. DW is a critical repository, especially for

How to cite this article: Shemsudheen, M.A. and Moosafintavida, S. (2021). Data Mining and Warehousing-A view & Case Study. *IJISC.*, **08**(02): 69-78.

Source of Support: None; **Conflict of Interest:** None

historical data and non-routine transactions. For example, consider historical data on customer purchase transactions at modern supermarkets. Keeping this type of data in a standard database would result in it being very large, resulting in slower results. For these reasons, historical data and non-routine transactions should be stored in a data warehouse for data mining[2].

Data warehousing and standard databases are designed in various ways. Dimensional modeling techniques are used in data warehousing, while an Entity Relationship Model is used in standard database design[3]. Multidimensional modeling (for example, star schema) improves performance[4]. Data Mining (DM) is a database and artificial intelligence hybrid that is used to provide valuable knowledge to both technical and non-technical users in order to help them make better decisions. It is commonly used as a decision-making aid[5]. The process of DM is not easy. There are many types of feedback, and the whole process can need to be replicated at times.

As a result, data mining is known to be an iterative process[6]. It is divided into six stages:

- Define the problem
- Prepare the data
- Explore the data
- Model
- Evaluate
- Deploy[7].

Data mining can help to automate the knowledge extraction process. This is why it is used in a variety of fields, especially science and business, where large amounts of data must be analyzed[8]. Web mining is one of the most popular applications of data mining. The Internet is becoming increasingly relevant and integrated into our daily lives. With terabytes of data being added every day, data mining techniques for extracting information are becoming increasingly relevant[6].

RESEARCH METHOD

In this paper, I used a combination of a literature review[9] on data mining[7, 10] and data warehousing[4, 11-13] as well as real-world findings from a case study[14]. Previous research on data mining and data warehousing assisted me in establishing a theoretical basis for this subject. My academic literature review improved my understanding of data warehousing and data mining, and then assisted me in identifying the key factors in the data mining process. The case study's real-world findings often highlight the knowledge gained from the literature review.

DATA, INFORMATION AND KNOWLEDGE

Data: facts or a summary Data consists of numbers and messages. Data is fed into computers as input. There are three types of data that computers can process: operational, non-operational and meta data[7].

Information: It can be given by establishing a link or connection between data. Computers transform data into information[7].

Knowledge: Patterns or connections between past and future data may be very helpful. For instance, combining historical sales data with consumer data will reveal details about customers' purchasing habits[7].

DATA WAREHOUSING

Present and historical data should be accessible for the data mining process in order to provide reliable information but holding historical data in a normal database will have a detrimental impact on the database itself. Old data is usually not used for day-to-day transactions, but it is used for data processing and monitoring purposes. Storing historical data in a regular database would result in a significant increase in its size, resulting in slower efficiency. Moving old data from various sites and integrating it all in a new archive known as a data center[12] is a good idea. There are three stages to moving data from operating databases to a data warehouse: Cleaning, transformation, and integration are the three steps[11]. There are several meanings for a data warehouse. The most famous definition comes from Bill Inmon, who says, "A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile repository of data in support of management's decision-making process."[1] Data Warehouse has following characteristics:

- subject-oriented: DW can be used to analyze any subject.
- integrated: DW integrates current and historical data from different sources.
- time-variant: DW keeps historical data of different time.
- non-volatile collection of data: content of DW should not be changed. It is historical data.

Unlike the Entity Relationship Model, which is used to construct standard databases, data warehousing is constructed using dimensional modeling techniques[3]. Modeling data warehousing is a difficult task. It necessitates familiarity with market methods, Understanding the structural and behavioral system's conceptual model and understanding of data warehousing techniques[15]. The dimensional modeling technique organizes all the data into 2 types of tables – fact table and dimension tables (Fig. 3). This technique makes the process of retrieving data from data warehouse easier and faster[4]. According to the representation of the fact table and the dimension tables, there are 3 types of architectures in dimensional model: (1) star schema, (2) snowflake schema, and (3) galaxy schema[16].

The architecture of data warehousing is incomplete without meta data. The data warehouse is defined by data. It's used to create, maintain, and instruct users about how to use a data warehouse. Any data mining process relies on this information[17].

DATA MINING

Data Mining (DM) is a database and artificial intelligence technique for extracting valuable information from large databases and assisting users in making smarter decisions. It's commonly used as a decision-making aid[5].

(A) Data Mining Usage

With such a large amount of data, humans find it impossible to interpret and extract valuable information. As a result, the use of Data Mining techniques is important. DM is used in a variety of fields to aid in

the extraction of valuable data and the subsequent making of better decisions. For e.g., DM may be used to promote a product.

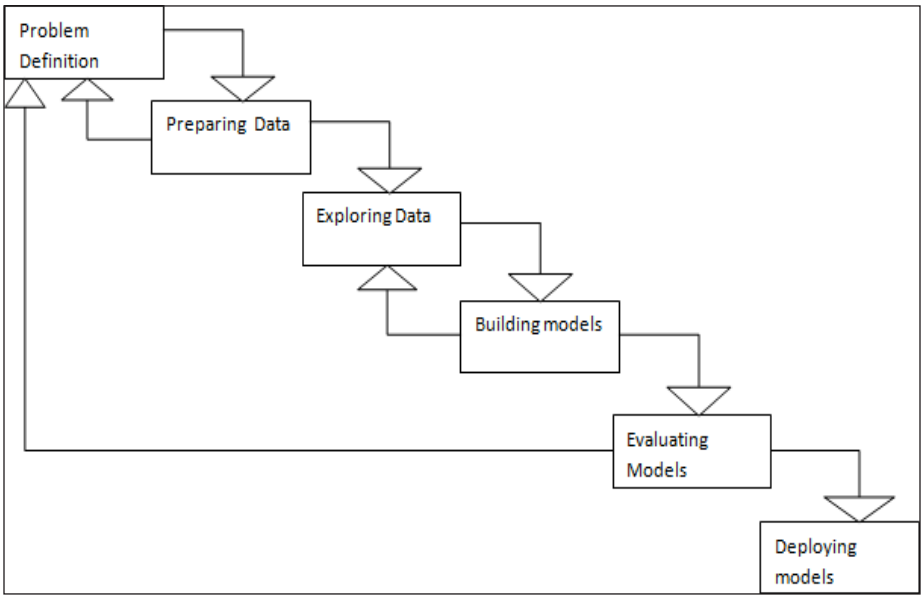

Fig. 1: Data Mining Usage

It may assist by providing valuable knowledge on the right media and timing to post a commercial in order to boost a product's sales. DM strategies (for example, correlation analysis) examine all past relevant campaign data and compare sales to generate insightful reports for policy makers to use in order to boost potential sales[18]. Web mining[19] is the most general use of data mining. With terabytes of data being added to the internet every day, a better way to evaluate web pages and collect valuable information is needed[6].

(B) Data Mining Process

The method of data mining is not easy. It is difficult to understand and has feedback loops, making it an iterative operation. The phases of the data mining method are depicted in Fig. 1[7]. It also demonstrates that the procedures can be reversed and that the whole procedure can often be restarted from the beginning. In fact, there are six steps to the data mining process:

- ❖ Problem definition
- ❖ Data Preparation
- ❖ Data Exploration
- ❖ Modeling
- ❖ Evaluation
- ❖ Deployment

(C) Explanation/Discussion of Model

Data Mining Process – Goal: Data mining is a method of querying data that is different from traditional querying. Instead of doing standard database searches, data mining goes a little forward by collecting more valuable information from a large number of datasets. It is not a simple procedure. It entails multiple stages of suggestions between them, and the whole process can be replicated from the beginning in order to have clearer responses that match the needs of the decision makers[20].

Problem Definition: Understanding the market challenge is the first step in the data mining process. The project goals and company criteria should be identified by the working team at this point[10]. This move can also describe the model parameters that will be used to test the model. The team should come up with a description of the data mining dilemma together[21].

Data Exploration: Domain experts gather, identify, and explore data in this process, as well as communicate with data mining and industry experts from the previous phase, in order to fully comprehend the metadata[10]. Understanding the data allows you to get a better understanding of the market, which will help you design the mining model[22].

Data Preparation: Domain experts should create data models and correct any data-related issues before developing mining models. Before going on to the next step, all bad data should be discarded, and any missing data should be brought in[23]. The final dataset should be prepared during this process[10].

Modeling: Data mining experts will begin the modeling process after completing the data discovery and planning phases by choosing modeling techniques and specifying the columns of data required to construct a mining framework, followed by the mining models[10]. The built model should be able to satisfy the requirements[24].

Evaluation: In this step, data mining experts create several models, test them, and then choose the best one. Until proceeding to the implementation process, the chosen model should be thoroughly tested before being deployed into the production area. If no model performs as predicted, the whole procedure, which is the problem description[10], should be replicated from the beginning.

Deployment: The optimal model will then be deployed into the manufacturing system after it has been evaluated. This process could result in the development of a data mining study[10]. Data mining activities can be completed after the deployment. Prediction activities, for example, will assist businesses in making informed choices. The paradigm that is implemented should bring value to the company[25].

RESEARCH OUTCOME

The developed model contributes to a better understanding of the data mining process by assisting in the identification of the process's key factors (steps). It also demonstrates how difficult it is. It is critical to have correct outcomes by reviews and iterations. This knowledge aids in the implementation of a real-world data warehouse and the subsequent application of data mining techniques. For the installation and observations, I choose a system with a large volume of data. The chosen system is a University Housing System which usually has huge amount of data like data related to objects as buildings, apartments, furniture, students, and many other things like maintenance requests. First, I implemented the operational database for the system. Then we built the data warehouse to apply the data mining techniques.

(A) Operational Database

Brief: Since the data for university housing must be kept in a finite order, it is preferable to use the assistance offered by a database, which makes working with the data very straightforward. Using a website for the Housing Department also serves to streamline the process. Fixing any malfunction or injury that could occur is one of the most important services that the Housing Department can provide in a timely manner. As a result, creating a database will help to speed up the process. Students may submit questions by filling out an electronic form with the information needed to be stored in the database. The requested details would then be sent to the employee who is in charge of receiving student maintenance requests. When the staff have completed their jobs, the database should be checked with all details relating to the reconstruction operation.

Functions of the System: The main functions that are done by this system are:

- ❖ Students can make requests in order to get their items repaired.
- ❖ The requests given by the students will go to an employee who is responsible to receive the requests.
- ❖ Student can know the price for the repair work done and the status of his/her repair request.
- ❖ Employees can list all the new requests.
- ❖ Employees can show the status of the repairing process.
- ❖ Employees can update the repairing information.

Design

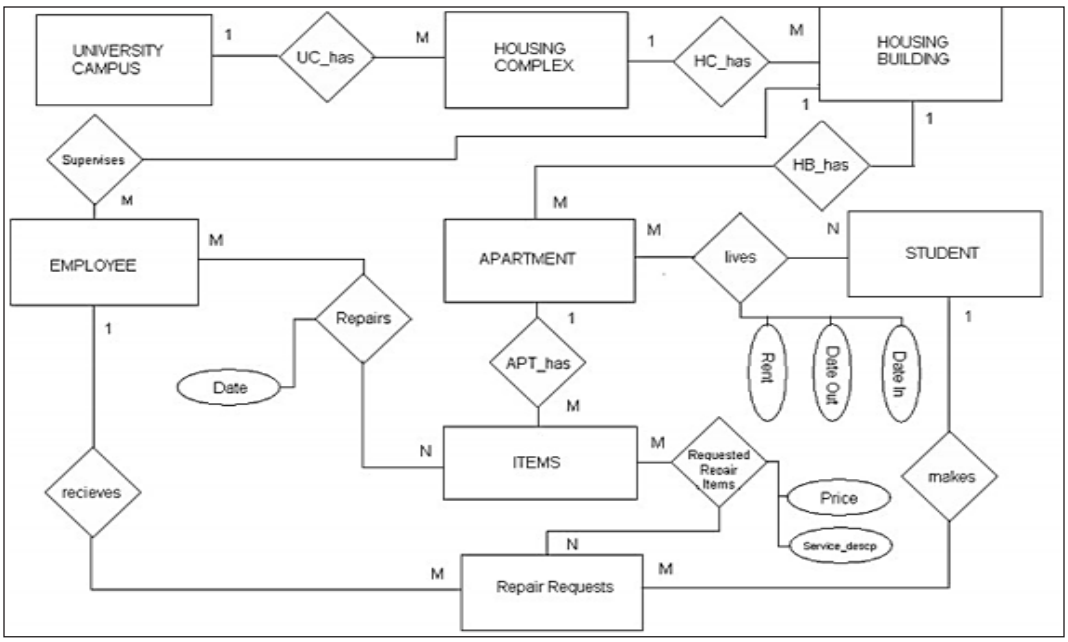

Fig. 2: Developed Entity-Relationship Diagram for the proposed Operational Database (ERD)

(B) Data Warehouse and Data Mining

Brief: The "repair Request Information" is the focus of this Data Warehouse. It is focused on demands for repairs submitted by students who reside in one of the university's housing buildings. This data warehouse stores the information provided by students asking to repair damaged items. Information includes:

- Students' information.
- Item's information.
- Location (campus name, complex name, building #, and apartment#).
- Request information.
- Employees Information who are going to take care of the damaged item.

Proposed results: Managers would be able to make smarter choices with the aid of data warehousing and data processing techniques. This method provides answers to a few key questions that will assist managers in making informed decisions.

For example:

- What items are requested to be repaired together?
- What items are requested to be repaired after each other?
- What are the most common things on which students submit maintenance requests?
- What were the most requested items to be repaired last-year?
- What was the name of the building with the most maintenance demands last month?

Design

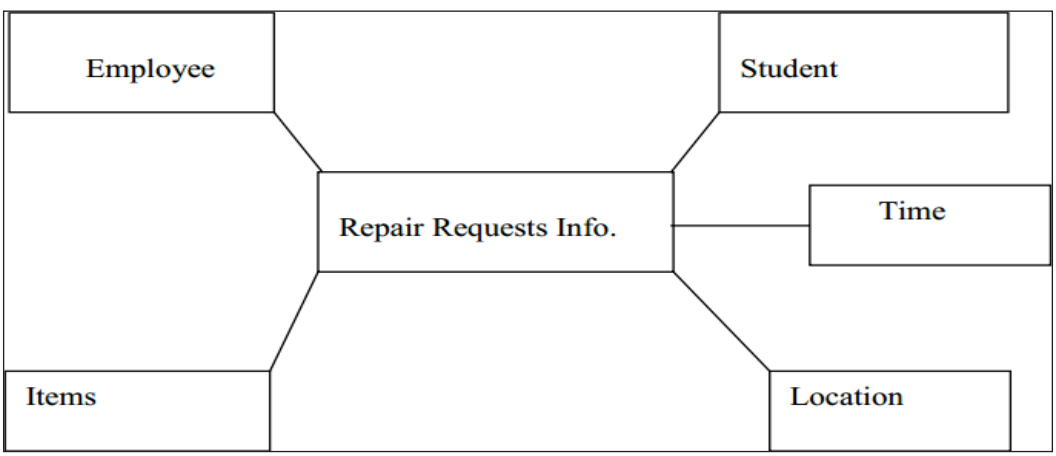

Fig. 3: Developed Star Schema for the proposed Data warehouse

CONCLUSION

We now have an immense number of records, necessitating the use of data warehousing and data mining. A data warehouse (also known as an operational database) is a central repository for a subject-oriented, integrated, time-variant, and non-volatile compilation of data from various sources[1]. Data warehousing organizes data in two separate architectures for faster performance: fact table and dimension tables[4]. As a result, modeling a data warehouse differs from modeling an operational database. A dimensional modeling is used to model the data warehouse (star schema, snowflake schema, or galaxy schema) but the operational database uses entity relationships diagram[3]. Data mining has evolved into a critical method for extracting valuable knowledge from the massive amounts of data available today. It may also aid in the extraction of data from the Internet, which has become an integral part of our lives. It involves six phases:

- Problem definition
- Data Preparation
- Data Exploration
- Modeling
- Evaluation
- Deployment[7].

It's an iterative process with feedback in between stages and the need to restart the process from the beginning on occasion. Iterations are needed during the mining process in order to provide clearer answers that users can use to make better decisions. The ability to automate data mining techniques and the added benefit of using them make it appealing for use in a variety of fields, especially science and business with large amounts of data[8]. Data mining is a sophisticated method of processing and querying data. It goes much further by detecting useful relationships in records, like secret trends or relationships[26]. One of the most popular applications of data mining is web mining[6]. It aids in the extraction of valuable knowledge from the vast amount of data available on the internet.

REFERENCES

1. Chen, Y. and L.-l. Qu. 2007. The Research of Universal Data Mining Model SYSTEM BASED on Logistics Data Warehouse and Application. In Management Science and Engineering, ICMSE 2007. International Conference on 2007.

2. Viqarunnisa, P., Laksmiwati, H. and Azizah, F.N. 2011. Generic data model pattern for data warehouse in Electrical Engineering and Informatics (ICEEI), 2011 International Conference on. 2011.

3. ElDahshan, K.A. and Lala, H.M.S. 2010. Mining uncertain data warehouse in Internet Technology and Secured Transactions (ICITST), International Conference for 2010.

4. Nimmagadda, S.L. and Dreher, H. 2009. On designing multidimensional oil and gas business data structures for effective data warehousing and mining. in Digital Ecosystems and Technologies, DEST '09. 3rd IEEE International Conference on 2009.

5. Trifan, M. *et al*. 2008. An ontology-based approach to intelligent data mining for environmental virtual warehouses of sensor data in Virtual Environments, Human-Computer Interfaces and Measurement Systems, VECIMS 2008. IEEE Conference on 2008.

6. Dongkwon, J. and Songchun, M. 2001. Scalable Web mining architecture for backward induction in data warehouse environment in TENCON 2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology.

7. Bora, S. 2011. Data mining and ware housing in Electronics Computer Technology (ICECT), 3rd International Conference on. 2011. IEEE.

8. Yi, L. and Yongjun, P. 2012. Application of Digital Content Management System Based on Data Warehouse and Data Mining Technology in Computational Intelligence and Communication Networks (CICN), Fourth International Conference on 2012.

9. LePine, J.A. and Wilcox-King, A. 2010. Editors' comments: Developing Novel Theoretical Insight from Reviews of Existing Theory and Research. *Academy of Management Review,* **35**(4): 506-509.

10. Huifang, Z. and Ding, P. 2010. A knowledge discovery and data mining process model in E-marketing in Intelligent Control and Automation (WCICA), 8th World Congress on 2010.

11. Zhen, L. and Minyi, G. 2001. A proposal of integrating data mining and online analytical processing in data warehouse in Info-tech and Info-net, Proceedings. ICII 2001 - Beijing. 2001 International Conferences on 2001.

12. Nimmagadda, S.L. *et al*. 2010. On new emerging concepts of modeling petroleum digital ecosystems by multidimensional data warehousing and mining approaches in Digital Ecosystems and Technologies (DEST), 4th IEEE International Conference on 2010.

13. Krippendorf, M. and Il-Yeol, S. 1997. The translation of star schema into entity-relationship diagrams in Database and Expert Systems Applications, Proceedings Eighth International Workshop on 1997.

14. Eisenhardt, K.M. 1989. Building Theories from Case Study Research. *The Academy of Management Review,* **14**(4): 532-550.

15. Usman, M. and Pears, R. 2010. A methodology for integrating and exploiting data mining techniques in the design of data warehouses in Advanced Information Management and Service (IMS), 6th International Conference on 2010.

16. Sung Ho, H. and Sang-Chan, P. 1998. Data modeling for improving performance of data mart in Engineering and Technology Management, Pioneering New Technologies: Management Issues and Challenges in the Third Millennium. IEMC '98 Proceedings. International Conference on 1998.

17. Yuekun, M. *et al*. 2010. Implementation of Metadata Warehouse Used in a Distributed Data Mining Tool in Challenges in Environmental Science and Computer Engineering (CESCE), International Conference on 2010.

18. Yun, Z. and Weihua, L. 2012. AHP Construct Mining Component strategy applied for data mining process in Information Science and Technology (ICIST), International Conference on 2012.

19. Xujuan, Z. *et al*. 2007. Using Information Filtering in Web Data Mining Process in Web Intelligence, IEEE/WIC/ACM International Conference on 2007.

20. Bianchi-Berthouze, N. and Hayashi, T. 2002. Subjective interpretation of complex data: requirements for supporting the mining process in Systems, Man and Cybernetics, IEEE International Conference on 2002.

21. Chieh-Yuan, T. and Min-Hong, T. 2005. A dynamic Web service-based data mining process system in Computer and Information Technology, CIT 2005. The Fifth International Conference on 2005.

22. Ding, P. 2009. A formal framework for Data Mining process model in Computational Intelligence and Industrial Applications, PACIIA 2009. Asia-Pacific Conference on 2009.

23. Tsumoto, S. *et al*. 2012. Exploratory temporal data mining process in hospital information systems in Cognitive Informatics & Cognitive Computing (ICCI*CC), IEEE 11th International Conference on 2012.

24. Ding, P. 2010. Data mining process model for marketing and CRM in Machine Learning and Cybernetics (ICMLC), International Conference on 2010.

25. Gang, K. and Yi, P. 2008. A Standard Process for Data Mining Based Software Debugging in Networked Computing and Advanced Information Management, NCM '08. Fourth International Conference on 2008.

26. Brohman, M.K. 2006. Knowledge Creation Opportunities in the Data Mining Process in System Sciences, HICSS '06. Proceedings of the 39th Annual Hawaii International Conference on 2006.