ASSESSING THE PERSPECTIVE TRENDS ON ONGOING ISSUES AND CHALLENGES IN DATA WAREHOUSING

P. Amuthabala[1], Dr. M. Mohanapriya[2]

ABSTRACT

With the rising demands of the customers, the emphasis on data has been shifted from storage to analysis most recently. In this regards, data warehousing has played quite a contributory roles only till the last decade. With the evolution of data volume and complexity from multiple heterogeneous data sources, the conventional data warehousing encounters some of the problems which went unnoticed. This paper has discussed mainly the unsolved challenges and ongoing issues in data warehousing technique with an aid of existing literatures, trends of technologies, and demands of customers. The outcome of the paper shows that traditional datamining approaches suffer from various issues which were not addressed in past research work. This paper will encourage addressing the benchmarking issues, less compatibility towards real-time growing data, data integration, and ineffective techniques, there is a need to carry out investigation on this issues.

[1]*Research Scholar: Dept of CS & E, Karpagam Academy of Higher Education, Coimbatore, Tamilnadu, India*
Email: amuthabala79@gmail.com
[2]*Professor & HOD, Dept of CS & E, Karpagam Academy of Higher Education, Coimbatore, Tamilnadu, India*

Keywords : Analysis, Business Intelligence, Data Warehousing, Datamining, ETL, Storage.

1. INTRODUCTION

With the growth of the enormous database, the challenge is more on storage and analysis. However, with storage going cost effective, the challenge is more on analytics over data. The term data warehousing is a form of database management system where emphasis is not to store the operational data but to store the summary of the operational data. The prime purpose of storing the extracted knowledge from operational data is to assist in decision making for the stake holder of the organization [1][2]. All the data stored in data warehouse are static and read-only type, which means it cannot be changed by user [3]. The data is basically the extraction of the knowledge owing to implementation of certain data mining technique possessing the hidden information about the operational data. Therefore, it is also named as summarized data. The prime applications of data mining tools are essentially to explore some hidden traits of the data for predictive analysis of business process [4][5]. Various forms of business application tools are used to predict the performance of the

products for specific set of customers. One of such tool of data warehouse is also called as Business Intelligent tool, which performs evaluation and transformation of the business data to generate business intelligence [6][7]. Some example of business intelligent tools are OLAP (Online Analytical Processing), spreadsheet, data warehousing, data mining etc [8]. Various forms of open source tools e.g. JasperReports, R, Eclipse BIRT, SpagoBI, KNIME etc are used for same purpose [9], [10]. The domain of data warehousing is investigated in present manuscript as it is still shrouded with various unsolved issues still there are many research attempts started from 1999 to till dates. The first problem of data warehousing relates to under evaluation of resources involved in data cleaning and loading to warehouse. The problem also exists from the source system where there is inappropriate feeding of the data to the warehouse or adoption of inappropriate technique to process the data. This problem occurs owing to lack of understanding of data for future analytics. In software industry, the problem mainly occurs when the actual requirement of storage or analysis is not understood properly by the designer, which leads to incorrect selection of tools and techniques for capturing the data. This potentially leads to the generation of false positives. Even if the sources are quite different from each other, the prime technical concept of data homogeneity, revolve around data uniformity. There is a strong possibility that some potential and sensitive data might be eliminated or corrupted during the process of data integration if carried out in wrong manner. The most rising problem is massively increasing data, where the demands of the customers are more on real-time data and less on historical data. This is because the existing warehousing supports only historical data and not real-time data, thereby making it more challenging to integrate data warehousing, with the existing technologies of data analysis. There is also vulnerability of data ownership as the confidential data needs to be fed within the data warehouse in order to perform analysis on the data. The maintenance cost of data warehouses is quite high and any change in business processes invites maximum expenditure. Development of data warehouse always consumes more than 3-5 years, which surfaces a question of Return-of-Investment.

Therefore, the prime goal of this paper is to understand the existing trends and issues of data warehousing techniques in the forms of research work done in the past to explore unsolved issues. Section 2 discusses about the fundamental important facts about data warehouses followed by illustration of data quality in Section 3. Section 4 discusses about the existing tools while description of challenges of real time data is illustrated in Section 5. Approaches in data warehousing from related work viewpoint is discussed in Section 6 followed

by identification of research gap in Section 7. Finally Section 8 summarizes the paper by its concluding remarks.

2. Data Warehousing Essentials

The concept of data warehousing is an important part of modern database management system which is more focused on analyzing data rather than storing transactional data like relational database [11]. Therefore, it can be said that data warehouse always stores historical data which is extracted from streams of transaction-based data for analysis. The prime responsibility is to discretize the operation of analyzing the data from the operational database to allow summarization or consolidation of data from multiple different database sources.

A conventional or standard data warehousing system doesn't consider the direct inputs from multiple databases of different types. The existence of multiple database system (example-operational system, ERP, CRM, etc) which is then subjected to a specific operation called as ETL or Extraction, Transformation, and Loading. Basically, the term Extract refers to an operation that pulls or extracts the specific data from multiple data sources [12]. The term Transform refers to performing mathematical transformation of the stored data in precise format as well as appropriate data structure in order to normalize the process of querying / analysis. The term Load refers to loading of the transformed data to the final warehouse. Interestingly, all the 3 operation are performed in a same time owing to time consumption on data extraction process. However, the purpose of all the three operations is to perform an integration of data from the multiple data source. A standard data warehouse is compose of metadata, summary data, and raw data, which are subjected to multiple form of operations e.g. OLAP (Online Analytical Processing) Analysis, reporting, datamining. The staging area is used for constructing summary of the data and performs general warehouse management.

There are 4 types of standard system involved in data warehousing principle e.g. i) Data mart, ii) OLAP (Online Analytical Processing), iii) OLTP (Online Transaction Processing), and iv) Predictive analysis. A data warehousing system focused on single functional area i.e. marketing or sales or finance etc, and then it is called as data mart, which is the very simple form of data warehouse. OLAP is a process of replying to complex and multiple forms of queries [13]. It is commonly used in Business Intelligence applications by integrating conventional data source, reporting, and knowledge extraction. Some of the frequently used applications of OLAP are forecasting, management reporting, budgeting etc [13]. The next application is called as OLTP that is required to process the query of various short online commercial transactions. OLTP is primarily used in data mining and its database

consists of current transactional data. Response time in terms of number of transactions per second is the main performance parameter. The final system of data warehouse is to perform predictive analysis that is used for exploring the hidden patterns of the data. However, the concepts of OLTP as well as standard data warehousing system are different from each other in following respect.

2.1 Data Editing

Basically, ETL performs the periodic updating of the data warehouse in bulk. The end users are not permitted to update the data warehouse. However, the end users are permitted to update the individual data to the database periodically.

2.2 Workload

It is feasible for data warehouse to perform adhoc query. It is not feasible for data warehouse to visualize and estimate the amount of workload in advance. In OLTP, the application supportability is fine tuned for any workload.

2.3 Data Management

All the massive numbers of rows are subjected to read operation in a conventional data warehouse, however, OLTP performs processing only smaller amount of record.

2.4 Structure Supportability

The conventional data warehouse supports only de-normalized as well as partially de-normalized structure e.g. Star structure for the purpose of enhancing the query processing. OLTP system supports only normalized structure for enhancing the updating, inserting, and deleting operation for ensuring consistency of data.

2.5 Chronological Data

The age of the oldness of data stored in data warehouse is much higher than that of OLTP. For an example, a data warehouse can store all the product sales for more than decades of years, while OLTP only stores and processes last week or last month sales data.

There are couple of advantages of data warehousing system e.g. i) integration of data originating from multiple and different sources, ii) it resists the problem of data separation level block due to increasing number of data processing, iii) capability to maintain data history even in absence of original source of operational database, iv) enhancing the quality of data by furnishing the consisting codes with indexing and even repairing the corrupted data, v) furnishes a uniform platform of data processing inspite of origination from heterogeneous data sources that potentially assists in decision making from organizational level [14]. Hence, data warehousing plays an important role in enhancing data quality too. The next section will discuss about the quality of the data in data warehousing.

3. Data Quality in Warehousing

Ensuring highest quality of data in warehouse is critically important as it doesn't stored operational data but it stores valuable mined data. Therefore, the operation of data quality concentrated on spontaneous enhancement of data quality [15]. The operation of enhancing quality starts from processing the inputs from operational data to generating extracted knowledge. An important point to be noted in data quality management is that, it is performed along with the flow of the data. A typical operation of enhancing the data quality can be seen in Fig.1 that shows that metadata management is quite necessary to ensure data quality. The evaluation of the data quality depends upon the richness of metadata retaining transformed processes as well as applied data structure. In the Fig.1, it shows a series of ETL operation which leads to generation of highly processed data, which are reposited over metadata. Based on the set of notification rules, the user performs analysis of the data based on some set of rules, which are also called as constraints (based on various real-world dependencies). Some of the frequently used constraint formulations are as follows:

- Linkage of one value with another
- Dependency of rows of one table on another row of another table
- Existence of time-invariant value.

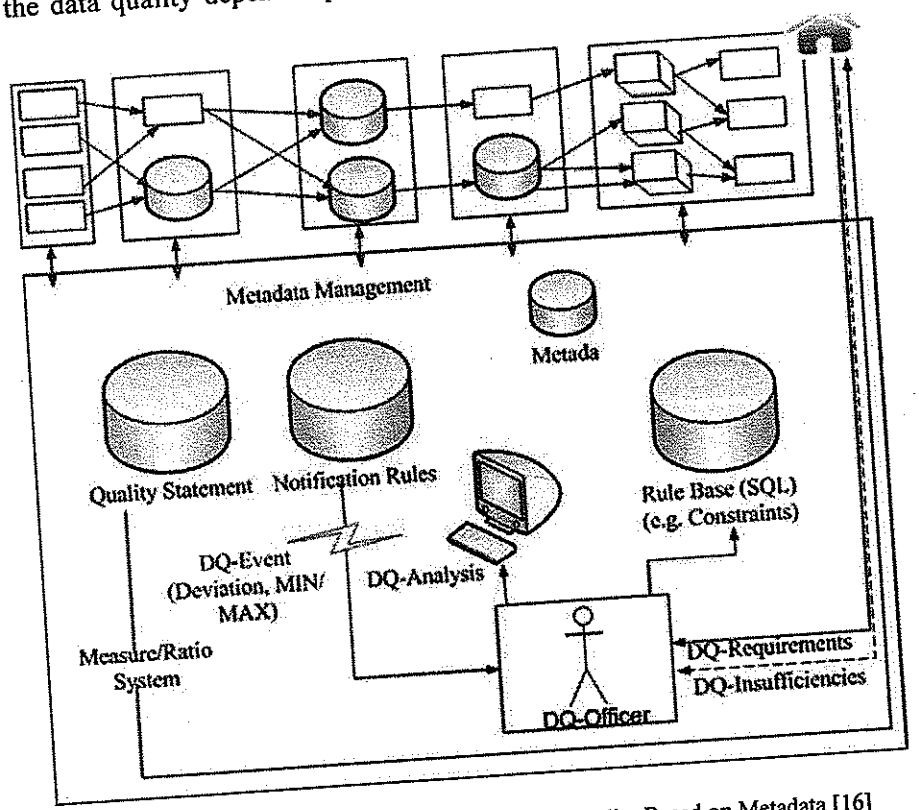

Figure -1 Standard Architecture of Data Quality Based on Metadata [16]

There are certain standards for testifying the quality of the data in data warehousing.

1.1 Information Product Map

This model was introduced by Shankaranarayanan [17] by deploying data from different sources and used a simple ETL process of numeral granularity degree. The deployment of this model is done based on source of data, processing of extracting information, noise cleaning process, process of integration, storing operation, a boundary of information system, consumer data, and monitoring block. This model was essentially introduced for maintaining quality of data pertaining to manufacturing sector.

1.2 AIM Quality Model

This model was basically the enhanced version of the model PSP/IP by Lee et al. [18]. The dimensioning process used in AIMQ is basically intended for bridging the gap of data quality as it provides both evaluations of data quality as well as empirical means to improve the quality of information. It was claimed that this model is known for 99% accuracy in terms of product quality in data warehousing. This model is also highly flexible to add new task along with existing one.

1.3 Data Model Quality

This standard of assessing quality of data was introduced by Moody [19]. The major goal of this standard is to understand the measuring factors of data quality inspite of declaring the factor of assessing when implicated to it. The testing of this model was carried out by the author considering real-time data.

1.4 Total Data Quality Management

This model of assessing data quality was introduced by Santos et al. [20] based on continuous data enhancement process. This framework was found to be widely accepted in industry and is slightly based on DMAIC process in Total Quality Management [21]. This model is extremely hypothetical in nature although its principle can be adopted in data warehousing.

1.5 Patient Assessment- Data quality Model

This model was developed by Donoghue et al. [22] and focused on enhancing patient related data in order to enhance the decision making in critical clinical condition for healthcare industry. The prime pillars of this data quality models are accuracy, timeliness of data, data completeness, and data consistency.

1.6 Data warehouse Development Life cycle

This model of assessing the data quality is presented by Kumar et al. [23]. The prime aim of this model is to address the importance of data quality to be done well in preliminary stage of data processing data warehouse and not at the end of the process.

The authors also claimed that considering of noise removal and various forms of data cleaning mechanism in the prior steps of data processing will significantly improve the data quality. The success of the model is based on an effective analysis of data, robust and practical planning of data, design and developing of data quality standards, implementation stage, and measuring or assessing the outcome of data quality.

1.7 De Lone and Mc Clean Model

The author De Lone and McClean [24] have introduced this model of enhancing the standards of data quality pertaining to data warehousing. It is designed based on four parameters i.e. quality of information, quality of system, quality of service, and quality of data inter-relationship. Apart from this there are various other models of data quality too. Each of them has focused on data classification, heterogeneity of data, data relationship, multidimensional data quality, etc

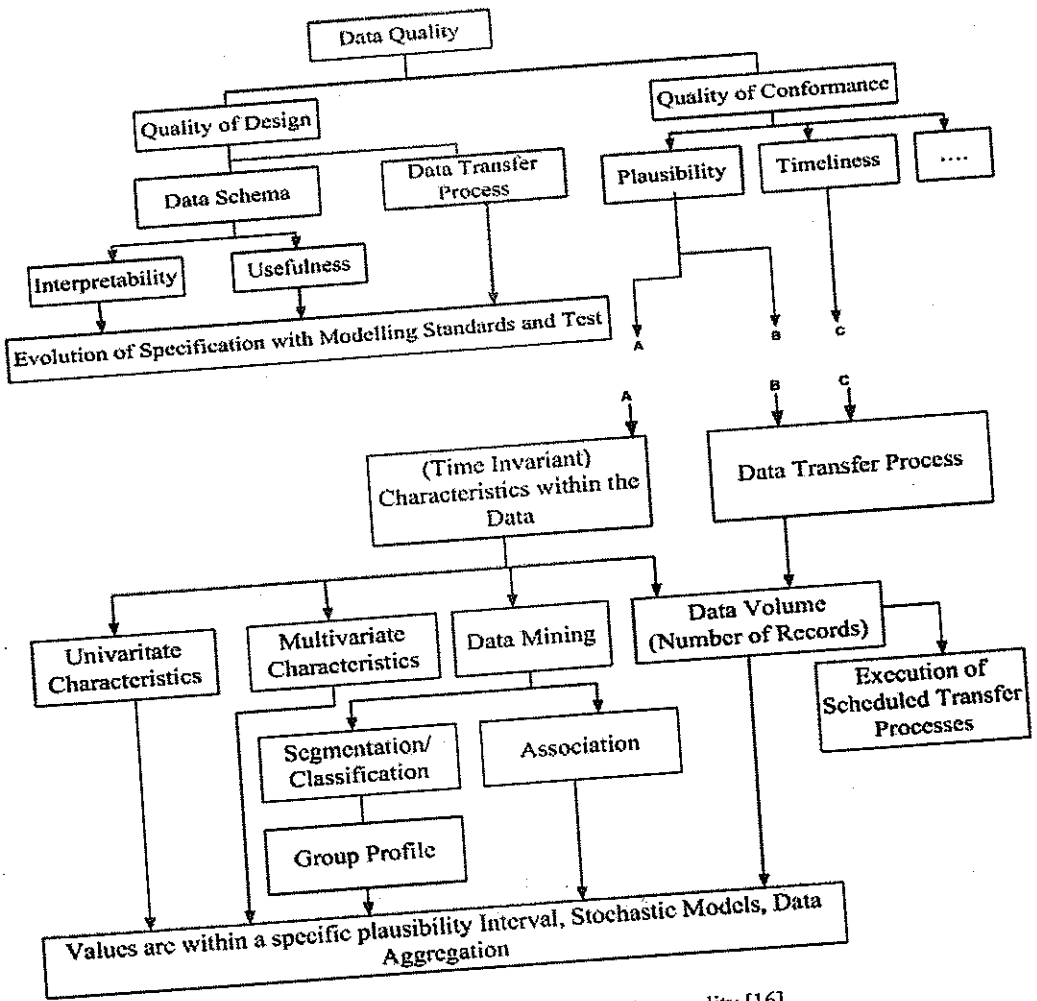

Figure - 2 Classification of study of data quality [16]

4 Tools of Datawarehousing

With more than a decade old, the data warehousing technology has no lack of tools. At present there are various tools that assist in performing data warehousing. The most frequently used tools of data warehousing are as follows [25] viz, Initio(AB), DT/ Studio(Embarcadero Technologies), Transformation Manager(ETL Solutions Ltd.), Elixir Repertoire(Elixir), Expressor Semantic Data Integration System(Expressor Software), DataFlow(Group 1 Software (Sagent)), DB2 Warehouse Edition(IBM), Cognos Data Manager (Formerly known as Cognos DecisionStream)(IBM), ETL4ALL(IKAN), Websphere DataStage(Formerly known as Ascential DataStage)(IBM), Data Migrator(Information Builders), Power Center(Informatica), CloverETL(OpenSys),Data Integrator (Formerly known as Sunopsis Data Conductor)(Oracle), Warehouse Builder(Oracle), SQL Server Integration Services(Microsoft), Data Integrator(Pervasive), Pentaho Data Integration(Pentaho),BusinessObjects Data Integrator(SAP),Data Integration Studio(SAS), Data Integrated Suite ETL(Sybase).

A simple scenario of tool usage in data warehousing with respect to Business Intelligence (BI) is shown in Fig.3. It intends to tell that it is important to understand what tool to use at what phase of data processing in data warehousing. The figure 3 shows different types of tool usage over source system, ETL layer, data and metadata repository layer, and presentation layer.

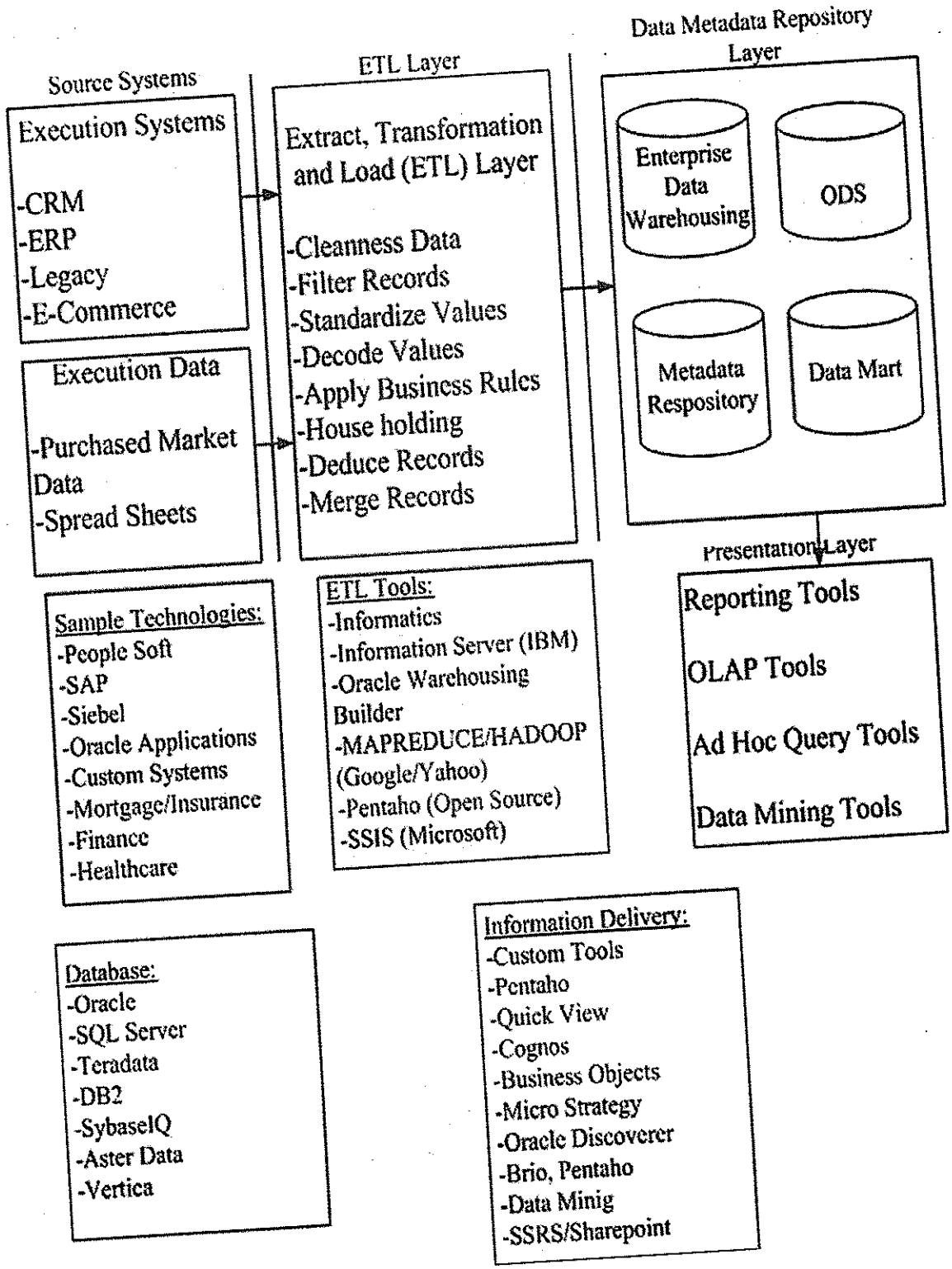

Figure - 3 Scenario of Tool usages in data warehousing

5 Challenges of Real-Time Data

By convention, the data warehouse doesn't store the present day data. Such data are formulated by storing it for a longer period of time ranges from weeks to years. Hence, theory of data warehouse is more and more inclined towards historical data, but there is not the thing which is required at present day. At present, owing to usage of various application and services, the dependency of data quality is more on dynamic database system rather than database system that emphasizes on historical data. For an example, a flight operator needs detailed information about passenger that has travelled most recently in case of suspicious event. In such case, much older historical data is of no use. Therefore, there is a bigger set of challenge for existing data warehousing system to address real-time data apart from historical data. Adoption of real-time data in the analytical modules in data warehousing could be highly utilized in detecting the fraudulent practices in enterprise applications or to understand the best need of customer based on their recent shopping behaviour etc. In order to support real-time application, the identified challenges are:

5.1 Designing Real-time Processing

The processing of the data is usually carried out by ETL operation, however, implementing ETL on real-time data is quite challenging although it supports simultaneous processing of data. It is known that majority of the ETL operations are carried out in batch mode, which considers the availability of data based on some specific update schedule. Only after getting the schedule, the ETL operation is carried out which has higher feasibility of data downtime barring the user access during data loading steps of ETL operation. Although, such downtime doesn't really effect the user as normally the data loading process happens at night time. Unfortunately, in the era of cloud computing which guarantees data availability, the process of loading real-time data will potentially affect legacy data warehousing process till date. So, when the user stays in different countries, there cannot be any downtime due to loading. So, there is a need of designing ETL tool that supports real-time data processing.

5.2 Developing Fact tables for Real-time Data

Owing to consideration of real-time data in existing data warehousing, the data modelling can undergo serious issues. For the data of older chronology, there is a significant time stamp over the data resulting in robust mining process. But for real-time data, the biggest challenge is to achieve a proper synching. For an example, when the aggregation of the heterogeneous real-time data is being carried out at multiple levels depending on a time factor, there is a higher possibility of not enough proper time for synching in the aggregated data in the warehouse.

This may cause outliers when subjected to datamining. There is also a possibility of ambiguity in data query processing for month-to-date or week-to-date that may potentially act in different manner with a half day of data where nature of the data changes continuously. In such case, neither it is possible to identify the best data structure for storage nor performing linking of the data in order to perform datamining over real-time data.

5.3 Developing OLAP Queries with Real-time Data

At present, the operations of every query processing as well as OLAP is meant to be working over historical data which is of read only nature. The challenging situation in this condition is after applying OLAP, the outcome is not subjected for any validation of data quality in case of data mobility that can occur only in real time data. The existing system only assumes that the data under observation is read-only type and so it is necessary to consider a backup plan for ensuring if the outcomes are adversely affected by the change of data concurrently owing to execution of query. Therefore, if the existing system of OLAP is subjected to dynamically changing data, it will result in data inconsistency and quite a ambigoius outcomes with presence of outliers. This problem is frequently encountered by the relational OLAP tools as they SQL based data process of multiple - pass. Usage of such technique results in two issues. The first issue relates to the fact that query outcome that consumes a specific duration of process cannot be considered to be precisely a real-time anymore. Such forms of latency of data may be acceptable for some application but they are never acceptable in case of finance prediction, meteorological forecasting, healthcare diagnostics etc. The second issue is that presence of data inconsistency when they are of real-time in nature. In such cases multiple passes of conventional SQL will also not be able to solve it especially if the data is of unstructured type. The problem becomes worst when real-time data for trend detection analysis is carried out.

5.4 Ensuring Scalability

The scalability of the existing system is measured with respect to quantity of the data being queried along with parallel execution of queries of same or other users in data warehouses. For a fixed size of data, the number of query response time is quite proportional to the quantity of the users online at the system. Therefore, response time becomes higher for more number of concurrent usages. Therefore, incoming real-time data packets will definitely increase the burden of spontaneously increasing data thereby limiting the scalability of the system.

5.5 Notification System

The existing notification system of the data warehouses is called report generation, which is

done by the automated email version of the report instantly after the data loading is carried out. However, it is quite a challenging task to generate a notification in the form of reporting system if the data are continuously changing its dimensions, types, formats etc. Especially, in the case of streams of big data, it is associated with higher volume, velocity, veracity, etc along with presence of semi and unstructured format of the data. Such data possess quite a challenging possess situation in terms of generating notification system.

Hence, it is seen that there are various issues of existing data warehousing technique to cater up the demands of the real-time data owing to much proliferation of ubiquitous computing and mobile networks. From more than a decade, there has been many research work in order to resist such issues. The next section will highlight some of the significant research attempts.

6 The Approaches in Data Warehousing

This section discusses about the existing research work that is focused on enhancing the performance of data warehousing. Majority of the existing approaches emphasizes on data quality using unique techniques. This paper involved in this section outlines the work of the discussion is based on various forms of approaches as discussed below:

6.1 Studies using Mining Approach

The existing system has witnessed the extensive usage of the datamining approach over the data being stored in warehouse repositories. A significant study towards emphasizing the benefit of datamining approach was discussed by Viktor and Motha [26] in 2002. The authors have also studied various techniques for assessing quality of the information. Rob and Ellis [27] have developed a mining approach exclusively for OLAP applications using conventional SQL based server. Taking a case study of OLTP, the study has developed a dimensional modelling using various tables, facts, and hierarchies of data using STAR schema. In 2009, Marketos [28] have investigated for issues of mobility of data with respect to mining approaches. The authors have introduced a trajectory data cube using conventional ETL methods in data warehouse. Tamilselvi and Gifta [29] have focused on the issues of mitigating adverse effects of duplicated data. The study has identified various records whose co-existence is

multiple in natures over the data table. Using threshold-based approach as well as certainty factor, the study has attempted various classifications of parameters. Finally, the study has used various parameters in order to formulate a rule for terminating the duplicated data from the data table using range factor, quantity of missed data, and lack of data quality, poor data representation, and threshold. The outcome of the study was evaluated with respect to reduction of duplicated records over increasing threshold values. Akintola et al. [30] have investigated on the OLAP tools pertaining to effective course management system using datamining tools. The study is almost similar to that of Rob and Ellis [27] in 2007; however, the authors have performed mining using association rule. Kochar and Chhillar [31] have discussed about a technique of normalizing data gathered from RFID-based applications. The authors have presented an empirical modelling to perform data cleaning, transformation, and loading based on java based environment. The outcome of the study was evaluated with frequency of occurrence and availability index. Faridi and Mustafa [32] have considered the case study of textile industry to perform investigation of mining approaches towards enhancing the decision making process. The authors have adopted empirical methodology to understand the significance of mining approaches. A simple design of data warehousing with mining approach is discussed by Bassil [33]. The authors have used MS-Access to develop the design pertaining to educational institution. In the modern data warehousing system, some of the research work was found to be inclined towards medical data. The research conducted by Feng et al. [34] has introduced a mining approach for investigation of heart disease. During the discussion, the authors have also discussed about various evolutionary techniques to performing mining. Essa and Christian [35] have presented a technique of performing mining approach for a specific type of queries of data. The authors have adopted an analytical schema to develop a mining approach over the data. A new form of mining can be seen in the form of BigData approach most recently. The studies conducted by Aydin et al. [36] have used various clouds computing modelling approach to perform mining of massively growing data. The authors have also used k-means clustering to perform mining approach for the large

and massive data over cloud. Mishra et al. [37] have presented a mining technique on sensory data over warehouse using three evolutionary techniques e.g. neural network, genetic algorithm, and fuzzy logic. The outcome of the study is benchmarked with respect to dimension matrix.

6.2 Studied for Enhancing Quality

The research work focusing on issues of data quality dates back on 1999. Rudra and Yeo [38] have presented a study focusing on the essentials parameters e.g. data inconsistency and data quality in data warehousing. The authors have used quantitative analysis to present their model based on Australian environmental data. Statistical parameters e.g. mean and standard deviation was used for evaluating the data quality. Gosain and Singh [39] have introduced a technique of enhancing data quality using goal-decision –information methodology in data warehousing. An interesting direction of research is being carried out by Almabhouh and Ahmad [40] have presented a study to find out the better mechanism of recognizing factors to ensure data quality. The authors have introduced 5 such data quality factors e.g. factors pertaining to organization, technical, project, environment, and infrastructure. All these 5 factors are also linked with 4 data quality parameters e.g. service quality, system quality, information quality, and relationship quality. A similar line of work is being carried out by Ali and Warraich [41] have presented a modelling for cleaning the Enterprise data for ensuring superior quality of information. The cleaning of the data modelling is performed using relational database system, Oracle warehouse builder, and PL-SQL. The investigation of this study was performed over data gathered from mobile network. The quality of data was checked with respect to number of rows, number of duplicated rows, standardization, nulls, garbage values, and lookups. Munawar et al. [42] have introduced a modelling of data warehouse using integrated analysis of requirements. The authors have presented various conceptual solutions to solve it. Idris and Ahmad [43] have focused on enhancing information quality pertaining to manufacturing industry. Similar line of study was also carried out by Sen et al. [44] where quality of data is associated with software engineering. Sidi et al. [45] have developed a comparative model for assessing the data quality. Based on the existing data quality models, the authors have presented a simple collaborative model of data warehousing for enhancing data quality. The issues of data quality model were also studied by Almeida et al. [46]. The authors have applied a quantitative analysis to understand the issues of data quality. Salim et al. [47] have proposed a study towards enhancing the data quality by focusing on pros and cons effect of requirement analysis. The authors have also put forward a practice evidence to enhance the data quality.

Table 1.Summary of the Mining Approaches

	Authors	Problem	Techniques Applied	Inference
MINING APPROACH	Viktor and Motha [26]	Challenges in mining	Qualitative discussion	Theoretical discussion
	Rob and Ellis [27]	Designing business intelligent	Analysis using SQL Server, OLAP	No comparative analysis
	Marketos [28]	Mining on mobility data	Trajectory based technique	No comparative analysis
	Tamilselvi and Gifta [29]	Duplicated data	Threshold-based duplicated data elimination	No comparative analysis
	Akintola et al. [30]	Course management	Analysis using SQL Server, OLAP	No comparative analysis
	Kochar and Chhillar [31]	Mining RFID data	Simple datamining technique with empirical approach	No comparative analysis
	Faridi and Mustafa [32]	Mining of textile data	Empirical method	No comparative analysis
	Bassil [33].	Mining of Educational data	MS-Access based warehousing	No comparative analysis
	Feng et al. [34]	Mining of Medical data	Qualitative discussion	Theoretical discussion
	Essa and Christian [35]	Mining of Infrastructure data	Analytical approach	No comparative analysis
	Aydin et al. [37]	Mining of massive cloud data	Computational Approach, k-means clustering, machine learning	No comparative analysis
	Mishra et al. [36]	Mining of sensor data	Evolutionary approach	Benchmarked, with better outcomes
DATA QUALITY	Rudra and Yeo [38]	Environmental data in Australia	Quantitative Analysis	No comparative analysis
	Gosain and Singh [39]	Data quality issues	goal-decision – information	No comparative analysis
	Almabhouh and Ahmad [40]	Exploring new quality factors	Analytical approach	Benchmarked, with better outcomes

	Ali and Warraich [41]	Data cleaning	Simulation-based study on mobile network data	Not effectively benchmarked
	Munawar et al. [42]	Requirement analysis	Conceptual modelling	No comparative analysis
	Idris and Ahmad [43]	Data Quality of manufacturing industry	Conceptual modelling	No comparative analysis
	Sen et al. [44]	Process maturity	Software engineering	Theoretical discussion
	Sidi et al. [45]	Comparative assessment of data quality model	Analytical modelling	Theoretical discussion
	Almeida et al. [46].	Multidimensional data quality	Quantitative approach	No comparative analysis
	Salim et al. [47]	Requirement analysis	Analytical modelling	No comparative analysis

7 Research Gap

All the literatures studied till date has their own advantages as well as limitations. The agenda of this part of the study was to find out those problems which were known and quite well-defined even during the times of existing research work, but it was found unaddressed. Therefore, this section will discuss about the research gap that has been derived by studying the existing trends and approaches of literatures.

7.1 Few Focus on Data Preprocessing and Data Generation

The discussion of the existing research attempts just start with a database, however, there is no evidence till date if the database has any real complications in order to evaluate the resiliency and robustness of the presented research idea towards process effectiveness. If the process is more focused on cleaning the data and making it more normalize before even subjecting it to ETL layer, the probability of data quality increases. However, no such attempts have been found.

7.2 Negligence towards Real-Time Data

Majority of the existing studies have used database to carry out the analytical processing in their studies, which is not applicable for real-time data processing application. Hence, existing research approaches towards mitigating real-time data is quite difficult to find. At present, the techniques are less to find that can mitigate the rising issue of semi/ unstructured data with massive volume, heterogeneity, uncertainty etc.

7.3 Very Less Emphasis on Data Quality

Majority of the work done till dates are conceptual, qualitative, or quantitative approaches to assess data

quality. Various problems (e.g. i) insufficient quality assurance, ii) incompatibility with real-time data, iii) integration and mining problems, iv) use of multiple way to represent data, v) insufficient mechanism of validation of data) are still unanswered from the solution presented most recently in data warehousing.

7.4 Less Efficient Data Quality Models

Section 3 and Section 6 has presented some of the standards of ensuring data quality in data warehousing. However, none of the existing system doesn't support timeliness of the matrices and neither considers problem complexity for classification. There is also less supportability of existing data quality approaches towards addressing structural complexity in the form of multiple dimensional data.

7.5 Lack of Benchmarking

It can be seen that majority of the existing research approaches towards data warehousing are found to discuss individual outcomes without any form of comparative analysis with other researchers. There is still no answer for the best work till date in this regards.

7.6 No Prototyping-based Approaches

All the approaches till date are non-prototyping based approach that doesn't consider adopting experimental-based research methodology. Therefore, reliability of the outcomes discussed (without benchmarked) is quite less for the existing literatures.

8 Conclusion

Data warehousing is highly important for decision makers and stakeholder in business. The paper has presented the fundamental discussion of the domain along with the problems associated with it in the modern era of pervasive computing and big data. After reviewing the existing tool, techniques, and unsolved issues, it is realized that resolving such issues is not an easy task. In the present era, with the evolution of datacenters, petabytes of data are being now stored in cloud clusters, which take the shape of operational database. Hence, the problems of performing data mining over the heterogeneous data are quite high. The structured data is arranged in row and column format and has specific data model. The same is not with unstructured data which is in much abundant in cloud computing. Organization like IBM have already started developing Big Data Analytics application with more focused on data mining and less focused on data storage. The mined data are only preferred to be stored in data warehousing, which poses a serious problem of incompatibility and various other issues illustrated in research gap of this manuscript. Hence in short it can be said that existing technologies of data warehousing is insufficient to meet the requirement of growing real-time data in present and

in upcoming times. Therefore, our future work will be to develop a unified model that performs heterogeneous data integration followed by mining and quality assurance.

9 References

[1] Terry H. Selected Readings on Database Technologies and Applications.2008.IGI Global Computers:564

[2] Ursino D. Extraction and Exploitation of Intensional Knowledge from Heterogeneous Information Sources: Semi-Automatic Approaches and Tools. Springer, Computers, 2003: 292

[3] Cios K.J. Pedrycz W. Swiniarski R. W. Kurgan L. Data Mining: A Knowledge Discovery Approach. Springer Science & Business Media: 606.

[4] Thierauf R.J. Effective Business Intelligence Systems. Greenwood Publishing Group Business & Economics, 2001: 370

[5] Zhou Z-H. Li H. Yang Q. Advances in Knowledge Discovery and Data Mining: 11th Pacific-Asia Conference, PAKDD, Nanjing, China, Proceedings. Springer Science & Business Media, Computers, 2007:1161

[6] Bulusu L. Open Source Data Warehousing and Business Intelligence. CRC Press,2012:432.

[7] Blokdijk G. Business Intelligence - Simple Steps to Win, Insights and Opportunities for Maxing Out Success. Emereo Publishing Social Science, 2015: 198.

[8] Agarwal B.B. Tayal S.P. Data Mining and Data Warehousing. Laxmi Publications, Ltd, Data Mining, 2009: 355

[9] Stuckenbrock S. Marktübersicht der Open Source Business Intelligence Systeme. diplom.de Computers, 2009: 47

[10] Rob P, Coronel C. Database Systems: Design, Implementation, and Management. Cengage Learning Computers, 2007: 704

[11] Jarke M, Lenzerini M. Vassiliou Y. Vassiliadis P. Fundamentals of Data Warehouses. Springer Science & Business Media. 2013: 195

[12] Pujari A.K. Data Mining Techniques. Universities Press, 2001: 288

[13] Rainardi, V. Building A Data Warehouse: With Examples In Sql Server. John Wiley & Sons, 2008: 540

[14] Hellerstein J.M, Stonebraker M. Readings in Database Systems. MIT Press Computers. 2005: 865

[15] Maydanchik A. Data Quality Assessment. Technics Publications, Business & Economics, 2007: 321

[16] Helfert M, and Herrmann C. Proactive data quality management for data warehouse systems. In DMDW, 2002: 97-106.

[17] Shankaranarayanan G. Towards implementing total data quality management in a data warehouse. Journal of Information Technology Management, 16, 2005: 21-30

[18] Lee Y.W. Strong D.M. Kahn B.K. and Wang R.Y. AIMQ: a methodology for information quality assessment. Information & Management, 40, 2002: 133-146.

[19] Moody D.L. and Shanks G.G. Improving the quality of data models: empirical validation of a quality management framework. Inf Syst, 28, 2003: 619-650.

[20] Santos G.D. Takaoka H and de Souza C.A. An Empirical Investigation of the Relationship between Information Quality and Individual Impact in Organizations. 2010.

[21] Janakiraman B. Gopal R. K. Total Quality Management: Text and Cases. PHI Learning Pvt. Ltd., Business & Economics. 2003: 260

[22] Donoghue J. O, Kane T. O, Gallagher J, Courtney G, Aftab A, Casey A, Torres J and Angove P. Modified Early Warning Scorecard: The Role of Data/Information Quality within the Decision Making Process. Electronic Journal Information Systems Evaluation. 14, 2011

[23] Kumar V and Thareja R. A Simplified Approach for Quality Management in Data Warehouse. International Journal of Data Mining & Knowledge Management Process (IJDKP) 3, 5, 2013.

[24] Delone W.H. and McLean E.R. The DeLone and McLean Model of Information Systems Success: A Ten-Year Update, J.Manage.Inf.Syst. 19, 2003: 9-30.

[25] "ETL Tools", Datawarehouse4u.Info, Link:- http://datawarehouse4u.info/ETL-tools.html, Retrived, 29th September, 2015.

[26] Viktor H. L. and Motha W. M. Creating Informative Data Warehouses: Exploring Data and Information Quality through Data Mining. InSITE - Where Parallels Intersect, 2002.

[27] Rob, A. Mohammad and M. E. Ellis. Case Projects in Data Warehousing and Data Mining. Issues in Information Systems VIII (1), 2007.

[28] Marketos G, and Theodoridis Y. Mobility Data Warehousing and Mining. In VLDB PhD Workshop. 2009.

[29] Tamilselvi J. Jebamalar, and Gifta C. B. Handling Duplicate Data in Data Warehouse for Data Mining. International Journal of Computer Applications. 15, 4, 2011.

[30] Akintola K.G. Adetunmbi A. O. and Adeola O. S. Building Data Warehousing and Data Mining from Course Management Systems: A Case Study of FUTA Course Management Information Systems. International Journal of Database Theory and Application, 4, 3, 2011: 13-24.

[31] Kochar B. and Singh R.C. An effective data warehousing system for RFID using novel data cleaning, data transformation and loading techniques. Int. Arab J. Inf. Technol, 9, 3, 2012: 208-216.

[32] Faridi M.S. and Mustafa T. Usability of data warehousing and data mining for interactive decision making in textile sector. Global Journal of Computer Science and Technology,12, 7, 2012.

[33] Bassil Y. A Data Warehouse Design for A Typical University Information System. arXiv preprint arXiv:1212.2071. 2012.

[34] Feng Y. Wang Y. Guo F. and Xu H. Applications of Data Mining Methods in the Integrative Medical Studies of Coronary Heart Disease: Progress and Prospect. Evidence-Based Complementary and Alternative Medicine, 2014.

[35] Essa R. A., Bach A., Christian. Data Mining and Warehousing. ASEE Zone I Conference, University of Bridgeport, Bridgpeort, CT, USA, 2014.

[36] Mishra N. Chang H-T. and Lin C-C. Data-centric knowledge discovery strategy for a safety-critical sensor application. International Journal of Antennas and Propagation, 2014 .

[37] Aydin G. Riza Hallac I. and Karakus B. Architecture and Implementation of a Scalable Sensor Data Storage and Analysis System Using Cloud Computing and Big Data Technologies. Journal of Sensors, 2015.

[38] Rudra A. and Yeo E. Key issues in achieving data quality and consistency in data warehousing among large organisations in Australia. In Systems Sciences, HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference, 1999:8

[39] Gosain A. and Singh J. Achieving data warehouse quality using gdi approach. In Applications of Digital Information and Web Technologies, ICADIWT, First

International Conference on the, 2008: 494-499.

[40] Mabhouh A.l. Eddin Al. and Ahmad A. Identifying Quality Factors within Data Warehouse. In Computer Research and Development, Second International Conference, 2010: 65-72

[41] Ali K. and Warraich M. A. A framework to implement data cleaning in enterprise data warehouse for robust data quality. In Information and Emerging Technologies (ICIET), 2010 International Conference, 2010: 1-6.

[42] Salim N. and Ibrahim R. Towards data warehouse quality through integrated requirements analysis. In Advanced Computer Science and Information System (ICACSIS), International Conference, 2011: 259-264.

[43] Idris N. and Ahmad K. Managing Data Source quality for data warehouse in manufacturing services. In Electrical Engineering and Informatics (ICEEI), International Conference, 2011: 1-6.

[44] Sen A. Sinha A.P. and Ramamurthy K. Data warehousing process maturity: An exploratory study of factors influencing user perceptions. Engineering Management, IEEE Transactions,53,3, 2006: 440-455

[45] Sidi F. Ramli A. Jabar M. Suriani Affendey L. Mustapha A. and Ibrahim H. Data quality comparative model for data warehouse. In Information Retrieval & Knowledge Management (CAMP), International Conference, 2012: 268-272.

[46] Almeida G.D. Wesley, R. T. D. Sousa, F.E.D. Deus, G.D. A. Nze, and F. Lucio, L. D. Mendonca. Taxonomy of data quality problems in multidimensional Data Warehouse models. In Information Systems and Technologies (CISTI), 8th Iberian Conference, 2013: 1-7

[47] Salim N. and Ibrahim R. Quality-based framework for requirement analysis in data warehouse. In Advanced Informatics: Concept, Theory and Application (ICAICTA), International Conference, 2014: 152-158.

AUTHORS BIOGRAPHY

P. Amuthabala has obtained her B.E in Computer science Engineering at Avanishilingam University in Coimbatore, Tamilnadu, India in 2002 and her M.E degree in Software Engineering at Bangalore University in Bangalore, Karnataka, India in 2011.She is working as an Assistant Professor in Information Science Department at Atria Institute of Technology, Bangalore, Karnataka. Her research areas of Interest include Data Mining Data warehousing and Cloud Computing. She is currently doing her PhD in Karpagam University.

Dr. M. Mohana Priya is currently the Professor and Head of the Department of CSE in Karpagam University. She completed her B.E CSE in the year 2002 at Sri Krishna College of Engineering and Technology, Coimbatore. She had completed her M.E CSE in Crescent Engineering College, Chennai. She holds a Doctorate Degree from Anna University, Chennai. She has more than 12 years of teaching experience to her credit. She has achieved best faculty award and the best paper award in the year 2013. She has pub lished more than 7 papers in International Journals and Conferences. She is also a reviewer of reputed journals namely, 'IEEE Transactions on Mobile Computing, Computers and Electrical Engineering, Elsevier Publications and Arabian Journal of Science and Engineering, Springer.