ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

Data Warehouse as a Generic Approach a Review

Shahid Bashir Dar[1], Ashish Sharma[2]

M. Tech Student, Dept. of CSE, BIMT, Mehli, Shimla, H.P, India[1]

Assistant Professor, Dept. of CSE, BIMT, Mehli, Shimla, H.P, India[2]

ABSTRACT: Digitization of data resulted in the generation of massive volumes of data in less time. The heterogeneous and disperse data sources makes the scene more complicated to handle. With the advent of 21st century enterprises realized the importance of data spread across disparate sources. Large efforts were made to integrate this data at one place for carrying out long term managerial decisions out of it. These efforts resulted in the development of data warehousing as a solution for data integration and data analytics. A number of warehousing solutions have been proposed in the last few years to analyze business data, meteorological data, clinical data, and so on. But least research has been done to develop a generic tool that can create a warehouse irrespective of enterprise and data.

KEYWORDS: data warehousing, data integration, data marts.

I. INTRODUCTION

A data warehouse is a repository of subjectively selected data from heterogeneous systems with the intent to provide strategic business information. Data warehouses are designed to facilitate answers to ad hoc, large and complex, statistical or analytical queries to carry out analysis and reporting. Enterprises need warehouse for effective business intelligence, strategic business formulation, and critical business decisions in order to survive in a globally competitive market where huge volumes of continuously growing heterogeneous data needs to be stored, processed and analysed. Traditional operational systems can't be used for such purpose as they are meant for day-to-day business operations, data from these operational systems flows into the warehouse where it is used for strategic decision making. Traditional operational systems stores current values optimized for transactions having high access frequency and large number of users. Access types are of read, update, and delete operation having low response time in the range of sub-seconds. On contrarily data warehouse stores archived, derived, and summarized data that is optimized for complex queries usually having low access frequency and small number of users. Read operation is the only access type in warehousing taking more time in giving response from seconds to minutes.

A. *Goals:*

A successful data warehouse should provide following features [12] to any organization.

- Fast and easy accessibility of information.
- Should present information consistently.
- Flexible and adaptive to handle day to day changes.
- Better decision making.
- High security.

B. *Data Warehouse versus Data Mart:*

Data from the warehouse flows into various departments for analysing their respective area. These individual departmental components refer to as data marts. A data mart is a logical subset of a warehouse that is targeted towards a single functional area [1][2][3] like sales, finance etc. Data warehouse is the union of all data marts which gathers data from broader subjects unlike data mart whose data comes from only few areas. Since data marts are small in size as compared to the data warehouse, so they are preferred for fast and easy analysis. The structure of a data mart is to suit the departmental view of data while as, a warehouse provides a corporate view of data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

C. *Design Considerations:*

The success of any data warehouse depends on how its design is complete and holistic to exhaustively cover all the possible heterogeneous data sources. It should handle all available data sources so as to meet the user requirements effectively. Inability to do so may lead to a design that is skewed towards a particular functional area [4], missing the other user requirements. Some of the important considerations while designing a warehouse are as:

- Heterogeneity of data sources and data types.
- Data integrity.
- Authenticity and reliability.
- Scalable with data growth.
- Appropriate selection of data.

D. *Data Integration:*

Data integration forms the heart of data warehousing and refers to the collection and transformation of data from disparate sources into a single unified structure. As data passes from operational databases to the data warehouse, various redundancies and inconsistencies arise in data that needs to be resolved to get an integrated and reconciled view of data of an enterprise. There are mainly three ways to the heterogeneous data integration; data warehouse approach, middleware approach and federated approach. Data warehouse method makes use of ETL technology to transform the data from a different to a unique form stored in a single centralized repository on which complex queries are executed. Middleware method makes use of mediator as a middleware between the query interface (client) and the disparate data sources. Mediator divides the query submitted by the client into sub-queries against the specific data sources. It also makes use of wrappers as query and result translators. Federated approach makes one to one connection between all pairs of data sources and is based on a decentralized architecture.

II. LITERATURE SURVEY

Bill Inmon primarily gave the idea of warehousing in 1992, in his book titled "Building the Data Warehouse". However, the roots of warehousing can be traced back to 1960 when G. Mills and D. College developed in their project the concept of dimensions and facts, which are the corner stone's of warehousing even as on date. It was not till 2000 when people realized the essence of warehousing and emergence of its implementation. This is the time when industry realized the importance of data integration and research in the area of warehousing became formalized and aggressive. The social media was germinated and massive data was being generated. This resulted in the rise of warehousing as a tool for effective data integration and data/ business analysis.

The evolution of data warehouse has not been exemplary; more than 50 percent of data warehouse implementations fail in achieving the specified goals [8]. Surajit Chaudhuri, and Umeshwar Dayal [13] in 1997 presented an overview of data warehousing and OLAP technologies and considered them as essential elements of decision support systems. Their paper also discusses back end tools, front end tools, and multidimensional data models required for a data warehouse.
A novel approach to conceptual modelling for source integration is in [14]. This model presented in 1998 suitably models the global concepts of the application, information sources, and their constraints. In 1999, Peter Chamoni and Steffen Stock [15] presented the essence of temporal structures in data warehousing. The temporal structure needs to be modelled in the data model of warehousing for the reasons of consistency. Time stamps can be used for such purposes. Vassiliadis [16] in 2000 debated the gap between researchers and practitioners. The issues in research and practice, and the extent up to which these overlap in the field of data warehousing is discussed. Diego Calvanese et.al [17] in 2001 presented a novel approach to data integration in a data warehouse. Information integration is one of the most important aspects of a data warehouse. When data passes from the sources of the application-oriented operational environment to the Data Warehouse, possible inconsistencies and redundancies should be resolved, so that the warehouse is able to provide an integrated and reconciled view of data of the organization. Their approach is based on a conceptual representation of the Data Warehouse application domain, and follows the so-called local-as-view paradigm. They propose a technique for declaratively specifying suitable reconciliation correspondences to be used in order to solve conflicts among data in different sources.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

Watson [19] in 2002 discusses the recent developments in data warehousing. Data warehouses can be developed in two alternative ways; the data mart and the enterprise wide data warehouse strategies, each having advantages and disadvantages. Depending on the business needs, either a relational or a multidimensional database technology can be used for the data stores. To get a multidimensional view of the data using a relational database, a star schema data model can be used. By the year 2003 researchers were concerned more about a comprehensive methodology that supports the entire process of determining information requirements of data warehouse users, matching information requirements with actual information supply, evaluating and homogenizing resulting information requirements, and establishing priorities for unsatisfied information requirements. Winter, Robert, and Bernhard Strauch [20] discussed such components as well as overall design based partially on literature review, but mainly on findings from a four year collaboration project with several large companies, mostly from the service sector.

Brandt et.al [21] in 2004 developed a patent for telecommunications integrated with web and providing a GUI for requesting, customizing, scheduling and viewing of various types of priced call detail data reports. Such an infrastructure performs an extraction process to obtain only those billing detail records of specific customers, and a harvesting process for transforming the billing records into a star schema format for storage in one or more operational data storage devices.

March et.al [22] in 2007 presented that supporting managerial decision making is critically dependent upon the availability of integrated, high quality information organized and presented in a timely and easily understood manner. Thomsen et.al [23] in 2008 presented with an idea of RiTe (Right-Time ETL), a middleware system that makes inserted data quickly available and providing bulk load insert speeds. A data producer (ETL) can insert data that becomes available to consumers on demand. RiTE includes an innovative main memory based catalyst that provides fast storage and offers concurrency control.

Santos et.al [24] in 2009, demonstrated that data warehouses must be able to enable continuous data integration, in order to deal with the most recent business data. Traditional data warehouses are not able to support any dynamics in structure and content while they are available for OLAP. Their data is periodically updated because they are unprepared for continuous data integration. They used table structure replication with minimum content and query predicate restrictions for selecting data, to enable loading data in the data warehouse continuously, with minimum impact in query execution time.

By the year 2010 social networking site had dominated WWW (internet), Ashish Thusoo et.al [25] discuss data warehouse and the underlying infrastructure at Facebook. They discuss how various open source technologies like Hadoop, Scribe, and Hive are used for data integration and data analytics. Kamil Bajda-Pawlikowski et.al [27] in 2011 discussed in detail the performance oriented query execution strategies for data warehouse queries in split execution environments, with particular focus on join and aggregation operations. By 2012 it was established that data warehousing (DW) requires huge investments, the data warehouse market is experiencing incredible growth. However, a large number of data warehouse initiatives end up as failures. The maturity of a data warehousing process could significantly lessen such large scale failures and ensure the delivery of consistent, high quality, single version of truth" data in a timely manner. However, unlike software development, the assessment of DWP maturity has not yet been tackled in a systematic way [26]. Cuzzocrea et.al [28] in 2013 explores the convergence of Data Warehousing, OLAP and data-intensive Cloud Infrastructures in the context of analytics over Big Data.

Majid and Muheet [5] in their research starting from 2006 to 2014 concluded that there should be a generic warehouse tool to create warehouse of any enterprise making use of fact and dimension tables. Data warehouse can be used to boost the educational sector by having insights into the academic data with the aim of predicting and forecasting new patterns and trends in the educational institutes. Any kind of a situation or a problem can be resolved efficiently. The critical need for the same and a customized model of a warehouse for educational institutions has been proposed in [7]. To resolve the problems of business processes like high costs in relying ad hoc solutions, dealing with frequent changes in business processes, [6] proposes a generic process warehousing solution to improve the business processes. Data warehousing an integral part of decision support systems are increasingly becoming more critical to the large enterprises.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

III. PROPOSED WORK

This research seeks integration of multiple data sources into single Data Warehouse irrespective of number and types of sources, this is achieved by designing, implementing and testing generic Data Warehouse tool. The design is given below.

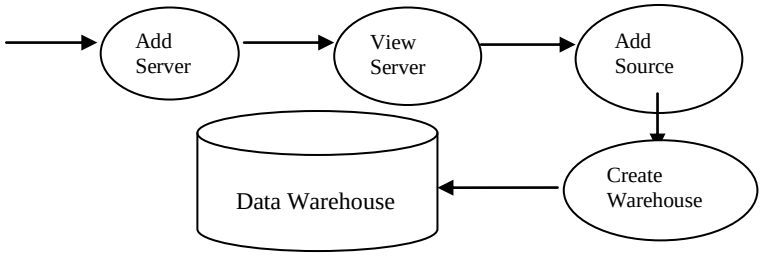

Figure 1. Data warehouse creation steps

The modules listed above are required for enlisting of various data sources and subsequently retrieving data from the said sources. This data is then transformed into the warehouse data using 'create warehouse' module.

A. *Objectives:*

i. Design and develop generic Data Warehouse approach for multiple data sources, the proposed solution will traverse data sources and create warehouse on runtime based on structure of existing source and design optimal structure of data warehouse.
ii. In the proposed work ETL would be completely automatic with minimal human intervention.
iii. The proposed solution will be completely generic- irrespective of type/and number of data source.
iv. Test proposed approach on multiple academic data sources, besides will determine effectiveness of this approach for efficient knowledge discovery.

IV. CONCLUSION AND FUTURE WORK

Enterprises across the globe are still struggling with data integration and warehouse design mainly because of heterogeneous data sources and over dependence on (lack) technical people in the organization. A number of warehousing solutions have been proposed in the last few years to analyse business data, meteorological data, clinical data, and so on. But least research has been done to develop a generic solution towards warehouse creation. We will implement the proposed work using java language.

REFERENCES

1. PaulrajPonniah, "Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals", John Wiley & Sons 2001, ISBN:0-471-41254-6.
2. Er. Majid zaman, Dr. S.M.K. Quadri, Er. Muheet Ahmed Butt, "Integrating Information from Heterogeneous Data Sources: Universityof Kashmir case study", Journal of Global Research in Computer Science, Volume 3, No. 5, May 2012, ISSN: 2229-371X.
3. Er. Majid zaman, Er. Muheet Ahmed Butt,"Information Integration: An Enterprise Solution", International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 1, January 2013, ISSN: 2319-5967.
4. C.S.R. Prabhu, "Data Warehousing: Concepts, Techniques, Products, and Applications", East Economy Edition, Second Edition, Prentice-Hall of India, 2006, ISBN: 81-203-2068-9.
5. Majid Zaman, Muheet Ahmed Butt,"Warehouse Creator: A Generic Enterprise Solution",International Journal Of Engineering And Science, Vol.2, Issue 12 (May 2013), Pp 65-68.
6. Fabio Casati, Malu Castellanos, UmeshwarDayal, Norman Salazar1, "A Generic solution for Warehousing Business Process Data", Copyright 2007 VLDB Endowment, ACM 978-1-59593-649-3/07/09.
7. Shaweta," Critical Need of the Data Warehouse for an Educational Institution and Its Challenges", International Journal of Computer Science and Information Technologies,Vol. 5 (3) , 2014, 4556-4559.
8. Debasish Mukherjeea& Derrick D'Souzab,"Think Phased Implementation for Successful Data Warehousing", Information Systems Management, Volume 20, Issue 2, 2003.
9. Marcus Costa Sampaio André Gomes de Sousa Cláudio de Souza Baptista, "Towards a Logical Multidimensional Model for Spatial Data Warehousing and OLAP", , November 10, 2006, Arlington, Virginia, USA.Copyright 2006 ACM 1-59593-530-4/06/0011.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

10. Zaman, Majid. "Health Information Systems Integration Plan: Databases Perspective." Journal of Global Research in Computer Science 4.4 (2013): 103-107.
11. Zaman, Majid, and Muheet Ahmed Butt. "Enterprise Management Information System: Design & Architecture." International Journal of Computational Engineering Research (IJCER), ISSN 2250 (2013): 3005.
12. Ralph Kimball, Margy Ross," The Data Warehouse Toolkit: the complete guide to dimensional modeling", second edition, Wiley Computer Publishing.
13. Surajit Chaudhuri &Umeshwar Dayal, "An overview of data warehousing and OLAP technology", ACM SIGMOD, Volume 26 Issue 1, March 1997, Pages 65-74.
14. Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Daniele Nardi, Riccardo Rosati , "Source integration in data warehousing", Database and Expert Systems Applications, 1998. Proceedings. Ninth International Workshop on, 25-28 Aug 1998, 192 – 197
15. Peter Chamoni, Steffen Stock, "Temporal Structures in Data Warehousing", DataWarehousing and Knowledge DiscoveryLecture Notes in Computer Science Volume 1676, 1999, pp 353-358
16. Panos Vassiliadis, "Gulliver in the land of data warehousing: practical experiences and observations of a researcher" DMDW. 2000.
17. Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Daniele Nardi, and Riccardo Rosati, "Data integration in data warehousing." International Journal of Cooperative Information Systems 10.03 (2001): 237-271.
18. Zaman, Er Majid, and ErMuheet Ahmed Butt. "Information Integration: An Enterprise Solution." International Journal of Engineering Science and Innovative Technology (IJESIT) 2.1 (2013): 315-317.
19. Hugh J. Watson, "Recent developments in data warehousing." Communications of the Association for Information Systems 8.1 (2002): 1.
20. Winter, Robert, and Bernhard Strauch. "A method for demand-driven information requirements analysis in data warehousing projects." System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on. IEEE, 2003.
21. Andre R Brandt, Barbara Frueh, Sajan J Pillai, Karl Rehder, Donald J Shearer, "Data warehousing infrastructure for web based reporting tool." U.S. Patent No. 6,714,979. 30 Mar. 2004.
22. Salvatore T. March, and Alan R. Hevner. "Integrated decision support systems: A data warehousing perspective." Decision Support Systems 43.3 (2007): 1031-1043.
23. Thomsen, Christian, Torben Bach Pedersen, and Wolfgang Lehner. "RiTE: Providing on-demand data for right-time data warehousing." Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on. IEEE, 2008.
24. Ricardo Santos Jorge, and Jorge Bernardino. "Optimizing data warehouse loading procedures for enabling useful-time data warehousing." Proceedings of the 2009 International Database Engineering & Applications Symposium. ACM, 2009.
25. Ashish Thusoo, Borthakur, Raghotham Murthy, Zheng Shao, Namit Jain, Hao Liu, Suresh Anthony, Joydeep Sen Sarma , "Data warehousing and analytics infrastructure at facebook." Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, 2010.
26. Arun Sen, K. Ramamurthy, and Atish P. Sinha. "A model of data warehousing process maturity." Software Engineering, IEEE Transactions on 38.2 (2012): 336-353.
27. Kamil Bajda-Pawlikowski, Daniel J. Abadi, Avi Silberschatz, Erik Paulson "Efficient processing of data warehousing queries in a split execution environment." Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, 2011.
28. Cuzzocrea, Alfredo. "Analytics over big data: Exploring the convergence of datawarehousing, olap and data-intensive cloud infrastructures." 2013 IEEE 37th Annual Computer Software and Applications Conference. IEEE, 2013.
29. Butt, Muheet Ahmed. "Integrating Information from Heterogeneous Data Sources." Journal of Global Research in Computer Science 3.5 (2012): 71-73.