A Survey on Data Warehousing: State-of-the-art Comparison

Dr. Poonam Yadav
D.A.V College of Engineering & Technology,
Kanina, Haryana 123027, India

Abstract*— Data warehouse is utilized for the decision making in industries, schools, and in other various fields. DW is implemented for accessing the data source, i.e., heterogeneous; clean, transform, and filter the data, understand, use and to keep all the information's as a structure and make it to access easier. These stored data are utilized in data analysis, query and reporting; thereby the suitable information in present state of construction can be extracted, and the decision-making process is executed. This paper explains the research work on 20 papers about data warehousing and its methods. In conclusion, the outcomes are analyzed with the tabulation and diagrammatic representation.

***Keywords*—Data Warehousing, DWP, Distributed DW, Query language.**

I. INTRODUCTION

The DW is defined as integrated, subject-oriented, nonvolatile and time-varying collected works of information that maintains the decision-making process in the management. DW is an essential constituent for each and every organizations, and it is said to be a intensive-data systems, thereby used in business and other works for analyzing the tasks like customer preferences statistics, cost/benefit relation statistics, analyzing sales/ profits statistics, and so on. The historical and operational information's are stored in the data warehouse in multidimensional cubes. DW consists of an implanted batch updates along with the read-only queries. In general, every organization has a centralized data warehouse structure. In an organization to store the overall information's, decision making and for responding to queries, takes place because of this centralized DW. When the data informatics amount and the queries amount get increased, then there is a requirement in distribution of accumulated information is generated, which develops the Distributed DW. Distributed DW system is said to be a collective interconnection of heterogeneous computers that are associated with the local and wide area networks

Multi-database systems are the combination of existed, independent, and perhaps may be a heterogeneous database structure. The multi-database with existed database structure is said o be a component or local database structure. Generally, multi-database is referred to a non-federated structure thereby connects several database structures that are heterogeneous. Moreover, it provides a query language (common) without global integrated module to specify the queries and transactions. Alternatively, the federal database structure offered a total or partial integration of data scheme with the module database structure. Queries include logical processing that consists of roll-ups, drill-downs dicing, and slicing, and aggregations were supported by these DW.

When the amount of data and the queries number increases, then it is said to be a more complex process for the distributed system. Therefore it has to be planned carefully to resolve the problems so that it gets up to the user requirements. While there is a need of change in the information of organization, then the change in DW requirements have to be taken into consideration thereby makes the process more difficult. The difficulties in DW process includes the preset expectation and goals that are not meet; over-budget, stalled projects or delayed; organizational politics; and technological challenges and obstacles. Therefore the algorithms have to be developed to solve these problems in the distributed DW processes.

In this paper, the data warehousing methods are reviewed underperformance measure, algorithmic analysis and by better performance. The arrangement of paper is as follows: Section II explains the Literature review. Section III is about the analysis review that includes the algorithm analysis and performance measure. Research gaps and challenges are defined in Section IV. Section V concludes the paper.

II. LITERATURE REVIEW

A. Related Works

In 2015, Luis Marco-Ruiza *et al.* [1] has said that at the time of health care delivery, the data reuse was an important one to gratify the clinical decision support systems and clinical research requirements. The major obstruction to reuse was because of the availability of the information in legacy format, and its data granularity was high when kept in a system named electronic health record (EHR). Thereby to permit the reuse of these data at any time of requirement, the methods were needed, thereby to aggregate, standardize and query data obscured in EHRs. The main objective was to produce a warehouse infrastructure of data by deploying archetype-based technologies, a standard query language that allowed the interoperability required to reuse the data. In this framework, the better archetype-based storage technology and alteration of data were implemented for generating a flow of work in extraction, load, transformation, and modeling of related EHR data with normalized data repositories. Archetype-based conversion was used to change the legacy of information and execute the patient-centered aggregations. Finally, the proposed model was worked at some use case query level in definite purpose aggregation.

In 2013, S. Krishnaveni and M. Hemalatha [2] have stated that the data warehousing was distributed primarily on the basis of, how the processing of data taken place and to differentiate the physical and dynamic distributed environment. By accessing the large database for discovering the applicable information, there need a lot of time and a huge laborious process. In the distributed data warehouse, these problems can be rectified by utilizing the query scheduling process which was a power-packed method to achieve this performance within a minute. Dynamic Fault-Tolerant Dependency Scheduling (DFTDS) algorithm was implemented depending on the dependency to schedule the queries. Thereby after examining the status of virtual machine, it systematically allocates the asset depend on the acknowledgment, which was said to be the user/client queries that have to be solved in data warehouse distributed technology. On comparing with the conventional methods, the proposed DFTDS algorithm proved efficient minimization in query processing and utilization of memory.

In 2015, Abhay Kumar Agarwal and Neelendra Badal [3] has defined that the enormous development in the amount of instruction and their storage function elongates these flat files system into Data Warehousing (DW), RDBMS and Distributed Data Warehousing technologies. A unique distributed data warehouse method was implemented for the excellent distribution in data warehousing, and it was named as Intelligent and Distributed Data Warehouse (IDDW). To construct the distributed data warehouse, the implemented method was developed with N-levels on the basis of top-down hierarchical design strategy. The IDDW process started by detecting the various locations to build the DW. This method presented a model for altering the relevant information from any DW in the upper level to the relevant DW in lower level. Here IDDW modeling, the IDDW internal association, it's modeling based architecture was also implemented through that entire functions inside IDDW were taken place.

In 2018, Jennifer Vander Weele *et al.* [4] has implemented a method for the prime diagnosis of diabetic ketoacidosis (DKA) by validating the enterprise data warehouse(EDW) intended for the cohort in hospitalized patients. By utilizing the EMR, the validation has taken place that analyzes the 10% admission. DKA clinical position and the diabetes type's classification have been evaluated within the EDW instruction and the EMR review.

In 2014, Alejandro Maté *et al.* [5] have said that the development of DW has been enhanced while executing the elicitation stages of requirements in which the requirement of users was constructed. Nowadays there has enormous proposals were evolved with different modeling requirements, in this proposals goal-oriented models and i* framework was considered as a major one. In i* model generally the proposed model have reduced modularity and thereby cause a negative defect on DW development, while required DW method be likely involves a large number of parameters and their traverse relationship among them. Hence the model's readability was minimized caused an effect in the maximization of error rate and the development time also harming the utility. The proposed model was an expansion of i* model to modularize the DW goals. The advantage of this implemented model was maximizing the modularity and scalability thereby maximize the error correction capacity and to develop trouble free understanding of complex methods to the normal users and the developers.

In 2008, Daniel L.Rubin and Terry S.Desser [6] have defined that most of the needed information for teaching and research about radiology was dispersed the information within the hospital systems all through the medical endeavor. This work aimed to explain the methodology, performance and the design of the information in the data warehouse to incorporate and made available the medical data types that were relevant to radiology teaching and research works. In the proposed method the radiology data warehouse was designed and presented by identifying the data (Radbank) to overcome this requirement. Radbank was constructed based on the Linux platform with open-source software tools with a relational database. A text report parsing component was formed which identifies the radiology report structure and made available the entire discrete segment to search and indexing. To connect the pathology and radiology reports, a database model has been constructed thereby allowed the users to recover the cases utilizing the queries that were flexible.

In 2017, Karsten U.Kortüm *et al.* [7] has explained that to attains the scientific usage of maximizing digital data from diagnostic devices and EMR; they implemented a near-real-time DW in a scholastic ophthalmologic center. The main aim was to develop the database by presenting a definite macular clinic user interfaces inside the institutional system. The EMR-linked picture-archiving and communications system throw an order for imaging modalities to the corresponding devices. The entire information was collected within the DW that operates on SQL database. The discovery tool was presented for the data; thereby an excellent search about patient was carried out regarding executed cataract surgery, minimum 10 intra-vitreal injections, and patients with age-related macular degeneration.

In 2006, Jane Zhao and Hui Ma [8] have described the logical function in business by On-line analytical processing (OLAP) systems. This function does not rest on Latest transaction updates, and the data that utilized within these OLAP systems were considered to be in data warehouse. Thereby divides the outcome of dialogue interfaces from the operational database input sources in OLAP. In the appliance of Abstract State Machines (ASMs), there was a challenging area in data warehouses and OLAP systems. The layered ground model specification was developed for OLAP system and data warehouses. This method is on the basis of the essential idea that explicitly divides the outcome to OLAP systems from the operational input databases. Some rules were defined, and they were formally right in the high-quality design of OLAP systems and data warehouses by performing the ASM formalism without known mathematical details. Moreover, the pragmatic guidelines were discussed in the application of that rules.

In 2011, Katerina Doka *et al.* [9] have presented the Brown Dwarf, which is a data analysis with distributed system that was constructed to query, update and store multidimensional data effectively within the network commodity nodes. The central indexing structure between peers on-the-fly was dispersed by Brown Dwarf, thereby implements parallelization to reduce the creation of cube and inquiry times. Unstructured Peer-to-Peer overlie was thus formed naturally by these analytical queries that carried out by cooperating nodes online. Expensive over-night process was generally neglected by these updates that carried out online. The proposed model was estimated on the basis of original testbed: the creation of cube accelerates was controlled and querying in enormous time measured with the centralized solution thereby utilized the capacity to work in parallel among the existing network nodes. During unexpected burst in load, it accommodates faster and left unaltered by significant fraction of loss in node failure.

In 2008, Neoklis Polyzotis *et al.* [10] have explained that the data warehousing with active techniques has came out to overcome the conventional warehousing techniques with respect to congregate the up-to-date information over huge-demand applications. Active warehouse was recharged online in a nutshell thereby attained a large reliability over the information that was stored and up to the recent data updates. In the execution of data warehouse transformations numerous challenges were initiated because of the requirement in online warehouse stimulation, in regards with their overhead and execution time that takes place during the process of warehouse. The proposed method focused on the method named the join of a fast stream S of source updates with a disk-based relation R, over a restriction of limited memory. This was the core operation for numerous usual transformations that involves duplicate detection, recognition of newly inserted tuples, or surrogate key assignment. Connected algorithm named as mesh join (MESHJOIN) was implemented to recompense the two combined input as (a) sharing the I/O cost of accessing R across multiple tuples of S and (b) relying entirely on fast sequential scans of R. the two main aim of this systematic model was: increasing throughput under a specific memory budget or reducing memory consumption for a specific throughput. The investigational studies examine the Maintaining the Integrity of the Specifications

In 2012, Arun Sen *et al.* [11] has stated that there was an enormous growth in the data warehouse market, though high investment has been needed for DW. But most of them ended in failure. Here, the maturity of data warehousing process (DWP) were discussed for moderating the huge-scale failures and make sure a reliable delivery, "single-version of truth", huge quality in a appropriate approach. For the period of over three years, the five-level DWP maturity model (DWP-M) has been designed and developed. The distinctive feature of this method was both the operations and the development of data warehouse processes were covered. The board of experts evaluates the last method that explained the outcome as the productivity, functionality, and usability were validated perfectly in this method. Here the first and the foremost model version of DWP-M were presented next to the different phases of maturity with the design of the numerous key processes.

In 2012, Lukasz Golab *et al.* [12] have modeled a update problem on the streaming warehouse based on the scheduled issues, thereby the work-related for processing load the new instruction inside the tables, and its aim was to reduce the data staleness under time. Then the authors proposed a scheduled model thereby the difficulties that were encounter by a stream warehouse was taken care of: data consistency, heterogeneity of update jobs inability to preempt updates, view hierarchies and priorities were generated by several inter-arrival times and volume of data between transient overload and different sources. The narrative characteristics of the proposed system were that the scheduled decisions do not rely on the update job properties (like deadlines) yet because of the cause of data staleness in update jobs. At last, the factors that affected the characteristics were solved by using the update scheduled algorithms and wide simulation experiments.

In 2011, Risa B. Myers and Jorge R. Herskovic [13] have investigated about, how to attain the exact counts of patient from the clinical data warehouse that consist of synthetic data of patients. For this, the probabilistic techniques were applied that was aided by May BMS probabilistic database. Here the synthetic Clinical Data Warehouse was presented that was populated through simulated data by deploying the generation engine with custom patient data. Then various techniques were evaluated, compared and implemented to achieve the count of patients. By performing the "stimulated expert analysis", the specificity and sensitivity of the billing were calculated. Further, the experts analyzed and labeled the diplomat sample of records and attain the fact of each record. The subsequent probability of the patient condition was calculated after every visit by 'Bayesian Chain'' on the basis of Bayes' theorem. The possibility of patient's with condition depends on the patient billing event was calculated by utilizing the second approach named as ''one-shot''. In general, by utilizing the Bayesian probabilistic techniques, it was concluded that the patient counts on replicated patient population were improved consistently. This possibility technique made sure the better outcome for applications and attained a clear adequate patient counts.

In 2008 K. Ramamurthy *et al.* [14] have performed a study about the development of structural method with 117 companies contributed, having two senior managers (one was IS and rest was the line function) to conduct the test of research theory. The outcome of this experiment stated that from the seven modules four were observed. This observation shown that quality of the project management process, complexity, organizational support, and compatibility considerably manipulate the infusion degree in DW and thereby considerably manipulate the stakeholder contentment and organization level advantage. Here this framework was a good performance in practice and research of DW and IT combination, and also for the impact over an organization-level combination of decision support and vast infrastructural methods that involves DW.

In 2006, Arun Sen *et al.* [15] have investigated about the factors that influence the awareness of process maturity in data warehousing. DW was said to be a software development process that includes components such as workflows and artifacts. This framework laid on the management observation of DWP. The outcome of this experimental study stated that the DWP maturity was affected by numerous factors like change management, data warehouse size alignment of architecture, data quality, and organizational readiness. The practical studies revealed that the outcome offered a technological and managerial helpful pointer thereby to achieve a lot more mature levels in the organization for elevating data warehousing processes.

In 2006, Ferenc Peter Pach, *et al.* [16] have developed a data-warehouse-based OSS that made the isolated production units and potential linking complex on the basis of the combination of the information that was collected heterogeneously from the complex production procedure. On the basis of the designed data warehouse the OSS was implemented subsequently followed by introduced data-warehouse focus-on-process design procedure. This referred the major concern with an information and material flow along the total enterprise. The OSS that resulted thus tag along the procedure along with organization and it does not focused on the task that was separated in the isolated process entity. The notion of this framework was explained on the basis of case study in industrial-level, while the OSS was introduced to control and monitor the high-density polyethylene (HDPE) plant.

In 2003, Nick Bassiliades, *et al*. [17] have described the combination of knowledge-based system and a multi-database system for performing the data integration module in the DW. The databases with numerous modules having general query language were combined in this multi-database system. Moreover the capacity for schema integration and additional necessary utilities were not provided by this in DW. Additionally, the second-order syntax along with declarative logic language was offered by this knowledge base system. Also, integrity checking, cleaning and to summarize the data into the DW, the deductive rules were used. Thereby the feature of this Knowledge Base System viewed an effective maximization in maintenance mechanism, which was utilized to refresh the DW with no data source query.

In 2003, Ling Feng and Tharam S. Dillon [18] have studied that the huge amount of data that was extracted and summarized from various sources integrated to form a DW, which was used for analyzing and querying directly. Thereby this offered an effortless access to decision-makers on attaining aggregated and historical data. Fuzzy technology was developed thereby it offered the development in aggregation and summarization of data in DW systems. This DW semantic method made up of three layers that involve qualitative (categorical) summarization, quantifier summarization, and quantitative (numerical) summarization. This was implemented in the semantics of DW for explicating and capturing of data's. Further, the SQL language was extended in order to access the queries that were flexible in opposition under improved DW.

In 2017, Yang Hu, *et al*. [19] has developed a nonrelational, DW, analytics environment (Energy-CRADLE) and distributed computing. Thereby the field and laboratory information in several heterogeneous photovoltaic (PV) test locations were analyzed. The various arrangements of the PV characteristics information and climatic telemetry time-series instructions that were collected from the PV outdoor test network were designed by these data informatics and analytics infrastructure. Energy-CRADLE does not consist of a predefined data table system; thereby the distributed DW used the Hadoop/HBase that allowed the intake of instruction in changed and diverse format. For allowing the Python MapReduce and to offer a graphical user interface, the Energy-CRADLE deploys the Hadoop streaming; thereby attained an effortless data retrieval and data ingestion, i.e. py-CRADLE. Energy-CRADLE demonstrated a secure, integrated, user-friendly and scalable informatics data and analytics schemes in PV researchers, by improving the HBase NoSQL database schema and Hadoop distributed computing platform within solar energy.

In 2015, João C. Ferreira, *et al.* [20] discussed an outcome in application research on the basis of a large information set for the eco-driving domain that was generated from the public transportation buses in Lisbon along the period of three years. These set of information's was enriched with GPS coordinates, road information and weather conditions and it was obtained automatically from control area network bus. The knowledge discovery (KD) techniques and online analytical processing (OLAP) was applied here. Thereby establish the main feature that controls the average fuel consumption and to process the large volume in this set of data, and to differentiate in accordance with the driving effectiveness of the drivers. As a result, the suitable driving styles and practices were identified. In this framework, the training sessions of driver were the major concern to be considered.

III. ANALYTICAL RESULTS

A. *Algorithmic Analysis*

The Fig. 1 illustrates the various algorithms that are utilized in this proposed model. The implemented algorithmic methods are as follows: Archetype-based data transformation is the algorithm that is deployed in [1]. The author in [2] explains the data warehousing using the DFTDS algorithm. In [3], the author imports the IDDW algorithm. The EMR is the algorithm that is utilized in [4]. In [5] the author extends i* profile algorithm for better performance. Rad Bank is the algorithm that deployed by the author in [6]. Macular clinic user interface is used in [7]. The author in [8] utilizes Layered ground model specification. Brown draft is the algorithm deploys in [9]. In [10] combination of two algorithm is used and it is named as MeshJoin. In [11], DWP maturity model is deployed. The author used the scheduling algorithm in [12]. The mixture of Archetype-based data transformation and MayBMS were utilized in [13]. In [14], the author developed a field study. CMM is the algorithm that presented in [15]. Traditional OSS algorithm is used in [16]. The combination of both the multi-database system and knowledgeable-base system are used in [17]. In [18] the author deploys the Fuzzy technology. Hadoop/HBase is the algorithm utilized in [19]. The OlAP along with KD is deployed by the author in [20].

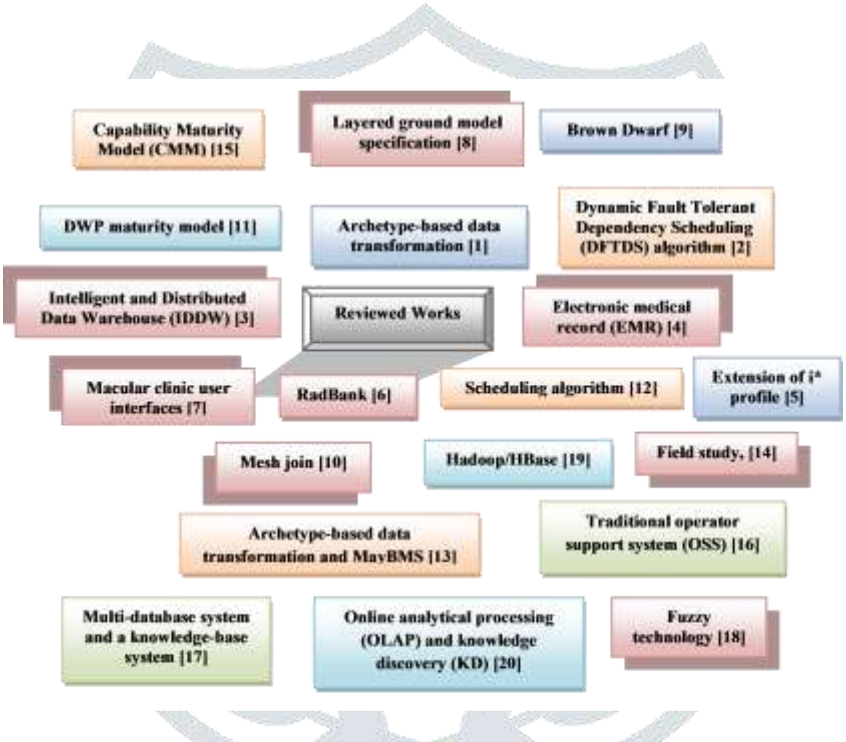

Fig. 1. Algorithmic Analysis of Reviewed Works

B. *Performance Analysis*

The performance analysis can be done by utilizing the performance measures that are analyzed in this framework and it is shown in table I. some of the measures that are deployed in this reviewed model is average time, scalability, reliability, mean and so on. In this overall contribution, 15% is utilized for the measurement of average time. Only 10% is the contribution of scalability, mean as well as reliability measures in the proposed one. Some of the other measures also used in this research work like POS, max service rate, relative lateness, dwarf size, std deviation, normalized MI, q-product, population and so on.

TABLE I. PERFORMANCE MEASURES OF REVIEWED WORKS

Citation	Average time	scalable	reliable	OPS	Max service rate	Relative lateness	mean	dwarf size	Std-deviation	Normalized MI	q-product	Population	others
[1]	□												
[2]		□											
[3]			□										
[4]													□
[5]	□	□											
[6]													□
[7]				□									
[8]													□
[9]								□					
[10]					□								

[11]													□
[12]						□							
[13]	□								□				
[14]			□				□						
[15]							□						
[16]										□			
[17]													□
[18]											□		
[19]													
[20]												□	□

C. Maximum Achieved Measures

The efficiency of the measured value with respect to the data warehousing is given in Table II. Here, the effective feature of the average time is 1.287. The scalability measure that is used in this model achieves 13% whereas the error-free execution and the query cost attain 3%, and 28%, respectively. The worst case and the best case reliability can be given a value as 0.285714 and 0.857143, correspondingly. 0.826 is the measure value that achieved by OPS. The tuples and dwarf size measures that are utilized in [9] have the value 10k and 1.2. In [12] the max service rate has attained 11250. The relative lateness measure value is 1.59. The Standard deviations that are used in this framework accomplish 39.8. The best mean value performance is given as 4.773. In [16], 0.32 is the value attained by the measure normalized MI. in the Q product sale the most sales achieved value is 0.6. The population value measured in [20] attains the value of 368 and 394 for AFC1 and AFC2, respectively.

TABLE II. MAXIMUM ACHIEVED MEASURE

Measure	Best performance value	citation
Average time	1.287	[1] [5] [13]
Scalability	13%	[2]
Error free execution	3%	[2]
Query cost	28%	[2]
worst case reliability	0.285714	[3]
Best case reliability	0.857143	[3]
Readability score	0.826	[5]
OPS	720721	[7]
Tuples	10k	[9]
Dwarf size	1.2	[9]
Max service rate	11250	[10]
Relative lateness	1.59	[12]
Std deviation	39.8	[13]
Mean	4.773	[14]
Normalized MI	0.32	[16]
Q product Most sale	0.6	[18]
Population AFC1,AFC2	368,394	[20]

D. Analysis on Subjective Data Warehousing Operation

This review shows the process incorporated in the data warehousing concept from the 20 papers and is illustrated in Fig. 2. This process includes data transformation, data cleansing, and data reuse. Here the data transformation contributes the maximum of 40% in the overall achievement. 20% is the contribution of the data reuse, and only 10% is the contribution of the data cleansing method. The remaining 30% is contributed by the rest of the measures such as validation, infusion, etc.

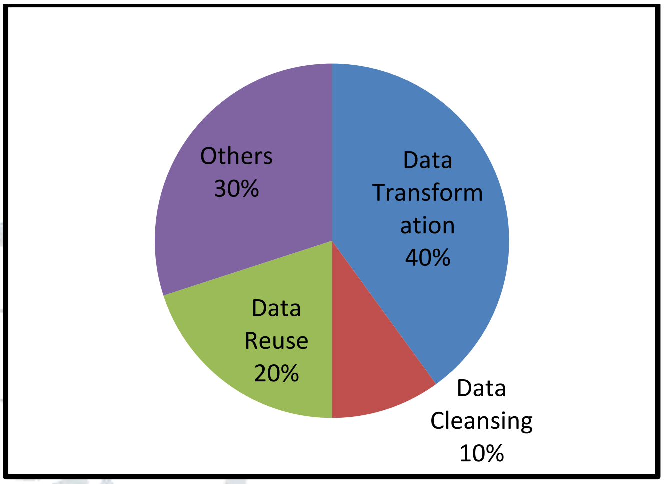

Fig. 2. Pie chart on Data Warehousing Operations

IV. RESEARCH GAPS AND CHALLENGES

The main challenges in data warehousing are the extra report work. If the organization is large, there is a need for more data and additional time and load for the data warehouse to run. The DWH is in need of the participation of every department that involved thereby creates an additional report work. Another challenge is inflexibility and homogenization in data. This will cause the limitation in data with respect to the relation that established at the time of aggregation and thereby generate difficulty in query speed. DW evaluation needs a huge amount of data resources for handling and managing the data from several sources. Thereby it increases the cost and benefits ratio in companies. One of the major challenges in DW is the updating ability. Because of this updates, the DW has to maintain regularly thereby increase the maintenance cost. It also has difficulties in compatibility.

Depending on the details of software, operating systems, and programs, it is difficult for the user to find out the proper use of the DW. The changed data that are captured by the time frame can be small because it is planned for the real time processing. Hence the window size is said to be small that incarnate low information for the operation. Another major challenge is the traditional SQL queries and functions are not supported. Hence while using SQL query, there may be a variation during the report generation. Even though the conceptual DW is still undergrowth in various fields, there remain some challenges that are mentioned above have to be rectified to recognize the powerful, flexible and efficient DW.

V. CONCLUSION

In this, the 20 papers are analyzed with regards to the data warehousing techniques and their challenges that have to be resolved. The reviewed outcome explains the various methodologies along with the analysis of the performance and the maximum attained measures. The tabulation reviews and the diagrammatic representation evaluate these reviewed works. Moreover, the challenges along with future work are explained in these papers.

REFERENCES

[1] Luis Marco-Ruiza,David Moner,José A. Maldonado, Nils Kolstrup ,Johan G. Bellika,"Archetype-based data warehouse environment to enable the reuse of electronic health record data,"International Journal of Medical Informatics, vol. 84, no. 9, pp 702-714,September 2015.

[2] S.Krishnaveni ,M.Hemalatha, "Evaluation of DFTDS algorithm for distributed data warehouse,"Egyptian Informatics Journal,vol. 15, no. 1, pp.51-58,March 2014

[3] Abhay Kumar Agarwal, Neelendra Badal,"A novel approach for intelligent distribution of data warehouses,"Egyptian Informatics Journal, vol. 17, no. 2,pp. 147-159, July 2016

[4] Jennifer VanderWeele, Teresa Pollack ,Diana Johnson Oakes, Colleen Smyrniotis, Vidhya Illuri, Priyathama Vellanki, Kevin O'Leary,Jane Holl ,Grazia Aleppo, Mark E.Molitch ,Amisha Wallia,"Validation of data from electronic data warehouse in diabetic ketoacidosis: Caution is needed,"Journal of Diabetes and its Complications, vol. 32, no. 7, pp. 650-654, July 2018

[5] Alejandro Maté, Juan Trujillo, Xavier Franch,"Adding semantic modules to improve goal-oriented analysis of data warehouses using I-star,"Journal of Systems and Software, vol. 88,pp. 102-111, February 2014

[6] Daniel L.Rubin MD, MS, Terry S.Desser MD,"A Data Warehouse for Integrating Radiologic and Pathologic Data,"Journal of the American College of Radiology, vol. 5, no. 38, pp. 210-217, March 2008

[7] Karsten U.Kortüm ,Michael Müller ,Christoph Kern, Alexander Babenko ,Wolfgang J.Mayer, Anselm Kampik,Thomas C.Kreutzer ,Siegfried Priglinger, Christoph Hirneiss,"Using Electronic Health Records to Build an Ophthalmologic Data Warehouse and Visualize Patients' Data,"American Journal of Ophthalmology, vol. 178, pp. 84-93, June 2017

[8] Jane Zhao ,Hui Ma,"ASM-based design of data warehouses and on-line analytical processing systems,"Journal of Systems and Software ,vol. 79, no. 5, pp. 613-629, May 2006,

[9] Katerina Doka, Dimitrios Tsoumakos,Nectarios Koziris,"Brown Dwarf: A fully-distributed, fault-tolerant data warehousing system,"Journal of Parallel and Distributed Computing, vol. 71, no. 11, pp. 1434-1446, November 2011

[10] N. Polyzotis, S. Skiadopoulos, P. Vassiliadis, A. Simitsis and N. Frantzell, "Meshing Streaming Updates with Persistent Data in an Active Data Warehouse," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 7, pp. 976-991, July 2008.

[11] A. Sen, K. Ramamurthy and A. P. Sinha, "A Model of Data Warehousing Process Maturity," IEEE Transactions on Software Engineering, vol. 38, no. 2, pp. 336-353, March-April 2012.

[12] L. Golab, T. Johnson and V. Shkapenyuk, "Scalable Scheduling of Updates in Streaming Data Warehouses,"IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 6, pp. 1092-1105, June 2012.

[13] Risa B.Myers, Jorge R.Herskovic, "Probabilistic techniques for obtaining accurate patient counts in Clinical Data Warehouses,"Journal of Biomedical Informatics

[14] vol.44, no. 1, pp. S69-S77, December 2011 K. Ramamurthy, A. Sen and A. P. Sinha, "Data Warehousing Infusion and Organizational Effectiveness," IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 38, no. 4, pp. 976-994, July 2008

[15] A. Sen, A. P. Sinha and K. Ramamurthy, "Data Warehousing Process Maturity: An Exploratory Study of Factors Influencing User Perceptions," IEEE Transactions on Engineering Management, vol. 53, no. 3, pp. 440-455, Aug. 2006.

[16] F. P. Pach, B. Feil, S. Nemeth, P. Arva and J. Abonyi, "Process-data-warehousing-based operator support system for complex production technologies," IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 36, no. 1, pp. 136-153, Jan. 2006.

[17] N. Bassiliades, I. Vlahavas, A. K. Elmagarmid and E. N. Houstis, "InterBase-KB: integrating a knowledge base system with a multidatabase system for data warehousing," IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 5, pp. 1188-1205, Sept.-Oct. 2003.

[18] Ling Feng and T. S. Dillon, "Using fuzzy linguistic representations to provide explanatory semantics for data warehouses," IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 1, pp. 86-102, Jan.-Feb. 2003.

[19] Y. Hu et al., "A Nonrelational Data Warehouse for the Analysis of Field and Laboratory Data From Multiple Heterogeneous Photovoltaic Test Sites," IEEE Journal of Photovoltaics, vol. 7, no. 1, pp. 230-236, Jan. 2017.

[20] J. C. Ferreira, J. de Almeida and A. R. da Silva, "The Impact of Driving Styles on Fuel Consumption: A Data-Warehouse-and-Data-Mining-Based Discovery Process," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 5, pp. 2653-2662, Oct. 2015.