A REVIEW OF DATA WAREHOUSES MULTIDIMENSIONAL MODEL AND DATA MINING

G. Sekhar Reddy [1], Dr. Ch. Suneetha [2]

[1] *Research scholar, Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, India*
golamari.sekhar@gmail.com

[2] *Associate Professor, Department of Computer Applications, RVR&JC college of Engineering, Guntur, India*
suneethachittineni@gmail.com

Abstract: Now days, the world technologies are an increasing development and observed in the usage of computer networking, social media and Internet of things (IoT). Research experiments generate large amounts of data that need to be collected, analyzed and managed. Researchers have found an increase in data containing useful and ineffective entities. The data warehouse finds difficulties in extracting useful information and increasing the amount of data generated. The data warehouse analysis developed as promising area in research which supports business organization like decision making. Also data mining supports information detection by determining associations, hidden patterns, building analytical models, and performing prediction and classification. Hence, this article gives the comprehensive review of data warehouse techniques, data warehouse demerits, and data mining approaches. Furthermore, it also explains the data warehouse appeared in decision making perspective and research problems in data warehouse and data mining

***Keywords*: Data warehouse, data mining, UML schema, conceptual model, and decision tree algorithm.**

1. Introduction

In 1990s period and ago, as a result of competitive world, the data's are needed to be analyzed to support the process of decision making. Traditional transaction databases do not meet the necessities for the analysis of data since which are intended to support day-to-day operations. But the historical data's are not included and not optimized to accomplish complex queries that contain large quantity of data. So as to overcome the limitations and issues the data warehouse were proposed. The data warehouse (DW) offers an arrangement which permits users to obtain accurate and effective complex questions. The various tools and systems are utilized to analyze and access stored data in data warehouses [1]. The DW is a collection of data based on subject, integrated, non-volatile and time-variant data to support their decisions management. For decision making, a DW design tool is to manage accurate time, logical information and deliver entire. The corresponding accepted information feature is most effective aspect that can bring substantial benefits to some organization. Data Warehouse is a computer-based data information system that enhances reporting tools and queries in database as its ability to evaluate data from different databases and in interesting methods [2].

Some of the interesting methods are top-down method, bottom-up method, hybrid method and federated. The DW is moving towards requirement demanding by means of huge data space with the huge information and its historical data. This leads to the outline of a slowly changing dimensional system. Hence, Kimball's suggested the technique of dimensional modeling, which stores the data in facts and dimension forms in Multi-dimensional Store (MD) [3]. It also develops the conceptual design for multi-dimensional systems, which outlines relevant issues related to the initial requirements of users, data sources availability, behavior of system and database programs. However, the design project in the MD model failed to address the necessary information, thus establishing poor communication among decision makers and DW developers.

The requirement of DW has grown significantly as a result of organizations distributes control from the middle management layer which has distributing and controlling business information. As users become more dependent on information obtained from information technology systems, the need to provide an information warehouse for the use of the remaining staff becomes increasingly important [4]. There are many technical reasons for having a data warehouse. Initially, the DW is intended to discourse the inconsistency of information and functional exchange systems. For these information systems, the two classes are intended to satisfy various inconsistent requirements. Simultaneously, information technology infrastructure is varying quickly and its abilities are increasing. To attain the higher performance, DW is associated with database system which supports Very Large Databases (VLDB). DW is an environment not a product and the architectural information construction are difficult to access in data stores traditional function [5].

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Every system or organization is played with good data which is normally stored and maintained but not able to find the valued, often earlier unidentified information data are hidden and to prevents this data from being converted to wisdom or knowledge. So as to satisfy these needs, some of the steps are to be used as i) Capture and integrate internal and external data for integrated data information into a comprehensive vision of 'mine'. ii) Organize and provide info and facts in ways that speed up the most complex function of decision making [6]. The DW application is classified into three types like information processing, analytical processing and data mining. In information processing, DW allows processing of information stored on it. For the processing of information, it is processed by reporting via tables, crosstabs graphs, querying and statistical analysis. In the analytical processing, DW supports the analytical processing of stored information on it. Furthermore, the data is analyzed by operations of OLAP, with drill down, pivoting and drill up. Then the data mining supports information detection by determining associations, hidden patterns, building analytical models, and performing prediction and classification.

The data mining is composed with DW which is essential to form historical information collected from the application of large-scale server. On the other hand, in various sectors data mining able to add the values into the organization information assets, via the data warehouses effective induction into client server [7]. Hence, on the large databases data mining advanced effective techniques in business environment is one of the interesting research areas. Generally DM is new and promising topology which is well-defined as the process of inventing new topologies and developments by mining high amount of data warehoused with data visualization, artificial intelligence, statistical, and machine learning techniques. The data mining application can be used in the fields of manufacturing, medical, chemical, aerospace industries and etc. Here, the knowledgeable new techniques are needs to support their decision making tactics [8]. For the new techniques enable the data predictive patterns, hypothesis testing and creation and visualizations. From the databases the useful information were extracted that which helps the data mining end users. In DW, the large databases are existent which means Data Mountain. In data mining, unknown and potentially useful information are obtained from data mountain because of DM is non-trivial mining of implicit [9]. But any of the industry, the DM is not specific because this necessitates intellectual techniques and to discover hidden knowledge that resides in the data.

The DW is logically designed as set of increasing designed data marts and every data marts has representing the business process view. The dimensional single model consists of dimension tables and fact tables. The bus configuration of DW attaches all the data marts to a logical DW. It is proficient by using confirmed facts and confirmed dimensions, which confirms various data marts grain is compatible. In recent times, some of the attempts are to modeling the data warehouse with Unified modeling Language (UML). The UML is the standard OMG modeling language which is mostly used in modeling of object-oriented. This states the general vocabulary for communication between users and designers. Here, many of the UML based database modeling is discussed and concentrating more on databases rather than the DW. OMG defines a typical warehouse meta-model that is utilized to guide modeling of general warehouse and it has inadequate meta-model for DW design. UML references are semantic in database models it is complex. For database groups, the techniques of traditional DW are straightforward and thought about database system people reveals the way. On the other hand, traditional codes ensure slight semantics and are data-based; they are not enough to deal with a complex data warehouse model. It was proposed to use conceptual design to implement the UML MD model to overcome the failure model structure, where the term conceptual modeling includes Entity / relationship based models.

The organization of the paper is described as follows: introduction about the research paper is presented in section 1, survey of recent existing methods and the challenges of existing methods are depicted in section 2 and the data mining and data warehousing techniques are surveys in section 3 and the conclusion of the article is presented in section 4.

2. Related works

Some of the information contents updating may be violates certain restrictions imposed on the schema. The classical technique to deal with this problem is to reject the requested update when its use leads to some restriction violations. An alternative approach aimed at automatically calculating the updates repairs, when used in conjunction with the requested update, the minimum additional changes that bring the database to a new level where all limitations are satisfied. Hence, Xavier Oriol *et al* [10] were proposed independent of language to define logical formalization based schema and constraints. Also, the UML and OCL (Object Constraint Language) had been applied mostly in the modeling of conceptual. It can be utilized to maintain the stability of information base after the application of certain updates and to deal with the non-performing functions problems. The OCL fragment was used to describe the restraints have the same relational algebra and detects its subgroup that provides some good properties in the repair-calculation process. Recently, conceptual schema validation and verification have the attracted tools to automate this process. In general, this is achieved by evaluating whether the schema satisfies the various types of desirable characteristics that ensure the schema is correct. Therefore, AuRUS topology has designed to analyze conceptual schema of UML/OCL to elaborate their correctness. If a property is satisfied, AuRUS provides instant

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

information on the model of the schema screening the specific circumstances in which the property has hold. In its absence, AuRUS provides an explanation for such dissatisfaction, which means a set of integrity constraints that conflict with assets.

Then the relational database schema of UML profile specification was investigated by Igor Tomic et al [11]. The specification of profile was based on the UML class map extension, which uses the functional signature to represent specific relevant database models like indices, single & multi column foreign and primary keys, etc. The suggested approach benefits are Minimize homogeneous additional attributes representing keys, database constraints better visualization and efficient forward database engineering. For the profile implementation, the platform was plug-in open source; it permits the manual modeling along with automatic relational database scheme with defined profile. It was an example of visualization. The information qualities were improved by the building of DW in order to attain particular business objectives as enhanced decision making and competitive benefits. Though, there is no systematic way to obtain a multi-dimensional scheme from a heterogeneous database recognized as standard by OMG and field professionals. So, for the design of DW Amine et al [12] was proposed a Model-Driven Approach (MDA). To use the MDA approach in DW construction method, a multi-dimensional meta-model and UML meta-model set of transformation described, which is fitted with a multi-dimensional meta-model. The language of Query View Transformation (QVT) was utilized to program the transformations rules. Additionally UML technique was build the DW to automate the method.

On the other hand, the evolution of medical data size should be modeled by DW. The medical database is a specific database aimed at decision support. The different medical databases were taken from other data sources and transferred into new arrangements, which makes the decision making process better. Thus, it is necessary to develop a systematic description of the medical database and DW models, the conceptual of all data generated in medical and healthcare institutions. In medical filed, classical conceptual modeling was not incorporating hence Mouhamed Gaith et al [13] had proposed UML profile based framework as medical profile. By using the UML extension algorithms, the processing of medical data was standardizing by the suggested medical profile. This model was linked with MDA, allows to defining and clarifying concepts related to the medical field, as well as annotation process of medical image.

In modern driven architecture, platform independent model (PIM) and platform specific model (PSM) were needed for the design of Computational Independent Model (CIM). For the representation of PSM and PIM methods, the suggested design can be explained through the mature technologies. Also, there is no standard for creating CIM. In this framework, data warehouse developments are unsuccessful because not much consideration is paid to the demand analysis stage. So as to overwhelmed these issues, El Beggar et al [14] were proposed CIM architecture for modeling data warehouse necessities by UML profile. It can extend the business process models and cases. From the methodologies, the performance of DW was ensured by hybrid approaches as MD schema to with adherent and data sources to user goals. Hence, to enhance the quality of schema and required efforts provided multidimensional schema to design the automation processing methodology. For the evaluation, hybrid method based case study was considered in DW schema design.

The development of temporal databases modeling approaches to facilitate the objects of temporal. Although UML is not presented to succeed this task, extended extension time and UML providing inadequate ways to express temporary constraints to variable data and the temporal objects dynamic behavior. For that reason, Soumiya Ain et al [15] had been offered a UML extension developed by features of OCL and understandability was improved by Bi-temporal dimension of UML/OCL model. The ability was always supports their attributes of temporal and its evaluation. The proposed Bi-temporal data was translating the temporal object-Relational database. One of the algorithms was suggested to transform from a conceptual schema enriched with bi-temporal features into a temporal object-related database model using subsequent various steps like classification, information extraction, attributes and relationships. In general, many techniques are interested in a specific feature of DW as ETL processing, storage, reporting, OLAP analysis and do not obscure its entire life cycle. Alternatively, MDA was support the software manufacturing and by promoting transformations and models among them up to code generation. So as to reducing the cost of software development, El Beggar et al [16] were proposed MDA-oriented UML profiles and described how to supply MDA in the whole development. Then started through eliciting the requirement and also lowers the multi-dimensional conceptual model.

Based on fuzzy inference method, Naveen Dahiya et al [17] had been investigated the DW conceptual model for ranking the metrics. The fuzzy based approach provides an accurate ranking system because of the ability to handle the inaccurate data involved in the ranking of measurements and the ambiguity in the expert decision-making process. However, the proposed work quality metrics validated through Manuel Serrano with particular identified parameters and estimation of criteria matrix utilized the permanent operation. In the industry analysis, the data of IoT with technologies of Business Intelligence (BI) has been developed as high importance. But applications of BI are more difficult for many of the reasons. Therefore, introduced conceptual model based UML profiles and MDA for BI application implementation and modeling.

Furthermore, it could help in the implementation process of the IoT subsystem through automatic code generation.

3. Data Warehouse Characteristics

The DW characteristics are classified into four types as Subject-Oriented Data, Integrated Data, Time Variant Data and Non-volatile data which are explained below.

Subject-Oriented Data

A DW is organized around key subjects as product, sales, supplier and customer. Instead of focusing on an organizations day-to-day transaction processing operation, a DW focuses on analyzing and modeling of data for decision makers.

Integrated Data

A data warehouse is fashioned using integrating data from varied, information systems heterogeneous databases like relational flat files and databases. For example, the male and female is represented by 0 & 1, M & F, or true & false. Generally, the irregularities are subtle and complex but the DW data is maintained by always consistent fashion.

Time Variant Data

In contrast to functional databases, historical data is of paramount importance in the DW world. The data's may be available at daily, weekly, monthly, quarterly or year collections. The time variant therefore calls for saving multiple copies of the basic details of different time and time frames. The strategy of time variant is necessary not only for performance, however maintaining the consistency of organizational units and contractions reported over time.

Non-volatile Data

The final primary feature of data warehouses is that after loading into a data warehouse, changes, insertions or deletions are no longer made. The data warehouse is then reloaded or added periodic based at night, weekly or monthly intervals with new, modified or compressed data. Apart from this loading process, the information in the DW is static. It is required to mark data at any time. If the data array is updated, the information will be erased. Maintaining institutional memory refers to DW which is the most important features, apart from the property of non-volatility also allows the database to be highly optimized for query processing. In addition to these four principals laid out a number of other things that went into DW Inmon's definition. Some of these policies were initially controversial, but are now generally accepted. However, in addition to these principles directly raised the function of DW in Inmon's definition, as well as the support for management decisions. The DW can be a treasure trove of information for companies, where employees can access data, along with accurate, timely and relevant information about products, consumers, products and technological enlargements. Creating and maintaining a valuable database is a major challenge for companies. If there is too much data in the warehouse, but the data is not kept relevant and accurate, the value of the warehouse is not sufficient if the tools used to access the data warehouse are not properly connected.

3.1 Architecture of a data warehouse

A DW system consists of two main architectures as the architecture of system and architecture of data flow. Architecture of data flow is describes how the data stores are organized within a database and how data flows from the source system to users via these databases. In system architecture, servers, software, network, storage and clients are the physical configurations. In a data warehouse, the configuration of data stores is the architecture of data flow within a database system, by arrangements for how users can use the data from source systems for end-users over these data stores. This contains how control the data flows monitored and logged along with the mechanism for ensuring data stores data quality. It is also differs from the architecture of data. The architecture of data is almost how data is organized in every data store and how designed a data stores to reflect process of business is also known as data modeling. In architecture of data flow, the important component is data stores. A database is a set of files containing one or more databases or DW data, organized into a specific format and involved in process of DW. The DW data stores are categorized into three types based on user accessibility.

❖ The data store of user-facing data store is accessible to end-users and is demanded by end-user applications and end-user.

❖ An internal data store is used locally for integration, cleaning, recording, and data processing by data warehouse components, and is not open for querying end-users and end user applications.

❖ A hybrid database is used for querying internal database algorithms and end-users and applications of end-user.

The three-level data warehouse structure allows the conceptual view (objective view) to be clearly separated from the user view and the storage view system. A data warehouse can be flexible and adaptable to separate the three views of data, and this flexibility and adaptability is data freedom. Then three types of DW views are user view, conceptual view and storage view which are explained in below.

User View

In the DW architecture, the user view is the top view among these three views. In this view, DW is only available for restricted portion to the end user. Since the DW is a shared resource, each user has a view of the real world as specified in the format known to him. The user view is top view which consists of analysis of multi-dimensional tools, data mining tools and query & reporting tools. For various users, creating a separate view of database is to ensure the security of the DW. When access is limited to data required by different users, access to the entire database is limited, reducing the risk of security breakdown.

Conceptual view

In the three views, this was the middle view of the architecture which is materialized view and stores the summarized and aggregated data. It does not comprise details of any storage status its independent of software and hardware. Software independent refers to does not depend on the database used to operate the database. Hardware independent refers to does not depend on the DW hardware. Thus the change in hardware and software ideologically does not affect the conceptual view of data warehouse.

Storage View

This is the bottom view of the architecture which is the actual physical storage of data. It defines how data is physically stored in a data warehouse. It also says the normal operation of the DW to achieve the optimal performance and storage space utilization required for effective DW.

3.2 Data Warehouse components

The data warehouse is consider as six types of components as data sources, data extraction tools of transformation, central repository, data modeling, front end tools and target DB.

3.2.3 Data warehouse design

The development life cycle of data warehouse includes some of the phases as i) requirements gathering and identification ii) dimensional model design iii) testing and iii) maintenance. The design phase is the utmost vital stage in the development life cycle of DW to create an effective database. Also the design of DW can be categorized into three types as

i) Conceptual design
ii) Logical design
iii) Physical design

i) Conceptual design

This is the complete building foundation and well-documented database based on specifications of user. This will also produce the expressive conceptual schema for the system. Conceptual design is aims to describe an attributes mapping for the ETL process from the data source to the objects of data warehouse. Commonly, it is based on the dimensional modeling concept which contains many number of categories or hierarchies and set of measures or facts for decision making. In DW conceptual design, fact constellation schema, snowflake schema and star schema are utilized. The dimensional fact model of conceptual based model was suggested by Rizzi et al [18] in data warehouse. The pattern based ETL modeling of conceptual design is using business-process modeling language (BPMN) and it was developed by Oliveira and Belo [19].

ii) Logical design

The development of conceptual design is logical design. At this point, conceptual model based implementation-oriented logical schema is established, which is particular to the selected tool of DW beneath a specific constraint.

iii) Physical design

This design is deals with concerns interrelated to implementing tools on behalf of data structures such as scheduling, allocation, and use of data marts.

3.3 Multidimensional Models

Multidimensional models are intended to describe the facts of the data warehouse and the different analytical dimensions. Lately, various articles have offered the multidimensional specific approaches for formalizing. Based on facts and dimensions, the information is structured in MD modeling. A fact is an object of interest to an organization, and it is described by the properties of cells or points in a data cube or by attributes called as fact attributes or measures. These measures are based on a set of dimensions that decide the granularity to represent the facts. Alternatively, dimensions afford the context in which facts should be analyzed. Furthermore, the dimension is characterized through attributes, so it is called as dimension attributes. Multi-dimensional data models within data analysis have three main application areas. Initially, the multidimensional models are utilized in data warehousing. In short, a data warehouse is a large repository of integrated data obtained from multiple sources in an organization for the specific purpose of data analysis. In general, multi-dimensional modeling is used for such data because it provides good support for data analysis. For the design of DW, MDA approach was presented by Amine Azzaoui et al [12]. To use the MDA approach to the construction process of DW, a multi-dimensional meta-model and set of transformations from the UML meta-model is described, which is mapped with a multi-dimensional meta-model.

For the modeling of MD, the general methods are requirement-driven, data-driven and next hybrid approach is introduced by advantages of both systems. In a demand (requirement)-driven approach, conceptual design is based on the needs of decision maker's information, while data sources are considered separately during ETL operations. In contrast, data-driven method is based on data sources detailed analysis and required information of decision makers are consider after the conceptual design implementation. In data warehousing, Maribel [20] has address some of the investigation challenges as the data sources and new business process integration, the good way to attain this integration, complex data management & performance enhancement, analytical functions automation, flexible and highly customizable visualization of their data provides an progressive decision making environment.

3.3.1 Data-driven approach

The data-driven method is designated in [21] starts with end-users that detect dimensions and facts to define an initial load. In the method of data-driven, user-requirements are informally and faintly characterized by language of natural and then it focused data source analysis. Before the conceptual design step, the most of the steps should be executed like integration and normalization of schemas. The conceptual design model is

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

centered on Dimensional Fact Model (DFM) and this model is created by dimensions and cubes and starts from the schema of relational (E/R). The released requirements do not refer the designer constraints, but only one type of recommendation to be utilized in the data modeling phase. Hence, conceptual design mostly depends on designer ability and experience. The method for conceptual modeling is semiautomatic, which is based on a method of creating an attribute tree which characterizes the integrated data source. The tree root is entity that arbitrarily selected the fact through designer. After, algorithms create a node for each attribute encountered when going back and forth through entities along relationships.

3.3.2 Requirement-driven approach

This method is grounded on the i^* framework, the information system functions are modeled by this method at engineering stage. The i^* framework is describe the relationships and tasks that exist between involved agents in database surroundings like multiple decision-makers and data warehouses, as it quickly emerges as an efficient goal-based and task-independent system. Hence to execute deep domain analysis, the i^* frameworks permits the designers and generate the formal model in decisional environs. Based on user needs, the conceptual schema is created by this framework. For this purpose, the required information from decision-makers should be transformed into multi-dimensional components as measures, facts, and dimensions, etc. specified along with the UML for the database. In fact, a great deal of effort is currently being put into expanding UML with profiles cover each aspect of the data warehouse life cycle. The [22] is an important piece of work to model how the database can be accessed by end users. For this framework, user needs are exploded to many detailed hierarchy of goals as strategic goals (objectives to reached organization), decision goals (to answer how the goals are satisfied), and information goals (to define need of information in decision making).

As a result of the complexity and variability nature of Supply Chains (SCs), companies need solutions that allow Big Data Analytics to integrate their large data sets, ensuring that efficient actions are taken rather than reactive actions. For the big data analysis, the methodology of data requirements elicitation was applied by António A.C et al [23]. It integrates the user-driven, goal-driven & data-driven methods in big data warehouse, data requirements elicitation to completing these different organizational scenarios in identifying relevant data to support the decision making process. The dynamic user demand information over demand assessment and prediction system was presented by the Yahui Wang et al [24]. This author proposed the concept of user requirements-oriented knowledge which was based on a four-level hierarchical graph model specializing in knowledge collaboration and communication.

3.3.3 Requirement analysis in DWs

A demand analysis level permits designers to create a DW that encounters the requirements of the companies, accordingly maximizing the DW tasks. For specific requirements, DW domain introduces some specified methods which are called as specialized methods. The major disadvantage is the inherent difficulty in understanding and referring to decision procedures, therefore huge gap among those who are authorities in the field and user needs on the decision makers and experts in design and DW constructions satisfies the requirements which means DW developers.

High-level abstraction is provided by conceptual model and purposes at attaining the deployment difficulties. So we are proposed a conceptual model based UML. The standard modeling language is UML and it was sustained by way of various tools. This is the main benefit of reusing and adapting existing techniques. UML profile is a set of methods and this technique for trying to convert UML to a particular application.

The UML profile is domain-specific concepts on technical view this is set of stereotypes. Since the suggested structure is based on merging the DW and temporal dimensional concepts. It designates basic concepts of MD as constellation which encompasses the meta-class package. The three types of classes are composed in a constellation namely fact, dimension and level attributes. For the dimension, it can be related one or other level attribute which means specific type association named as Hierarchy. Furthermore, the same hierarchical "level attributes" are associated with a particular compound named as "rolls up". The class of UML is made up of a set of dynamic and static functions. The static properties state the fact and dimension properties in DW context.

3.3.4 Querying temporal data

Temporal data warehousing is one of the development model which is not capable of manipulating time effectively without proper query language. In standard, a temporal query can be formatted straight in a related program by standard SQL; however this can be long and complicated even for a talented user.

3.3.5 Designing Temporal Data Warehouses

Temporal DW is widely accepted that designing a DW system necessitates completely different techniques from the generally accepted techniques for designing functional databases. Also the designation problem is time. So Rizzi et al [18] stated by developing design methods skillful of taking into account time which is one of the undefended issues of DW research. Pedersen and Jensen [25] acknowledge that manipulating time is essential for multi-dimensional models. Further time dimensions significant are constantly esteemed and constantly growing, and related with multiple user-defined calendars. Also proposed temporal data warehouses design methodology featuring logical design which is characterized by a temporal validity and its design discourses the effective storage and access.

The temporal hierarchies play within OLAP queries and data marts; it is worthwhile to follow temporal approaches to modeling them not only logically but also from conceptual view. However, the conceptual models for data mart permits the temporary steps to be characterized in the same way as the other steps, which is to provide ad-hoc ideas for modeling time to provide the best approach to our knowledge. It is based on multidimensional conceptual model extension suggested by Malinowski & Zimányi [26]. The various kinds are permitted temporality as namely, transaction, loading time, valid time, lifespan, and temporal support for hierarchies, properties, levels and measures are granted. To end with, numerous solutions can be followed in the occurrence of delayed metrics depending on the flow or participation of events and queries types to be performed. The temporal databases and data warehouse field is joined to make temporal data warehouses filed. So, the Bi-temporal DW model was introduced by Canan Eren and Gözde Alp [27] in which transaction time and valid time are involved to the attributes. The objects and cubes of DW are created using multi-dimensional bi-temporal relational database.

Spatial and temporal data are two factors that seriously affect decision making and strategies of marketing and numerous applications need modeling and specialized treatment of these types of data because they cannot be efficiently conducted within a conventional multi-dimensional database. The main spatiotemporal application is industries of telecommunication and quickly ruled by massive data. Garani et al [28] were stated modeling of DW schema which integrate unified spatial and temporal data in usual DW framework. Spatial and temporal data integration is very important as the size and distribution of data grows.

3.3.6 Temporal Support in Data Warehouse

The temporal support provided in a DW essentially depends on the needs of the temporal maintenance and analysis provided by the source OLTPs. At this time, describing the different prospects concerning in OLTP timestamps availability and provision for specific timestamps in DW. To analyze the temporal data management and handling of a DW changes, thus create taxonomy to compile existing methods based on their features. According to the taxonomy, in order to clearly distinguish features of surviving methods, a DW is separated from handling temporal support changes and their support in business tools. Hence, this has the initial level of taxonomy. The TDM contains different properties and measures and it have dimension and fact table which is entered in the tables columns. On the other hand, those techniques not stated the current requirements of DW infrastructure like visualization of DW activities, supporting temporal dimensions, record keeping for long-term and etc. To overwhelm these challenges, offers a UML based DW system using temporal dimensional modeling. This structure designs user requirement based DW supporting time-dependent dimensions thereby allowing the end-user to accumulate the history of variations up to required time interval.

3.3.7 Demerits of Data Warehousing model

Although a DW is best for maintaining and storing data in a central warehouse and offers a number of advantages, there are demerits to using a DW. Some of these are provided because databases are not the optimum environs for un-configured data and data should be mined, transferred and weighed down into the warehouse, so this is a time consuming process. Integrating data from different sources is more complicated, time consuming and requires more effort. The data warehouses schema design and data marts is complex and dynamic. Selecting the optimum no. of dimensions in the DW is difficult. During their lifetime, data warehouse can have high-cost, which means maintaining cost of a DW can be high whether it is not accomplished correctly. Therefore, this study supports in selecting the role of the DW, definite functions in designing the database structure, and the responsibilities for designing the advanced version of the DW architecture. For these drawbacks, the data mining task is very useful for massive and complex data sets. Hence, we discussed data mining approaches.

3.4 Data mining concepts

Data mining is the method of data warehousing, which is extract the hidden details from huge databases besides the potential with new development technologies to help the organizations focused on data warehouses significant information. The techniques of data mining works from pattern recognition, parallel algorithms, statistics, visualizations, machine learning, database, computer performance and etc. mostly many of these techniques are usually used in image processing, vision, language understanding, handwriting recognition and speech recognition. Still, scaling and automated business intelligence solutions distinguish data mining from other statistical modeling and machine learning applications. Presently, we focused on the data mining most common techniques. For numerous ways in data mining, basically adaption from machine learning approaches to applications of business. For many of the approaches, the statistics are the foundation in data mining construction for example, discriminant analysis, standard deviation, regression analysis, standard variance, standard distribution, confidence intervals, and cluster analysis. Those techniques are utilized its relationships and to study the data. The data mining and data warehouse are alternate tools which depend on the robust data structure. Mustafa Erkayaoglu and Sean Dessureault [29] have been explored the data-driven architecture for modern mining and offerings data mining activity in real-time mining related data to predict explosive performance. On integration of DW, the adaptive boosting and random forest algorithms are utilized and to determine the efficient operation performances.

3.5 DM classification techniques

Several data mining topologies and systems has developed and designed. The several techniques are categorized based on the techniques to be used, database and the knowledge to be revealed. For the classification methods, several methods were suggested. Various database systems are based on the database which is utilized in organizations like spatial database, relational database, multimedia database, object-oriented database, transaction database, web and legacy database. The data mining system has been categorized and designed based on its database types. For example, if the system detects knowledge from the relational database which is the relational data mining scheme, and if it detects structure knowledge from the object-oriented database.

The different kinds of knowledge are discovered from data mining systems knowledge including characteristics rules, association rules, deviation analysis, evolution, clustering and classification rules. Also it can be classified along with discovered knowledge abstraction levels. For this knowledge might be categorized into multiple level knowledge, primitive knowledge and general knowledge. Based on different data mining technique, which can be classified as stated by driven approach like data-driven mining, autonomous knowledge mining, interactive data mining and query-driven mining techniques. Otherwise, it was classified based on underlying mining method, such as pattern-based mining, integrated approaches, generalization-based mining and statistical/mathematical based mining methods. Data mining is a process of discovering against big data to find patterns in decision making. The classification is one of the techniques in decision making. In data mining, classification is a technique and the decision tree method is applied to form the data, then the C4.5 decision tree algorithm is utilized to classify the data in the form of tree. The constructed system has better performance and minimum error in the system that differentiates the anomaly traffic by normal traffic. Robbi Rahim et al [30] were presented the data mining inventory system applications to minimize the production costs.

3.6 DM techniques

The most significant data mining techniques are reviewed in this section. In the **evaluation** of knowledge extraction, sampling and data selection, the indispensable component is statistics. It has been used to estimate the data mining results to isolate the good data from the corrupt data. In the data cleaning process, statistics provide tactics for discovering 'outliers' softening data while needed, and to assessing noise. Thereby utilization of estimation techniques, the statistics is also deals with missing data. For tentative data analysis, the clustering techniques and design experiments are worked. However, work on statistics has generally emphasized the theoretical features of models and techniques. Thus, critical search on data mining received little attention. Additionally, the most significant issues are database interface, techniques for handling large data sets and efficient data management techniques in data mining. However, these issues are beginning to gain attention in the statistics. In data mining, artificial intelligent techniques are mostly used technique as neural networks, machine learning and pattern recognition. AI other techniques for example representation of knowledge, search and knowledge acquisition are related to the numerous process activities in DM. In data mining, the major issue is classification of data. The classification is data sets diving into equally exclusive groups in which each group members are as near to each other as probable and the various group members are as far as probable from each other.

3.7 Decision tree approach

Sets of decisions are represented by decision tree it's like tree shaped structure. For the classification of dataset, decision tree tactic is generating the classification rules. The C4.5, ID3 and Classification and Regression Trees (CART) are the particular decision trees in data classification. It provides a set of rules that could be used on un-classified data to predict which records will contain a given result. CART generally wants fewer data production than ID3. From the training tuples of labeled classes the decision tree induction is learned by decision trees. Decision tree is a flow chart like tree structure where each inner/non-leaf node is represents an attribute, every branch is denotes the test outcome and every leaf node/terminal node holds as class label. The tree topmost node is root node, sometimes to buy a computer it denotes the concept of computer, which means predicts if a customer possible to buying a computer. The leaf nodes are indicted by ovals and rectangle form is denoted the internal nodes. Some decision tree mechanisms only create binary trees (each internal node branching to exactly two another nodes), while others can create non-binary trees. The algorithm of C4.5 is adopting the greedy approach and this is constructed based on conquer manner and top-down recursive manner. Many of the decision tree induction follows the top-down method in which tuples training set and the class labels are associated. Ross qiunlan have been developed decision tree based C4.5 algorithm. The researchers made improvements in decision tree algorithm. However the problem is, for the construction of decision tree, it requires sorting of data collection and multiple scanning. So as to improve the accuracy and classification, Kemal Polat and Salih Gunes [31] were proposed the hybrid approach of C4.5 decision tree classifier. To classify the multi-class problem included the various UCI (University of California Irvine) data sets. Further improvements in classification accuracy, the hybrid optimization based decision tree algorithms are used. The below section reviews the population based approaches in data mining classification technique.

3.8 Population based algorithms

The importance of C4.5 decision tree algorithm search rules efficiency has been the attention of many scholars. Hence, enlargement should be carried out to develop a novel, effective method; however it cannot be detached from the accuracy analysis as the results of the method. So, Genetic algorithm (GA) [32] was utilized to optimize the classification rules and simply the complex combinations. Hence the problem each possible solution is represented as population of rules which is initially created as randomly. The two kinds of rules are associated to generate offspring for next generation. In each generation, the members' genetic structures are randomly modified by mutation process. At next, the system runs at maximum iteration and process is terminated as to attain the optimal solution or particular criteria. This algorithm is suitable for many of the optimization approaches and also applied in data mining problems. In training package, the amount to be reduced is often the no. of classification errors. Larger and more complex difficulties require a faster system to get the right solutions at the right time. The mining of large data sets by genetic means has recently come into practice due to the availability of high speed. Several optimization algorithms have been used to classify data in a data mining methodology.

Using C4.5 with a hybrid genetic algorithm to find the most effective rules requires good understanding and a long time. But the use of both methods is often effective if the cases are more complex, and taking many branches and high accuracy. Here, Kun-Huang Chen et al [33] were proposed combined decision tree with gene selection by Particle Swarm Optimization (PSO) algorithm for classification performance. The cancer classification of microarray data were effectively classify the data and to identify the cancers. To extract the classification rules, the Ant Colony Optimization (ACO) algorithm was applied with decision tree induction. It was proposed by Fernando E.B et al [34] for combining ACO and conventional decision tree algorithm.

DM can be used through several profitable organizations, non-profit, educational institution & even for various sectors: marketing companies, retail, finance, and communications for their long-term survival and success with employee satisfaction and development, strong consumers focuses on their retention, and the modernization of a particular organization. Process of data mining automates of discovering predictive statistics on huge databases. There are different types of algorithms and learning approaches for prediction and rule extraction. The proposed classical approach of C4.5 decision tree algorithm is combined with Selfish Herd Optimization (SHO) algorithm to tune the gain of the given data set. In this technique, the optimal weight for the information gain will be updated based on SHO. The data set is classified with quadratic entropy calculation and the information gain of the UCI data sets. The suggested approach is to attain the higher accuracy performance than the conventional approaches.

Real world data is very large and is stored in a DW in multi-dimensions, so the size of this metric which mean the number of rows and attributes can be very poorly expressed and complex, which is very challenging to sustain and handling. Apart from this, the warehoused data may perhaps numerous false values or be incomplete. Thus, designing and implementing mining techniques can be very exciting and time consuming. In any of the purpose, small data warehouse was created for the simplicity. The DW size is well-defined through the no. of attributes and records which is named as example database. Therefore, the proposed methods will be successfully implemented as a training data set in this example database, but this does not apply to a data warehouse containing thousands of attributes and records that are presented in real-world data.

4. Conclusion

The main objective of this study was to investigate the data warehouse architectures, multidimensional models, conceptual schema and data mining techniques and its challenges have been studied. Moreover, it enlightens the data warehouse, data mining and related works. Finally, it supports data mining approaches, challenges in data and analysis, and effective decision making with an industrial perspective. This survey motivation is to support the research initiative, in what way can large data be integrated and transformed with data sitting within a data warehouse and mining to perform efficient decision making. We have not been able to determine the limits accurately, so the analysis of the study results has not yet been completed however, the size of the model and the use of data collection is to be surveyed. In future, our focus will not be limited to integrating a data management platform to improve decision making. But also in business models that are of particular importance in analytics. This survey is also the overview prospects of data warehouse and data mining analysis for the effectiveness of large amount of data.

References

[1] Prat, Nicolas, Jacky Akoka, and Isabelle Comyn-Wattiau. "A UML-based data warehouse design method." Decision support systems 42, no. 3 (2006): 1449-1473.

[2] El-Sappagh, Shaker H. Ali, Abdeltawab M. Ahmed Hendawi, and Ali Hamed El Bastawissy. "A proposed model for data warehouse ETL processes." Journal of King Saud University-Computer and Information Sciences 23, no. 2 (2011): 91-104.

[3] Yusof, Sharmila Mat, Fatimah Sidi, Hamidah Ibrahim, and Lilly Suriani Affendey. "A study of multidimensional modeling approaches for data

warehouse." In AIP Conference Proceedings, vol. 1761, no. 1, p. 020063. AIP Publishing LLC, 2016.

[4] Vassiliadis, Panos, Alkis Simitsis, and Spiros Skiadopoulos. "Conceptual modeling for ETL processes." In Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP, pp. 14-21. 2002.

[5] Bimonte, Sandro, Leandro Antonelli, and Stefano Rizzi. "Requirements- driven data warehouse design based on enhanced pivot tables." REQUIREMENTS ENGINEERING (2020).

[6] Erkayaoglu, Mustafa, and Sean Dessureault. "Improving mine-to-mill by data warehousing and data mining." International Journal of Mining, Reclamation and Environment 33, no. 6 (2019): 409-424.

[7] Çığşar, Begüm, and Deniz Ünal. "Comparison of data mining classification algorithms determining the default risk." Scientific Programming 2019 (2019).

[8] Saritas, Mucahid Mustafa, and Ali Yasar. "Performance analysis of ANN and Naive Bayes classification algorithm for data classification." International Journal of Intelligent Systems and Applications in Engineering 7, no. 2 (2019): 88-91.

[9] Arellano, Aldo Ramirez, Juan Bory-Reyes, and Luis Manuel Hernandez-Simon. "Statistical Entropy Measures in C4. 5 Trees." International Journal of Data Warehousing and Mining (IJDWM) 14, no. 1 (2018): 1-14.

[10] Oriol, Xavier, Ernest Teniente, and Albert Tort. "Computing repairs for constraint violations in UML/OCL conceptual schemas." Data & Knowledge Engineering 99 (2015): 39-58.

[11] Tomic, Igor, Drazen Brdjanin, and Slavko Maric. "A novel UML profile for representation of a relational database schema." In IEEE EUROCON 2015-International Conference on Computer as a Tool (EUROCON), pp. 1-6. IEEE, 2015.

[12] Azzaoui, Amine, Ouzayr Rabhi, and Ayyoub Mani. "A Model Driven Architecture Approach to Generate Multidimensional Schemas of Data Warehouses." International Journal of Online and Biomedical Engineering (iJOE) 15, no. 12 (2019): 18-31.

[13] Ayadi, Mouhamed Gaith, Riadh Bouslimi, and Jalel Akaichi. "A framework for medical and health care databases and data warehouses conceptual modeling support." Network Modeling Analysis in Health Informatics and Bioinformatics 5, no. 1 (2016): 13.

[14] El Beggar, Omar, Khadija Letrache, and Mohammed Ramdani. "CIM for data warehouse requirements using an UML profile." IET Software 11, no. 4 (2017): 181-194.

[15] El Hayat, Soumiya Ain, Fouad Toufik, and Mohamed Bahaj. "UML/OCL based design and the transition towards temporal object relational database with bitemporal data." Journal of King Saud University-Computer and Information Sciences 32, no. 4 (2020): 398-407.

[16] El Beggar, Omar, Khadija Letrache, and Mohammed Ramdani. "Towards an MDA-oriented UML profiles for data warehouses design and development." In 2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA), pp. 1-6. IEEE, 2016.

[17] Dahiya, Naveen, Vishal Bhatnagar, and Manjeet Singh. "A fuzzy based matrix methodology for evaluation and ranking of data warehouse conceptual models metrics." Int. Arab J. Inf. Technol. 15, no. 2 (2018): 202-212.

[18] Golfarelli, Matteo, Dario Maio, and Stefano Rizzi. "Conceptual design of data warehouses from E/R schemes." In Proceedings of the thirty-first Hawaii international conference on system sciences, vol. 7, pp. 334-343. IEEE, 1998.

[19] Oliveira, Bruno & Belo, Orlando. (2012). BPMN patterns for ETL conceptual modelling and validation. 7661. 445-454. 10.1007/978-3-642-34624-8_50.

[20] Santos, Maribel & Costa, Carlos & Galvão, João & Andrade, Carina & Pastor, Oscar & Marcén, Ana. (2019). Enhancing Big Data Warehousing for Efficient, Integrated and Advanced Analytics: Visionary Paper. 215-226. 10.1007/978-3-030-21297-1_19.

[21] Di, Tria Francesco, Ezio Lefons, and Filippo Tangorra. "Academic data warehouse design using a hybrid methodology." Computer Science and Information Systems 12, no. 1 (2015): 135-160.

[22] Reddy, G. Sekhar, and Ch Suneetha. "Conceptual Design of Data Warehouse using Hybrid Methodology." International Journal 9, no. 3 (2020).

[23] Vieira, António AC, Luís Pedro, Maribel Yasmina Santos, João Miguel Fernandes, and Luís S. Dias. "Data requirements elicitation in big data warehousing." In European, Mediterranean, and Middle Eastern Conference on Information Systems, pp. 106-113. Springer, Cham, 2018.

[24] Wang, Yahui, Suihuai Yu, and Ting Xu. "A user requirement driven framework for collaborative design knowledge management." Advanced Engineering Informatics 33 (2017): 16-28.

[25] Pedersen, T., Jensen, C., & Dyreson, C. (1999). Supporting Imprecision in Multidimensional Databases Using Granularities. Proceedings of the 11th International Conference on Scientific and Statistical Database Management (SSDBM'99), (pp. 90-101). Cleveland, USA.

[26] Malinowski, Elzbieta, and Esteban Zimányi. "Hierarchies in a multidimensional model: From conceptual modeling to logical representation." Data & Knowledge Engineering 59, no. 2 (2006): 348-377.

[27] Atay, Canan Eren, and Gözde Alp. "Modeling and querying multidimensional bitemporal data warehouses." International Journal of Computer and Communication Engineering 5, no. 2 (2016): 110.

[28] Garani, Georgia, Andrey Chernov, Ilias Savvas, and Maria Butakova. "A Data Warehouse Approach for Business Intelligence." In 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), pp. 70-75. IEEE, 2019.

[29] Erkayaoglu, Mustafa, and Sean Dessureault. "Improving mine-to-mill by data warehousing and data mining." International Journal of Mining, Reclamation and Environment 33, no. 6 (2019): 409-424. Rahim, Robbi & Zufria, Ilka & Kurniasih, Nuning & Simargolang, Muhammad & Hasibuan, Abdurrozzaq & Sutiksno, Dian & Nanuru, Ricardo & Anamofa, Jusuf & Ahmar, Ansari & GS, Achmad. (2018). C4.5 Classification Data Mining for Inventory Control. International Journal of Engineering & Technology. 7. 68. 10.14419/ijet.v7i2.3.12618.

[30] Polat, Kemal, and Salih Güneş. "A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems." Expert Systems with Applications 36, no. 2 (2009): 1587-1592.

[31] Damanik, Irfan Sudahri, Agus Perdana Windarto, Anjar Wanto, Sundari Retno Andani, and Widodo Saputra. "Decision Tree Optimization in C4. 5 Algorithm Using Genetic Algorithm." In Journal of Physics: Conference Series, vol. 1255, no. 1, p. 012012. IOP Publishing, 2019.

[32] Chen, Kun-Huang, Kung-Jeng Wang, Kung-Min Wang, and Melani-Adrian Angelia. "Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data." Applied Soft Computing 24 (2014): 773-780.

[33] Otero, Fernando EB, Alex A. Freitas, and Colin G. Johnson. "Inducing decision trees with an ant colony optimization algorithm." Applied Soft Computing 12, no. 11 (2012): 3615-3626.