International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 11 (2019) pp. 242-254

Received: 20 April, 2019; Accepted 8 Jan, 2020; Published: 29 Jan, 2020

An Approach for Implementation of Cost Effective Automated Data Warehouse System

Sachin Sharma[1], Kamal Kumar[2] and Sandip Kumar Goyal[3]

[1] Computer Science Department, MMD University,
Mullana, Ambala, Haryana, India
er.sachinsharma@gmail.com

[2] Computer Science & Engineering Department, National Institute of Technology
Srinagar, Uttarakhand, India
kamalkumar@nituk.ac.in

[3] Computer Science & Engineering Department, MMD University,
Mullana, Ambala, Haryana, India
hodce@mmumullana.org

Abstract: Technology has resulted in humongous growth in generation of data and need of Data Warehouse. The performance of business decisions depends upon the optimality of the operations performed on data and supporting data warehouse system (DWH). Mostly, data warehouse system design methodologies focus on Extraction, Transformation and Loading (ETL) and ETL processes. Big Data paradigm is emerging as a big challenge to ETL and causing performance degradation and frequent re-configuration to meet day to day operational tasks. **Recent research claims** on automations of ETL processes **have** emerged as possible solution in Big Data paradigm. Through this proposal, data warehouse system process is considered as a complete unified tool covering all intermediate steps **from source data** to report generation; for analysis and decision making. All steps such as, source data extraction automation, transformation & loading automation and reporting automation are attempted. **Cost effective data warehouse system solution is proposed by avoiding use of commercial tools without compromising on performance.**

***Keywords*:** DWH, Automation, maintainability, ETL, Data Science, Analytics

I. Introduction

Data is backbone of every decision. A trending quote by Clive Humby establishes the spirit of statement [1].

"Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value."

The soul of this statement is to give a pathway for data scientists to consider data as most valuable asset for every organization. Data exists everywhere, but we need to understand data in holistic view – Why to store? What to store? How to Store? Where to store? What am getting out of it? Is the results aligned with the organizational goals? When we think about these questions and then start planning for the strategy, it will lead to the excellent results. In-fact, why data is stored is really driving argument. Stored data allows to have insight of data and to take the necessary analytics followed by actions which are based on the results. Ever increasing volume of data is not as simple, as it is coming all the way in the forms of structured data and unstructured data. Structured data is business data having well defined structured which generally supports "Schema on Write" paradigm. Structured data can easily be stored in the relational database management systems whereas for unstructured data, there is a requirement of many tools and utilities like Hadoop and No-SQL database management systems.

Big data is an emerging field and interchangeably used with heavy data sets. The definition says "Big data is a collection of large or complex data sets that cannot be processed by traditional data processing applications and tools". Data warehouse is a term used to describe container where data is stored. Relational database management systems are used for storing structured data. Recently evolved framework like **Hadoop is used to store large data sets using distributed processing by utilizing programming models.** The term Big Data is frequently associated with the unstructured data. Data analytics act on the top of the data warehouse for decision making. **Both relational data warehouse system and big-data based data warehouse system has unique properties of their own. Relational data warehouse system can only handle structured data whereas big-data based data warehouse system can handle unstructured and semi-structured data.** In-fact both data-warehouse systems (Relational and Big Data) should work in tandem with each other so as to get the decision metrics. Although there is a lot talking about Big-data and its usage, there is indeed a need to discuss big data as nearly 80% of the data is unstructured data and is like raw crude oil, but the importance of the relational data (structured data) cannot be ignored.

Data has to be refined, intermediately stored for processing, adding some other ingredients for desired product, different variants passed to different apartments and finally got the desired product. The desired product adds value to organization or help making decision for the better productivity.

Extraction, Transformation and Loading (ETL) converts data into useful and desired forms. **ETL process is a sub process to design the data warehouse system and can be performed with either commercial tools or by customized programming at OS and database level.**

Post ETL operations, data is available in data warehouse and meaningful data can be retrieved out of it using metric based logical reporting. Reporting is considered an additional area over and above data warehouse system and ETL, and is a key factor towards decision making. An efficient and in-time reporting serves its purpose for maximum profitability of any organization so should be dealt with very carefully. The reporting from DWH should be well in-line with the data warehouse system. When it comes to the large data sets, data processing in data warehouse system goes on as ongoing activity; heavy and un-optimized reporting mechanism can have a severe impact on data-warehouse.

Researchers have put in their efforts to build up a data-warehouse, integrate the processes in [1] to have a view of various data warehouse systems, and in [2] data warehouse; problem has been taken as big data problem and various tools have been deployed for solution. Many researchers have put in their efforts in the field of data warehouse and its processes by their own ontologies.

It has been observed that, the whole process of data warehouse system till reporting has not been viewed as unified tool which should take an input and give output in the most desired way. All sub-processes should be aligned and well integrate with minimum manual intervention. The business user/technical users of the framework need not necessarily know the internal technical details of the system and can focus on the business result. There is a need of relational data-warehouse framework which is completely automated in nature from source data extraction, transformation, loading till optimized reporting. **Data retrieval refers to extracting data from data warehouse as per business requirement.** The whole process can be easily achieved using various commercial utilities and tools which need heavy investment and license cost.

In this research paper, we have provided an automated framework to set up the relational data warehouse system for organization of any size without use of the commercial tools and utilities. The implementation can be done in any programming language of programmer's choice with knowledge of database systems. The complete automated data warehouse system is categorized in mainly three parts – source data extraction automation, data transformation and loading automation, data retrieval automation.

Success of a data warehouse project depends heavily on smoothness of the processes incorporated in it post deployment on day to day basis. Not only development of data warehouse system should be focused, but stability of its operations, process automation and tuning should be focused as continuous activity during whole life span of the data warehouse project. A data warehouse process consists of data warehouse processes and data warehouse itself.

Setting up a data warehouse is a tedious task for small and medium scale industries due to cost and complexity involved in maintaining the system. Motivations of this study is to put in place a robust data warehouse system architecture with complete automation and cost effectiveness with minimum manual intervention required for maintenance. In this paper, process automation and maintainability issues are addressed for data warehouse project without use of commercial tools. Study is oriented towards providing a system that is self-sustainable and self-healing for handling hick-ups during data processing. **Self-sustainable system is a system where less manual intervention is required and** self-healing system refers to a system where majority of abnormality during operations are handled by **the** system itself. These properties are key ingredients for **automating the** system.

Remainder paper is organized as follows: Section II reviews a wide range of related works in the realm of data warehouse system for automation. Section III gives brief detail on preliminary for data warehouse system automation. Section IV describes the problem definition. In section V, proposed approach, methods and materials are given respectively. The performance and results are presented in Section VI. Finally, Section VII summarizes the conclusions and future directions of this work.

II. Literature Review

NoSQL based data warehouse system is proposed which contains social networking data. Automated transformation rules are used for transition of conceptual database to NoSQL database. The automated transformation rules are evaluated using TPC-DS benchmark [1]. Document based NoSQL data stores are used for NoSQL based data warehouse system approach. Agile development approach is utilized to develop data warehouse system. Cross platform de-normalization technology is harnessed for building data warehouse system [2]. Big data warehouse refers to a data warehouse where complex, huge and variety of data is stored using big data technologies. Hadoop is proposed as alternative for building big data warehouse system [3]. Massively parallel processing (MPP) system is proposed for building scalable data warehouse system. MPP database and data warehouse systems are designed. The MPP enabled data warehouse system is evaluated using IBM Datastage ETL tool. MPP enabled data warehouse system is applied in telecommunication domain [4]. With emerging need to process unstructured data, non-relational database management systems (RDBMS) are required. Many legacy and current application uses RDBMS as database and now same applications are expected to handle unstructured data. There is heavy cost involved in replacement of applications to use non- relational database, a mechanism is required for enabling data transformation and keep application compatibility with database. An automated framework is presented for migration RDBMS data to Not only SQL (No-SQL) based database with compatibility of application with NoSQL database [5]. Data preparation is considered as a processing steps for data including data discovery, data extraction, data profiling, format transformation, source selection, matching, mapping, data repair, duplicate detection and data fusion. Three questions are considered for data preparation automation a) can we automate? b) Should we automate? c) Must we automate? [6] To achieve automated big data ETL framework, an extendable

ETL framework is presented to address issues associated with big data. For big data ETL, user defined function (UDF) component is introduced. UDF component enables ETL developers with reusable algorithmic layout and cost model generates effective ETL workflow execution plan [7]. Continuous or real time automated auditing using audit data warehouse system and data marts is presented. Audit data warehouse and data marts contains near real time records of financial transactions to enable auditors to achieve continuous auditing [8].

Authors have proposed a framework for federation data warehouse and further proposed algorithms for pre-processing, query decomposition and results [9]. The ultimate aim is to build up a system up and above various data warehouse systems and called them federation data warehouse. The group of data warehouse federation system containing semantic information and users has a unified view for various data warehouses. Authors have been able to achieve data generalization and give a prototype of their approach by making an application to use the same.

Proposal has identified data warehouse problem as big data problem in which data is increasing and focused on speed and number of users [10]. Authors have proposed a new data warehouse system framework named General Engineering and Manufacturing Data Warehouse (GEM-D) in which highly effective data gathering techniques are deployed so as to achieve "just in time" data gathering. To achieve the objectives, authors have proposed a model having source systems, staging layer, logical layer, analysis tools and Business Intelligence (BI) tools and reporting tools.

Researchers have put in their efforts and addressed **an issue to design** data center for tobacco industry. The data center focus on effective data mining and data warehouse system techniques [11]. The aim is to provide a system specific to data warehouse and data mining which can facilitate trend analysis and future forecasting. Authors have proposed a framework which is production process control centric so as to improve the processes based on data mining and data warehouse system techniques.

A recent work has proposed an improved model for enterprise data warehouse storage system named, "UStore" [12]. The storage system aimed at problem of data storage around 3 peta bytes in China's bank card enterprise "UnionPay". The "UStore" framework is having key features of Storage Unification, query efficiency and user transparency. It is based upon single physical table multiple view architecture which facilitate efficient and effective query execution. Ustore also helps in handling data redundancy without any changes in existing query logic.

Another work put(s) emphasize on the scalability aspect of big data sets without investing much in the hardware improvements [13]. Boolean queries are mapped to SQL for clinical research data warehouse system. Custom query generation has been used and compared with default query generation in clinical research data warehouse system. Authors have compared execution time for SQL queries generated by custom Common Table Expression (CTE) generator and default SQL query generator. Authors have observed significant improvements in results for CTE based scalable solution without costly hardware upgrades.

Peer Review Markup Language (PRML) for education sector data warehouse system focused on peer assessment [14]. Authors have addressed the difficulties while comparing the functionalities of two different peer review systems supporting different protocols. Authors have put in their efforts to generate a generic data set and then comparing the different peer assessment systems using Peer review markup language.

A comparison among different modeling techniques among Kimball data modeling approach, Inmon subject modeling approach and data vault approach for scalability and flexibility analysis was attempted [15]. The quality aspects of data warehouse system upon context based approach were focused recently [16]. Authors in [17] have tried to address need of small and medium enterprises. A data warehouse system model that is based on service oriented architecture has been proposed.

Data warehouser is like an artist who can examine data warehouse schema pictures in detail [18]. Authors have suggested that a good data warehouser should have in depth knowledge to implement data warehouse. Authors have put efforts in the field of higher education by business intelligence framework [19]. Emphasis has to be on extraction, loading and transformation along with data integration to develop and manage higher education centric business intelligence solution. Authors have proposed a data warehouse system framework aimed at decision support system for Egypt's tourism sector. The data warehouse can be used by multiple agencies / entities to decide on their key decisions and help to mitigate the risk [20]. A quality driven data warehouse system should be objective and keep quality aspect in the center of the ETL processes while building data warehouse [21]. The quality approach is to look for corrupt data before loading and approach is validated through a case study using Oracle Semantic database source. Particle swarm optimization is focused with respect to swarm intelligence [22]. Basis data mining terminologies are presented with past and ongoing work of swarm intelligence in data mining. Data mining framework is presented for mining weather data [23]. Self-organizing data mining approach called enhanced Group Method of Data Handling (e-GMDH) is presented.

In this paper, we discuss a framework to design an automated data warehouse system at optimized cost. Although, there are many challenges and issues that data warehouse system needs to handle, here we only concentrate on issues related to automating data warehouse system with cost effectiveness. In this context, several parameters need to be considered. The parameters are cost effectiveness, time taken for data processing, self-healing property of system and retrieval time metrics. **These issues drive the choice of data warehouse system framework, which determines how efficiently organization utilizes its data.**

To address the cost effective automation problem of data warehouse system we need to split data warehouse system into sub processes and work on their automation. First, data warehouse system is split into three processes for automation and then, automate individual processes. Existing techniques and tools are not specifically tailored to accommodate the need of an automated data warehouse system without taking help of commercial tools in data warehouse system development and its reporting.

Reporting automation has not been considered as integral part of data warehouse system development process.

Table 1. Summary Table of Existing Work

No.	Title of Contribution	Published In	Major Contribution
1	Automatic transformation of data warehouse schema to NoSQL data base: comparative study	Procedia Computer Science. 2016 Jan 1;96:255-64.	Automated transformation rules are used for transition of conceptual database to NoSQL database. The automated transformation rules are evaluated using TPC-DS benchmark [1].
2	Towards nosql-based data warehouse solutions	Procedia Computer Science. 2017 Jan 1;104:104-11.	Cross platform de-normalization technology is harnessed for building data warehouse system [2].
3	Research in big data warehousing using Hadoop	Journal of Information Systems Engineering & Management. 2017;2(2):10.	Hadoop is proposed as alternative for building big data warehouse system [3].
4	Implementation of database massively parallel processing system to build scalability on process data warehouse.	Procedia Computer Science. 2018 Jan 1;135:68-79.	Massively parallel processing (MPP) system is proposed for building scalable data warehouse system.
5	A framework for migrating relational datasets to NoSQL.	. Procedia Computer Science. 2015 Jan 1;51:2593-602.	An automated framework is presented for migration RDBMS data to Not only SQL (No-SQL) based database with compatibility of application with NoSQL database [5].
6	Automating Data Preparation: Can We? Should We? Must We?.	http://ceur-ws.org/Vol-2324/Paper00-InvTalk2-NPaton.pdf	Three questions are considered for data preparation automation a) can we automate? b) Should we automate? c) Must we automate? [6]
7	Next-generation ETL Framework to Address the Challenges Posed by Big Data	InDOLAP 2018.	UDF component enables ETL developers with reusable algorithmic layout and cost model generates effective ETL workflow execution plan [7].
8	Continuous Auditing: Building Automated Auditing Capability.	InContinuous Auditing: Theory and Application 2018 Mar 8 (pp. 169-190). Emerald Publishing Limited.	Audit data warehouse and data marts contains near real time records of financial transactions to enable auditors to achieve continuous auditing [8].
9	A framework for data warehouse federations building	Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, 2897–2902. https://doi.org/10.1109/ICSMC.2012.6378233	Authors in [9] have proposed a framework for federation data warehouse and further proposed algorithms for pre-processing, query decomposition and result.
10	Heading Towards Big Data	Better, B. A., Warehouse, D., More, F., Speed, M., Users, M., & Goss, R. G. (2013). 220–225	Authors in [10] have identified data warehouse problem as big data problem in which data is increasing and focus on speed and number of users.
11	Research and Design Data Center for Tobacco Based on Data Warehouse and Data Mining	Wenzhi, W., Qide, H., Lin, L., &Xinqin, H. (2016). 166–170	Researchers have put in their efforts and addressed an issue of to design data centre for tobacco industry in China aimed at data mining and data warehouse [11].
12	An optimized storage system for enterprise data warehouses at UnionPay.	Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016, 1574–1578. https://doi.org/10.1109/BigData.2016.7840766	Authors in [12] have proposed a improved model for enterprise data warehouse storage system named “UStore”.
13	Using Common Table Expressions to Build a Scalable Boolean Query Generator for Clinical Data Warehouses	IEEE Journal of Biomedical and Health Informatics, 18(5), 1607–1613. https://doi.org/10.1109/JBHI.2013.2292591	Authors in [13] put emphasize on the scalability aspect of big data sets without investing much in the hardware improvements.
14	A markup language for building a data warehouse for educational peer-assessment research.	Proceedings - Frontiers in Education Conference, FIE, 2016–Novem, 0–4. https://doi.org/10.1109/FIE.2016.7757600	Authors in [14] have proposed Peer Review Markup Language (PRML) for education sector data warehouse focused on peer assessment.
15	Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault	2016 International Conference on System Reliability and Science, ICSRS 2016 - Proceedings, 95–99. https://doi.org/10.1109/ICSRS.2016.7815845	Authors in [15] have compare different modeling techniques among Kimball data modeling approach, Inmon subject modeling approach and data vault approach for scalability and flexibility analysis.
16	Data Quality in Data Warehouse Systems	Serra, F., &Marotta, A. (2016).	Authors in [16] have focused on the quality aspects of data warehouse upon context based approach.

maintaining cost effectiveness of overall system. Majority of research is carried out by focusing on ETL and its workflow. ETL is an inevitable sub process of a data warehouse system but it is not a data warehouse. Data warehouse system is a combination of many sub systems

17	Building a Data Warehousing Infrastructure based on Service Oriented Architecture.	Proceedings of 2012 International Conference on Cloud Computing, 82–87. https://doi.org/10.1109/ICCCTAM.2012.6488077	Authors in [17] have tried to address needs of small and medium enterprises. A data warehouse model that is based on service oriented architecture has been proposed.
18	Datawarehouser: A data warehouse artist who have ability to understand data warehouse schema pictures	IEEE Region 10 Annual International Conference, Proceedings/TENCON, 2205–2208. https://doi.org/10.1109/TENCON.2016.7848419	Authors in [18] have proposed a view point of data warehouser. Data warehouser is like an artist who can examine data warehouse schema pictures in detail.
19	Significance of data integration and ETL in business intelligence framework for higher education	Proceedings - 2015 International Conference on Science in Information Technology: Big Data Spectrum for Future Information Economy, ICSITech 2015, 181–186. https://doi.org/10.1109/ICSITech.2015.7407800	Authors in [19] have put efforts in the field of higher education by business intelligence framework.
20	Building Data Warehouse System For The Tourism Sector.	Conference Paper ·December 2015, (January), 0–8.	Authors in [20] proposed a data warehouse framework aimed at decision support system for Egypt's tourism sector.
21	A Quality-Driven Approach for Building Heterogeneous Distributed Databases: The Case of Data Warehouses	2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), Cartagena, 2016, pp. 631-638.	A quality driven data warehouse system should be objective and keep quality aspect in the center of the ETL processes while building data warehouse [21].
22	Swarm intelligence in data mining	InSwarm Intelligence in Data Mining 2006 (pp. 1-20). Springer, Berlin, Heidelberg.	Particle swarm optimization is focused with respect to swarm intelligence [22].
23	Self-organizing data mining for weather forecasting.	InIADIS European Conf. Data Mining 2007 Jul (pp. 81-88).	Data mining framework is presented for mining weather data [23].

III. Preliminaries on Data Warehouse Automation

In general, data warehouse system automation takes input in the form of heterogeneous data from multiple data sources and output is unified data available at destination source. Destination source may be database or flat file or other data repository type. The resultant data is utilized by different verticals like decision making, analytical techniques, business intelligence and others. The resultant output of data warehouse system is unified data. Further use of this data depends upon different business unit requirements. Data warehouse system is considered implemented if unified data is available at production fact tables. Beyond it, reporting/retrieval is considered as next phase after data warehouse system development. To achieve data warehouse system automation, ETL job automation and workflow optimization is considered. In next section, problem definition is detailed.

IV. Problem Definition

On reviewing literature in the realms of self-computing and data warehouse, only a few works have tried to address issues related to automation of data warehouse system with retrieval mechanism and at the same time mainly of extraction, transformation and loading. In related

literature, research is performed to shift paradigm towards fully automated ETL framework. Automation of data warehouse system can be achieved as a result of sub systems automation and binding of all sub system seamlessly. Research is carried out to automate migration of RDBMS to NoSQL. Extendable ETL framework is used to achieve automated big data ETL framework. Automated auditing is examined for research by making use of near real time data warehouse. ETL workflow execution plan is utilized for automation. It is observed that, ETL tools (like IBM Datastage or Informatica) are used frequently for data warehouse system automation. ETL workflow automation is interchangeably considered as data warehouse system automation. When ETL tools step- in during data warehouse system development, it becomes costly affair due to cost associated with the tools. To manage the cost, NoSQL based databases, big data approach and open source technologies are worked upon with ETL tools. There are some strong points of RDBMS and specific strengths are associated with NoSQL based databases. Both type of databases have specific purpose and utilities. RDBMS cannot be replaced by NoSQL

databases and vice versa but can be used together to achieve specific goals.
End result of data warehouse is reporting. Reporting and retrieval are used interchangeably. Separate tools are available for ETL and reporting. Sometimes, ETL tools and reporting tools are part of an enterprise business suit. For reporting (which is an ultimate aim for DWH), an organization needs to either procure reporting tool or build in-house reporting application for specific reporting needs. There is a need of integrated data warehouse system framework which keeps in mind reporting requirement from beginning of the data warehouse system development process. Data retrieval handle organization's specific reporting requirements and data warehouse is designed in such a way that reporting is quick and just in time.
By keeping in view thoughts and studies in relevant field, some gaps are identified towards cost effective automated data warehouse system goal which are mentioned as:
a) Not only ETL process automation, but overall data warehouse system process automation needs consideration for research.
b) Data warehouse system process should consider reporting/retrieval process as integral part in addition to ETL process.
c) Explore possibilities to develop automated data warehouse system with inbuilt utilities to avoid commercial ETL and reporting tools.
These issues are addressed in coming sections of this paper.

V. Proposed Approach: Methods & Materials

Data warehouse consists of three major processes which are:
Data extraction from source,
Transformation and
Loading.
Another term "Retrieval" is added to the data warehouse development process which contributes to decision making. The processes to automate are identified and are segregated for seamless and efficient automation.
a) Source Data Process Automation,
b) Data Transformation and Loading Automation and
c) Data Retrieval Automation.

The sub processes automation is optimized and integration of sub processes is carried out in closely coupled way for desired output. The individual sub processes improved automation is explained in following subsections. The flow of sub systems automation is shown in Figure 1.

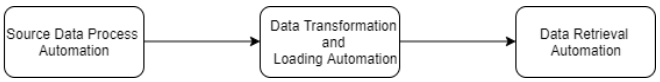

Figure 1. Sub Systems Automation Flow

Source Data Process Automation:
Data warehouse data originates from different data sources having heterogeneous technology platforms. ETL processes are involved to build up a data warehouse. Extraction means extract data from different sources and give it as feed to transformation engine. In this section, we focus at automating the whole process of data extraction using various tools and utilities. We propose a mediation system, which interacts with various data sources having data in different formats like hexadecimal formats, binary format or any other device specific formats. The mediation system is a dedicated system having specialization in handling incoming data streams and subsequently converts it into desirable format and can output the multiple streams to different staging locations in parallel. Error handling has been incorporated as an integral part of mediation system that detects logical expression and data quality issues in incoming data streams. Erroneous data is identified and reprocessed to ensure complete data integrity for staging area. Filtering logic is implemented to obtain selective data at staging area and eliminating other data. The parsers are integral part of mediation system which convert data into desirable format. The required data streams generated from mediation are fed to staging area at specific locations in raw format; so that they may be available for transformation and further for loading. The process becomes automated using parsers in mediation system which can be coded in C language on Unix/Linux platform. The mediation system is getting continuous data streams from various sources and parsers are processing data streams to the configurable data formats. The data is pushed (still raw data in view of data warehouse) to the staging area for further processing.

The raw data is available at staging area for further processing. Source data process automation is shown in figure 2.

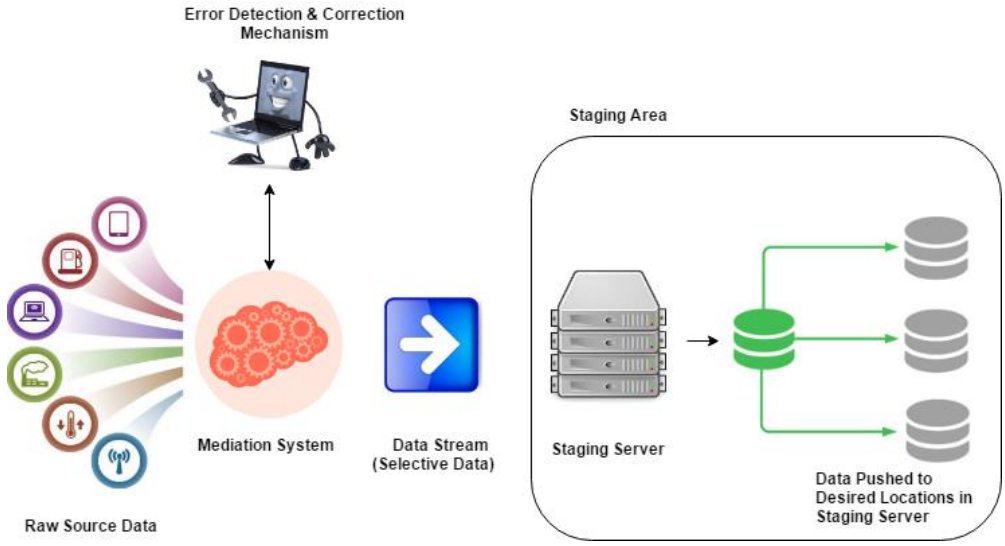

Figure 2. Source Data Process Automation

Data Transformation and Loading Automation:

Raw data is automatically pushed towards staging area by mediation system. The data is available at predefined directory locations of staging area. Our target is to transform raw data to desired format and make it available for loading. Data transformation and loading automation refers to continuous flow of incoming data from different sources so that transformation take place on the fly. Proposed approach is to get data streams from different data sources at staging server in the form of physical files using OS storage. At the same time, the data available at Operating System (OS) storage is viewed in the form of table by making use of external table concept. Loading engine fetches transformed data from staging area and load it into the fact area. The automation is achieved in such a way that it supports large volume of data arriving at high velocity. Data transformation logic is written using external table and using view functionality for transformation so that as and when data arrives in the files at OS level, it would be auto transformed on the basis of logic defined in the view on external table. By doing so, data is automatically available in the form of tables as and when data arrived at staging area in the form of flat files.

Data Retrieval Automation:

In data retrieval automation, various functions have been proposed including:

Function parallel_thread which sets the maximum threshold for reporting to manage load on data warehouse database.

Function select_list checks for the report requirement conditions and accordingly logic is built to get desired result from the data warehouse.

Function request_processing define report location at data warehouse server and execute optimize SQL generated on the basis of logic in previous function. Required report results are streamed to the specified location.

Function system_clear checks for the storage capacity and delete obsolete data based on pre-specified conditions.

Function report_submission updates the incoming request status and check for maximum threads to execute or put in queue depends upon current running reporting threads. Required data location is identified precisely and query aimed at the only partition where data is expected. Structured query language is generated and finally, highly tuned SQL executed for report data.

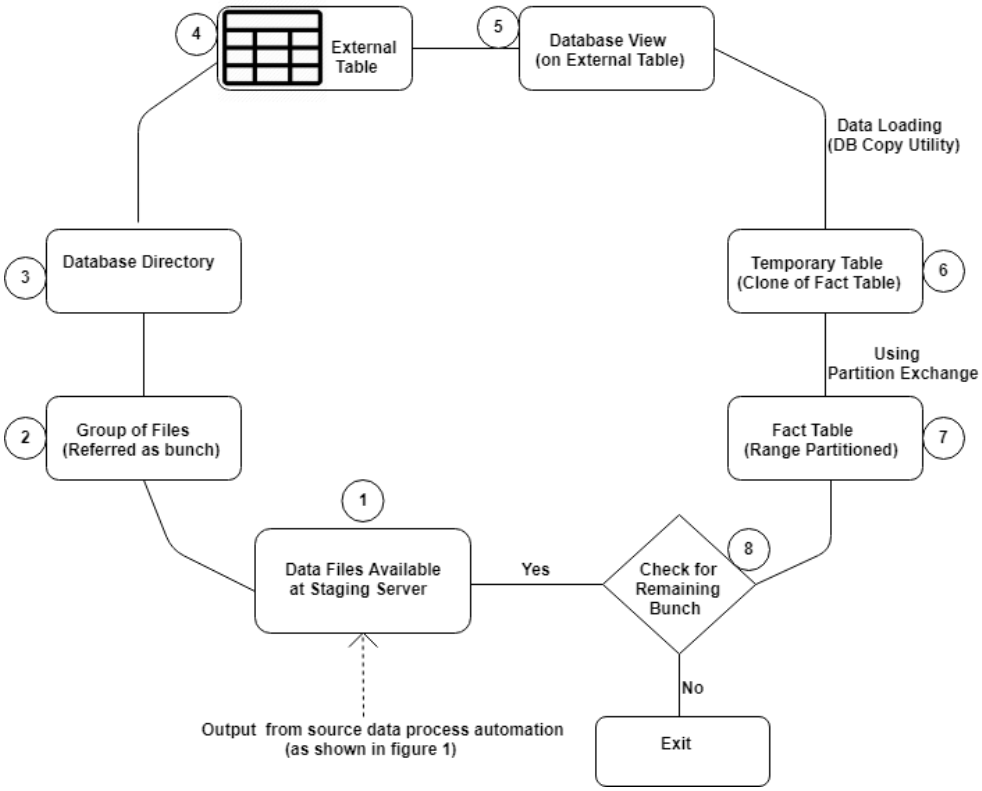

Figure 3. Data Transformation and Loading Automation

The ultimate output of experiment is improved data process automation framework that facilitates highly tuned data processing for large data volumes. Data Transformation and Loading automation framework is shown in figure 3. As detailed in figure 3, data files are available at staging server as shown in step 1. In step 2 of figure 3, group of files are identified which is referred as bunch. In step 3 of figure 3, there are many bunches of files placed at predefined database directory which points to a physical directory at operating system. In step 4 of figure 3, data available in form of flat files at database directory is reflected as tabular. The tables are called as external table. In step 5 of figure 3, data available in external tables is transformed using database view functionality. Transformed data from database view is copied to temporary table which is exact clone of final (fact) table as shown in step 6 of figure 3. In step 7 of figure 3, temporary table is exchanged with final (fact) table using partitioned exchange technique. The process is performed for remaining bunches. Automation process viewed as unified tool is shown in figure 4.

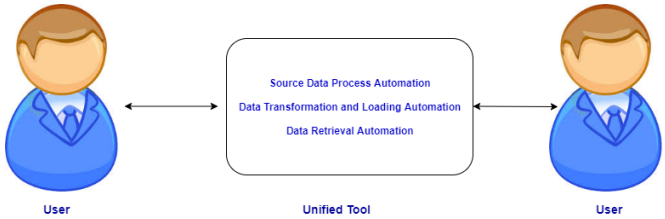

Figure 4. Automation as Unified Tool

Major verticals for data warehouse system of an organizations is ETL and we have added additional dimension to it which is "retrieval". A data warehouse system without retrieval techniques is of no use and retrieval should be planned and taken as the integrated component for building data warehouse system. The processes have been automated and form an enhanced data warehouse system in tandem with other processes. Post individual process automation, the processes are tailored in a highly cohesive environment through programming technology platform such that every data unit moves from one process to other in highly efficient manner to produce the desired output. The integrated process automation framework resulted in an improved data warehouse system automation model which takes cares of the reporting paradigm also.

During model development, it has been observed that when flow of data is high, in the form of data stream/files, data extraction has becomes a vital task to handle and its automation handling involved many complex sub tasks. Automation should take care of the error handling and self-reviving (system recovery) techniques and optimization and logging so that system overhead for next sub process can be minimized. The automation framework without use of commercial tools needs careful attention for transformation, loading and retrieval too.

Advantage of using this framework:

- Improvise end to end automation framework
- Retrieval is considered as core component of a data warehouse system.
- All subsystems of DWH including retrieval are integrated and automated to view as black box.
- Low cost solution without use of commercial tools

- Highly customizable and improved performance.

It has been presented as a complete automated framework which binds sub processes into cohesive way.

V. Performance and Results

The results of the experiment are automation of sub frameworks that majorly contributes to build a data warehouse system. In addition to the existing sub-processes, retrieval has been identified as an integral part for a successful data warehouse system and retrieval sub process has been improvised for automation. The results are as discussed below:

- Source data process automation: The automated resultant framework for source data process automation has mediation system in place which has inbuilt utility for error detection and correction mechanism. The data streams of raw data coming from various sources have been collected at the receiving end of the mediation system. Only selective data after filtering the raw data and its refinement has been sent further to the output data streams at different locations of the staging servers. Its duty of the mediation system to control the flow of incoming data and to ensure that the number of data streams received should be sent to the desired locations at staging server. In case of any issues like network crunch or transfer fail, mediation system ensures to resend the correct data. The result of the sub process automation is smooth flow of data from source system, its filtering and refinement, buffer control and send to the desired staging locations for further processing.

- Data Transformation and Loading Automation: The optimized automation framework has used number of technologies in auto-switch mode to process the data towards desired tables for reporting. Raw data has been configured in the form of external tables based on the logic. The transformation logic has been written logically on external tables and switching of the raw files and further loading to the tables in iterative way such that data flow from raw files to the fact tables has been performed in a fine way. The process is iterative in nature and keeps on loading data and file switching has been performed in highly efficient way so that no manual intervention is required.

- Data Retrieval Automation: The retrieval process automation is to automate the most crucial part i.e. reporting part based on data warehouse. The framework resulted in the data request raised by any interface to be handled in a way that framework has described. The data request in desired format has been updated in a meta-data table and the request parameters have been stored. The parameters may consist of filter conditions and other indicative constraints for a report. The parameters are stored in the variables and number of requests have been checked whether they are permissible or not with number of parallel threads. The optimized logic has been identified for the report and tuning parameters have been incorporated. The request has been submitted in queued way and the request is executed and report is stored at desired location.

A. Observations about the results (compare / contrast between experiments)

Metrics are identified to evaluate the proposal. The research is aimed at reducing cost of overall data warehouse system process automation. **Therefore,** cost effectiveness is taken as one metrics to evaluate the result. The automated framework should ensure efficient data processing without compromising performance. **Therefore,** time taken for data processing is considered as a metrics for result evaluation. During automation, system should avoid manual intervention in most of the cases. **Therefore,** self-healing property of system is considered for evaluation. Reporting is considered as integral part during data warehouse system automation. Retrieval/reporting time metrics is considered for result evaluation.

Therefore, results are evaluated on the following parameters:

- Cost Effectiveness
- Time taken for data processing
- Self-healing property of system.
- Retrieval time metrics.

The resulted output has shown an obvious movement of the data streams that not only facilitates automated flow but use of commercial ETL tools has been eliminated. Error detection and correction mechanism have been incorporated at every sub-processes of the framework to ensure self-healing property of the model. The results have been evaluated to automate the entire process and keeping the above mentioned parameters in consideration to compare the results for optimization/improvement. Observations about the results on parameters are discussed in below sections.

1) Cost Effectiveness

The resultant framework has not used any commercial tools and is based on customized blend of technology mix. The sub processes of the model can be implemented for data warehouse system automation by making use of programming skills of the developer with database conceptual knowledge. The model is highly cost effective and can be easily integrate(d) with DBMS software (open source as well) that gives facility of external table and partitioning concept. It's been found that at any of the sub processes, no commercial tool has been used and framework supports open source that can be integrated with industry of any scale without have to worry about cost. Cost is calculated as design cost, development cost and maintenance cost. Development cost

includes ETL cost and reporting cost. Further, the cost components are divided into activity list which includes design cost, extraction automation cost, transformation cost, loading cost, reporting cost and maintenance cost.

Cost Comparison –

- Design Cost
- Development Cost (ETL + Reporting License)
- Maintenance Cost

Activity Cost Estimates – Activity List for Cost Estimation:

- Design Cost
- Extraction Automation Cost
- Transformation Cost
- Loading Cost (All Above 3 are ETL Cost)
- Reporting Cost
- Maintenance Cost

To compare the cost, parametric estimation technique is used. **Program evaluation and review technique (PERT)** is used for comparison of cost with commercial tools versus proposed framework. Using PERT technique there are three scenarios for cost calculation – most likely, pessimistic and optimistic. Final cost is calculated using PERT formula which is (O+P+4×M)/6 where O denotes optimistic, P denotes pessimistic and M denotes most likely.

Cost is evaluated for a data warehouse project having one year duration that includes development and operations support. Cost is evaluated for two scenarios a) Data warehouse with commercial ETL Tools and b) Data warehouse system with proposed automated framework (without commercial tools). Parameters **are** taken for cost consideration including number of tool licenses, development time, operations time and report/retrieval users. For PERT evaluation; optimistic cost, pessimistic cost and most likely cost is calculated for DWH **using** commercial tools and DWH **using** proposed framework. Comparison chart is prepared as shown in Table 1 and observations are discussed.

Pert formula = (O+P+4×M)/6

Cost using two frameworks are calculated using PERT formula as given above.

So, by putting values –

Total Cost using commercial tools = (1060240+4261200+4*2118560)/6

Estimated Cost of Project= $2299280

Total Cost using automated framework = (962400+2628000+4*1924800)/6

Estimated Cost of Project= $1881600

Total Cost Benefits using automated framework - $2299280 - $1881600 = $417680

Thus saving cost to build a Data warehouse for an enterprise.

2) Time for Data Processing

The automation framework has been analyzed on the time parameter for data processing. The improved automation module should complete the automation in a highly efficient way and in time bound manner. Data processing mainly taking the area after the files arrived at transformation and loading sub process and is available at the fact tables with indexing on the tables. The resultant time has been analyzed number of files; number of records with three indexes is as shown in figure 5 and Time taken for processing (Using Traditional Approach) is captured and statistics are shown in figure 6. Application of the framework is telecom data warehouse in which enormous number of data files are generated in the form of Call data Records.

Table 2. Cost Estimation Using Pert Technique (Parametric Estimation)

	With Commercial Tools (a)			**With Automated Framework (b)**		
	Optimistic	**Pessimistic**	**Most Likely**	**Optimistic**	**Pessimistic**	**Most Likely**
Number of License	1 ETL License	8 ETL License	4 ETL License	NA	NA	NA
Development Time (In Days)	20 Days	60 Days	30 Days	60 Days	120 Days	90 Days
Operations Time (365- Development Time) in Days	345 Days	305 Days	335 Days	305 Days	245 Days	275 Days
Report/Retrieval Users	50 Report User	150 Report Users	100 Report Users	Supports Upto 200 Report Users	Supports Upto 200 Report Users	Supports Upto 200 Report Users
Total Cost	**$1060240**	**$4261200**	**$2118560**	**$962400**	**$2628000**	**$1924800**

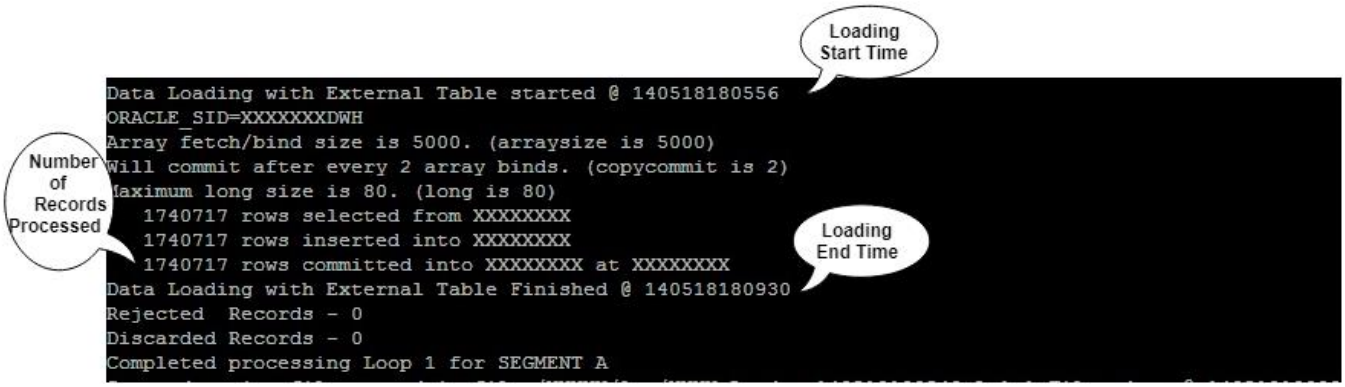

Figure 5. Time Taken for Processing (Using Automated Framework)

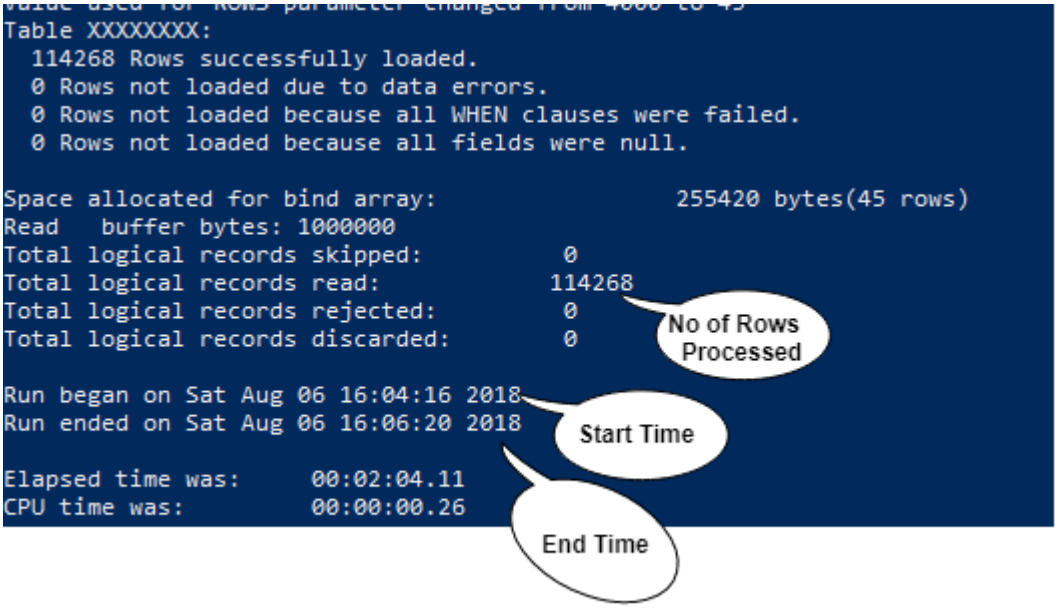

Figure 6. Time taken for processing (Using Traditional Approach)

3) Self-Healing/Revival Property of System

The automation framework has been evaluated on self-healing capabilities of the model. The data stream has been broken while transmitting for result. The auto revival property of system should ensure that the data available at source system should be available at the transformation and loading area for automation. It's been observed that the error detection and correction mechanism is in place at source data process automation. Network communication has been cut for a while to check whether broken data stream has been successfully transmitted to next layer and redundancy has been handled or not. The data stream volume at source system and at transformation and loading area has been checked and verified. In this functional approach, self-healing property of proposed system is presented. In function run_sql_loader, technology mix technique with combination of external table, copy utility with parallel execution on partition exchange tables is first used for data processing. It take group of files presented at a time in an external table **and load(s) it** to the destination database. The groups are referred as bunches and there may be multiple bunches processing at same time by making use of parallel thread technology. In case of issue in a particular bunch, whole bunch processing is suspended and group of files are moved automatically to different location for analysis and processing. For suspended bunch, different type of data processing is used in which row by row loading took placer rather than group of files at one go. Benefit of using this approach is that, only in case of error in bunch processing – this technique is used and we can pinpoint the error records and get it fixed.

Error handling and revival auto revival logic is incorporated.

4) Retrieval Time metrics

Statistics have been checked and retrieval time logs are analyzed. The function report_submission handles retrieval by identifying table partition of required data and locating indexes available on it. SQL query is optimized dynamically by adding partition and indexes as **in-line** hints. Optimized query is passed to **the** database for report generation. The performance statistics are shown in figure 7 and figure 8.

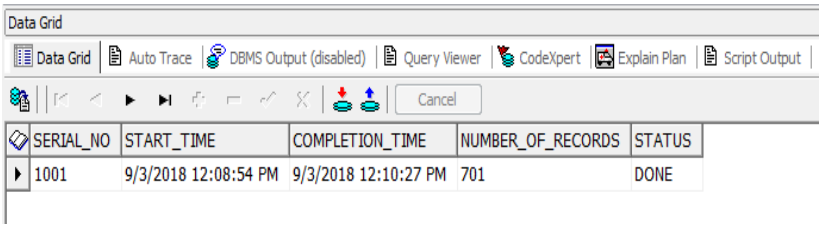

Figure 7. Query Time and Completion Time

```
$ more REPORT_OUTPUT.log
1 row updated.

Commit complete.

Current Threads : 4 MaximumThreads:20
Submitting request for 1001
CIRCLE PARAMETER IS :XX
1 row updated.
Commit complete.
Report Creation Started 030918120854

OUTPUT File for Report Serial No. 1001 is XXXXXXXX.csv
1 row updated.
Commit complete.

Report Creation Finished at 030918121027
```

Figure 8. Query Processing Time Log

The report submission function runs every even minute. This total time taken to generate report 27 seconds from heavy data warehouse table containing billion of records (in Terabytes). The reporting is achieved without using existing reporting tools and saving cost on it.

VI. Conclusion

Automation framework proposed in this work has proved to be an effective solution for data warehouse system automation. Existing studies in the field of data warehouse system focused largely on extraction, transformation and loading while other parameter "reporting" should be taken into account. Data warehouse system for an organization has never been viewed as a single unified tool where reporting was also considered as its integral part. In this paper, data warehouse system process is provided as a complete unified tool covering all intermediate steps from data source to report generation. A minimized maintainability and manual intervention has been reported post implementation of framework. Framework resulted in an improved automation system which includes all steps such as, source data extraction automation, transformation and loading automation and reporting automation. Framework deeply and firmly automates all major ETL processes.

Improved automation framework is evaluated on various parameters and significant improvement in all aspects is observed during result analysis phase. Encouraging results have been recorded on implementation of the automation framework in the context of cost, self-healing property, data processing time, maintainability and report retrieval time.

The automation framework not only automates the entire data warehouse system without using commercial tools, but also gives improved performance characteristics on all sub-systems of a data warehouse system. The framework can be used as a roadmap for any organization which is looking for setup its data warehouse system. Organizations can be benefitted by referring the framework and setting up of data warehouse with complete automation. Every aspect of setting up data warehouse is realized and worked upon to give comprehensive in-depth solution for a data warehouse. This work will be extended for scalability and fitment in the world of big data and analytics. The scope of the framework will be to expand in the area of big data and Hadoop like infrastructure to cop up with emerging data volumes of data driven industry.

VII. References

[1] Yangui R, Nabli A, Gargouri F. Automatic transformation of data warehouse schema to NoSQL data base: comparative study. Procedia Computer Science. 2016 Jan 1;96:255-64.

[2] Bicevska Z, Oditis I. Towards nosql-based data warehouse solutions. Procedia Computer Science. 2017 Jan 1;104:104-11.

[3] Sebaa A, Chick F, Nouicer A, Tari A. Research in big data warehousing using Hadoop. Journal of Information Systems Engineering & Management. 2017;2(2):10.

[4] Bani FC, Girsang AS. Implementation of database massively parallel processing system to build scalability on process data warehouse. Procedia Computer Science. 2018 Jan 1;135:68-79.

[5] Rocha L, Vale F, Cirilo E, Barbosa D, Mourão F. A framework for migrating relational datasets to NoSQL. Procedia Computer Science. 2015 Jan 1;51:2593-602.

[6] Data B. Automating Data Preparation: Can We? Should We? Must We?.

[7] Ali SM. Next-generation ETL Framework to Address the Challenges Posed by Big Data. InDOLAP 2018.

[8] Rezaee Z, Sharbatoghlie A, Elam R, McMickle PL. Continuous Auditing: Building Automated Auditing Capability. InContinuous Auditing: Theory and Application 2018 Mar 8 (pp. 169-190). Emerald Publishing Limited.

[9] M. Maleszka, B. Mianowska and N. T. Nguyen, "A framework for data warehouse federations building," *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Seoul, 2012, pp. 2897-2902.

[10] R. G. Goss and K. Veeramuthu, "Heading towards big data building a better data warehouse for more data, more speed, and more users," *ASMC 2013 SEMI Advanced Semiconductor Manufacturing Conference*, Saratoga Springs, NY, 2013, pp. 220-225.

[11] Wang Wenzhi *et al*., "Research and design data center for tobacco based on data warehouse and data mining," *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, Chengdu, 2016, pp. 166-170.

[12] H. Chai *et al*., "UStore: An optimized storage system for enterprise data warehouses at UnionPay," *2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2016, pp. 1574-1578.

[13] D. R. Harris, D. W. Henderson, R. Kavuluru, A. J. Stromberg and T. R. Johnson, "Using Common Table Expressions to Build a Scalable Boolean Query Generator for Clinical Data Warehouses," in *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1607-1613, Sept. 2014.

[14] Y. Song, F. Pramudianto and E. F. Gehringer, "A markup language for building a data warehouse for educational peer-assessment research," *2016 IEEE Frontiers in Education Conference (FIE)*, Erie, PA, USA, 2016, pp. 1-5.

[15] L. Yessad and A. Labiod, "Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault," *2016 International Conference on System Reliability and Science (ICSRS)*, Paris, 2016, pp. 95-99.

[16] F. Serra and A. Marotta, "Data quality in data warehouse systems: A context-based approach," *2016 XLII Latin American Computing Conference (CLEI)*, Valparaiso, 2016, pp. 1-12.

[17] Y. Sharma, R. Nasri and K. Askand, "Building a data warehousing infrastructure based on service oriented architecture," *2012 International Conference on Cloud Computing Technologies, Applications and Management (ICCCTAM)*, Dubai, 2012, pp. 82-87.

[18] H. L. H. S. Warnars and R. Randriatoamanana, "Datawarehouser: A data warehouse artist who have ability to understand data warehouse schema pictures," *2016 IEEE Region 10 Conference (TENCON)*, Singapore, 2016, pp. 2205-2208.

[19] N. A. H. M. Rodzi, M. S. Othman and L. M. Yusuf, "Significance of data integration and ETL in business intelligence framework for higher education," *2015 International Conference on Science in Information Technology (ICSITech)*, Yogyakarta, 2015, pp. 181-186.

[20] T. A. Abdulaziz, I. F. Moawad and W. M. Abu-Alam, "Building data warehouse system for the tourism sector," *2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, Cairo, 2015, pp. 410-417.

[21] S. Abdellaoui, L. Bellatreche and F. Nader, "A Quality-Driven Approach for Building Heterogeneous Distributed Databases: The Case of Data Warehouses," *2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, Cartagena, 2016, pp. 631-638.

[22] Grosan C, Abraham A, Chis M. Swarm intelligence in data mining. InSwarm Intelligence in Data Mining 2006 (pp. 1-20). Springer, Berlin, Heidelberg.

[23] Onwubolu GC, Buryan P, Garimella S, Ramachandran V, Buadromo V, Abraham A. Self-organizing data mining for weather forecasting. InIADIS European Conf. Data Mining 2007 Jul (pp. 81-88).

Author Biographies

Sachin Sharma He is a Ph.D Scholar born at Shahzadpur District Ambala Haryana India on April 22, 1987. He did Bachelor of Technology in Computer Engineering from Kurukshetra University, Kurukshetra,Post Graduate Diploma in IT from SCDL,Pune and M.Tech in Computer Science and Technology from MMDU, Mullana, Ambala Haryana. His area of research is Data Analytics, Data Warehouse, Big Data and other fields of data science.

Dr. Kamal Kumar He is working as Assistant Professor in Dept. of Computer Science & Engineering, National Institute of Technology, Uttarakhand-246174. He did Ph.D from Thapar University Patiala, M.Tech and B.Tech from Kurukshetra University Kurukhetra. His research area of interest are WSN, Security, Cloud Computing, Deep Learning and Artificial Intelligence.

Dr. Sandip Kumar Goyal He is professor and head of computer science engineering department at Maharishi Markandeshwar Deemed to be University, Mullana Ambala Haryana. He did Ph.D from MMDU Mullana, M.Tech and B.Tech from Kurukshetra University Kurukshetra. His area of interest are load balancing and cloud computing.