International Journal of u- and e- Service, Science and Technology
Vol.8, No.4 (2015), pp.9-16
http://dx.doi.org/10.14257/ijunesst.2015.8.4.02

A Survey on Data Warehouse Constructions, Processes and Architectures

[1,2]Muhammad Arif
[1]*Faculty of Computer Science and Information Technology, University of Malaya 50603 Kuala Lumpur, Malaysia*
[2]*Computer Science Department, Comsats Institute of Information and Technology Islamabad Pakistan*

Abstract

Data warehouse is the repository of historical data. Real time data ware housing has been achieved by different tools and techniques. This paper is based on survey on different dimensions of real time data ware housing in different manner. The survey is based on real time data ware house loading methodology in which data is load in real time and its update, different aspects of peer to peer network for cashing of OLAP results, self-maintenance view of different multiple views in data ware house scenarios, for checking the research either it is dead or alive for modeling and design of data ware house, view maintenance in ware housing and also focus the study on the temporal requirements of real time DWH. And also provides the comparison between these specific studies by giving the tables.

Keywords:

1. Introduction

Data warehouse the warehouse where all the data has been integrated and stored along with the historical data. There are few available tools to provide the real time data warehousing.

By supporting the easy and available query at any time we can use real time data ware housing and it empowers the end nodes by giving the most up to date information. By using the data marts and ETL services from the source system helps the users to give the query results on run time. Real time data warehouse can be applied in crime mining [14], telecommunication, universities and in multinationals compnies.

Real time data warehousing mainly consist of three tiers that are integrated each other [1]

i) Presentation tier provides the interface between the user and data ware house.
ii) Architecture layer gives the structure of data ware house in a flexible manner.
iii) Middleware tier holds the data ware house as a glut.

In real Time Data ware House Architecture, Operational data in which data is gathered from different resources and then send it for the change and transformation of data. And then move it to the data ware house.

ISSN: 2005-4246 IJUNESST
Copyright ⓒ 2015 SERSC

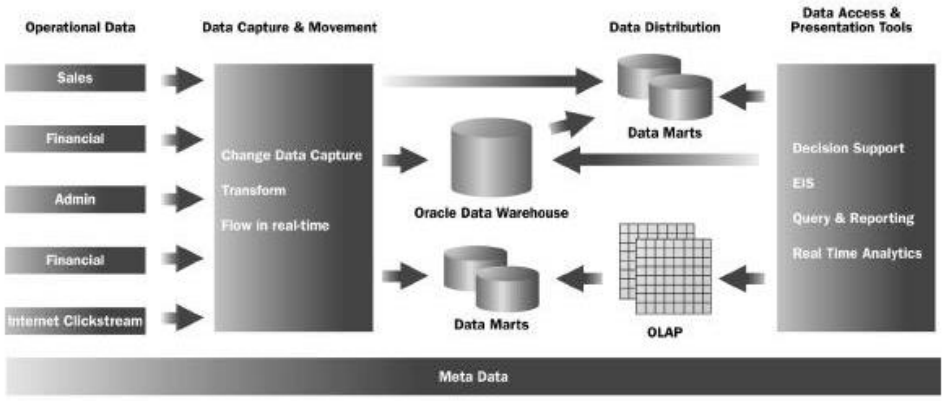

Figure 1. Real Time Data Ware House Architecture [1]

Real time data ware housing playing vital role in field of databases. Typically real time data ware housing provides reflection of business in real time and it defines the system as the refection of business in real time and also we can define real times data ware housing as it is the process of delivering information regarding different operations in business. In this research survey different real time methodology are discussed such as loading methodology used in real time data ware housing, different temporal requirement, maintenance of data and its multiple views, real time data warehousing model and designing, and also discusses the adoptive measure of peer to peer network in real time.

2. Literature Review

A lot of work has been done on the dataware house architecture and gives different techniques for improving and better performance in architecture.

The research found that real time data ware housing provides the synchronization between warehouse and transactional data and redefine the static implementation. Traditional data ware house has static structure in schema and are not able to have dynamic structure. Due to periodic up gradation nature it cannot allow for integration of data. So to overcome the real time data from promising, a methodology is proposed for how to adopt OLAP queries and schemas for integration of real time data efficiently.

To resolve this problem of integration the researcher provides the solution in integration of data ware house during execution has been enabled with no effect on performance and minimize the delay on decision support database between the transaction of information and data updating. The researcher focus on the loading processes and usage of data area for efficient integration. to get new real time schema of data ware house by doing replication of data structure and getting instructions of query by using the simulation of TPC-/h bench mark by performing the simultaneously data integration at various synchronization level that increase the time of query during execution [2].

Researcher found that there are less work has been done by employing system of database on peer to peer networks. Peer OLAP architecture proposed for the queries of OLAP a cache is used to connect the peer 2 peer network for the answering of queries and this cache acts on the peer OLAP and have the advantage of client side cache. By decreasing the cost of query to make the system distributed and reconfigurable. In this paper [3] the research gives the idea of core components of peer OLAP and gives the result by doing simulation and installation of prototypes.

For effective decision making global environment plays an essential role in the data ware housing. Previous research in this area discussed value of multidimensional space in the model, different Lattice graph with nodes, using aggregations as materialized views both in static and dynamic approach, cache manager called as watchman, Dynamat (OLAP cache manager), cache results are in the form of chunks, OLAP works in proxy

 Copyright ⓒ 2015 SERSC

server by extracting them in OLAP cache servers and peer OLAP network. Peer OLAP networks in posing queries in set of peer by accessing data ware housing.

Peer OLAP networks processing OLAP queries in the set of peer networks by accessing data ware houses and is the combination of data ware house with seven peers network.

Peer architecture has the different work nodes connected with each other and also with data ware house takes the OLAP queries and gives desired results and also the cost model is present in its architecture.

Query processing is either by two different ways:

a) Processing of Eager Query (EQP)
b) Processing of Lazy Query (LQP)

In this scenario it has a defined cache policy that is used for execution of queries as well as minimizing the cost of query such as:

i) Admission & algorithm of replacement also called Least Benefit First (LBF) which focus the objects and its benefits
ii) Isolated Caching Policy (ICP) gives advantage to peers.
iii) Caching policy of hits aware (HACP) gives hits to other peers by decreasing the cost of the query and last is
iv) Voluntary caching for avoiding any wastage of results.

By creating the neighbors virtually the network has reorganize the peers and caching the results of query. Experiment evaluation and execution is done by having client side cache architecture, optimization of queries strategies, caching policies evaluation and execution and reorganization of networks and its effects [3].

Data ware house consists of different vies by attempting different queries for getting the desired results and have different relation between the views. User queries are executed by using these views. On these views we can generate different queries and updates the relations by suing external resources so it become expansion to reduce its cost by maintaining the views and updating the records. May be it will not maintain its view into some specific situation and its effect of update on its state. So a question arises here that whether and how the view will be maintained.

For this problem the researcher provides the solution proposed an algorithm that resolves the problem of maintenance. The algorithm generates the queries of SQL and its answer is based on the maintenance of the view in the multiple view presence involving base data and gives the idea of self-maintenance by decreasing it in the query containment.

In most ware housing system the view are generated from that initial stage to resolve the maintenance problem and gives the maintenance solution by the limited use of external base data. Defining self-maintenance problem in the multiple view have different dimensions such as strict self-maintenance, generalized self-maintenance, notation and query containment.

Different queries are generated if the view is known and is in self-maintenance to become update by giving the examples and there proves. After that it decides the self-maintenance and then queries are generated to examine the self-maintainability view by using algorithm.

So the algorithm that is proposed in this paper [4] gives the maintainability of views on the base access by generating the tests and SQL queries expressions. It can help the efficiency of ware house [4].

There are a lot of techniques and modeling involves in the design of data ware house but have fewer consensuses on the method of design. In this paper [5] issues of design and modeling of data ware house has been discussed such as issues in design methods,

Copyright ⓒ 2015 SERSC

conceptual and logical model issues and some issues that are carried out by doing the new architecture and models. Some architecture has defined multidimensional and non OLAP modeling techniques.

This paper [5] arises a wide discussion on the current trends ware housing in the workshop " DATA ware housing at the cross roads".

Conceptual Models gives the obstruction in architecture and work s back bone in developing a data base according to the user requirements. Multidimensional & ETL modeling is mainly two view points for conceptual modeling. In this modeling the issues are still there such as lack of standards, security of modeling and mining-aware design. On the basis of conceptual modeling, logical modeling has transformed for a specific targeted system in main issues that arises in logical modeling are semantic gap and ETC modeling.

Different methods and techniques are used to design the conceptual and logical models and several methods are for the automating a database such as requirement analysis, schema evaluation and quality metrics.

By discussing the impact or architecture and new applications design and modeling mostly on spatial data ware housing, web ware housing, real time data ware housing and BPM also in distributed data ware housing and suggesting the ad-hoc techniques to resolve the issues in designing and modeling. It is concluded that the research of modeling and design of data ware housing are not dead. Some techniques are used to solve specific problem [5].

Data ware house implements views that are maintained and updated are the materialized views. These views are different from traditional views and decoupled the base data. It results in anomalies in the traditional implementation of algorithm. A new algorithm called ECA (Eager Comparison Algorithm) is proposed to expel anomalies and is comprises of incremental algorithm of maintenance view with more other queries. By giving the two versions and extensions of ECA for specific views and updates along with initial study of performance that gives the comparison in order to get transition of messages, transfer of data and cost of I/O.

For the correct updating to the views from the source in materialized views and its assurance is the main problem to illustrate this by using different examples and discussions such as correct view maintenance, anomaly of view maintenance and anomaly deletion.

A lot of mechanisms are used to avoid the anomalies but the research is focus only on scenarios where source, which may unsophisticated system with no performance of any view management. Only responsible for the updates and resulting the queries by using and recomputed the view, store at the ware house relations in views and its copies and also eager computing algorithm. To improve these we can use two improvements in basic ECA to ECA- Key algorithm and ECA local Algorithm.

By checking correctness of environment where the source activity is decouple in ware house by the events both on source and warehouse end. By using the algorithm which have the properties of convergence, weak consistency, strong consistency and completeness.

To present the algorithm we must have to know about the queries and views by handling two types of update in which deletion and insertion are involve. By using sign on tuples, these tuple signs are used in relational operations and the other is query expression that is used to present the algorithm.

ECA algorithms of maintenance have two extensions of ECA key and ECA local having the properties such as

i) No need to start from scratch gives the updates from the source.

ii) No additional burden on source.

iii) When frequency of update is low the query comes back before next updating.

 Copyright ⓒ 2015 SERSC

The performance evaluation criteria define the performance on the basis of number of messages to the source that consist of queries and performance is also based on data transferred and on input/output [6].

Temporal requirements are the important part in real time data ware housing. Distribution of electronics documents are widely performing and publishing over the World Wide Web. In this paper [7] the discussion is focus on the feeding of data ware house in real time. By decreasing the delay time of web pages and it's changing on internet and system detects or notify the this change during the updating process by maintaining the temporal consistency of data that are extracted from the source information and also gives the monitoring of the system by delivering the requirements of users.

All data such as manual help, forms, profiles, new policies, of different government agencies are presents on their websites same as with the stand alone organizations. So the daily updating of records on the web pages are required autonomously. But for the user queries there in not any availability of standard interface, not any structure of result, and embedded domain knowledge.

In this paper [7] data is capturing from the web and then feeding it in real time data ware housing for further operations and maintenance by using the Data Extraction with Temporal Constraints.

In real time data ware house requirements there are three characteristics [8]:

i) Integration of data from operational resources.
ii) Engines of active decision for recommendations and decisions.
iii) Analytical environment for accessing and loading.

The components presents in the real time data ware housing involves capturing real time environment, delivery of real time, transformation engine, data ware house, aggregator of increment and preparation engine [9].

Wrapper [10] is the module which is use to translate the information from the source by using the W4F as a tool kit which generates fast wrappers for web. These wrappers mostly follow the POST and GET methods of HTML in the timely schedule manner by using algorithms of coherency between cached data and source data [11,12].

These wrapper output is used for many purposes such as real time user and client delivery in real time, storing the web page for future use and its conversion to XML is possible, integration with the information, data mining and analysis [13] of real time data ware house and also inform users for the specific value of threshold. This technique is used to maintain the temporal requirements by the integration and retrieval of information.

Table 1. Explains the Type, Tools and Challenges Handled by Data Warehouse

Ref#	Author name	Published date	conference / journal paper	DWH Type	Tool name	Challenge Handled
[2]	Ricardo Jorge Santos, Jorge Bernardino	September 2008	Conference	Real Time DWH	Technique is used for table structure replication and query predicate restrictions for selecting data	Yes
[3]	Panos Kalnis, Wee Siong Ng, Beng Chin Ooi, Dimitris Papadias, Kian-Lee Tan	June 2002	Journal	Distributed DWH	Peer OLAP Architecture	yes
[4]	Nam Huyn	April 1995	Conference	Real Time DWH	Algorithms for general class of view	Yes

Copyright ⓒ 2015 SERSC

[5]	Stefano Rizzi, Alberto Abello	10 Nov 2006	Journal	Modeling of DWH	Survey	Yes
[6]	Yue Zhuge, Hector Garcia-Molina, Joachim Hammaer, Jennifer Widom	1995	Journal	Real Time DWH	Algorithm ECA for view maintenance	Yes
[7]	Francisco Araque	-	Conference	Real Time DWH	Technique to minimize the delay between time a web page changes and change is detected	Yes

Table 2. Explains the Approach, Type, Analysis, Tools of Data Ware House

Ref#	Gui Schema Yes/no	Data warehouse approach	Data ware house type	Application type	Analysis type	Layered approach	Tool	ETL Yes/no
[2]	No	Methodology to adopt DWH Schema	Real Time	Methodology	Demonstrate by query performance using bench mark TPC-H	No	Technique	Yes
[3]	Yes	Peer OLAP	Real Time	Simulation	Prototype installation running on geographically remote peers	Yes	Simulator and prototype on peer networks	Yes
[4]	No	Multiple View self-maintenance	Traditional DWH	Algorithm	generates SQL updates for maintenance of views	No	Algorithm	Yes
[5]	No	Modeling and Design	Both	Follows Discussions	Focus on conceptual model, logical model	No	No Tool	Yes
[6]	No	View Maintenance in DWH	Traditional DWH	Algorithm	Versions of ECA Algorithm	No	Algorothm	Yes
[7]	No	Real time data warehousing with temporal requirements	Real Time DWH	Wrapper Technique	System that minimize delay between time a web page changes on internet and the time this change is detected	No	No Tool	Yes

Table 3. Explains the Experimental Environment, Related Architectures, and Future Work

Ref#	Experimental environment	Related architecture / work	Future idea
[2]	Efficiency of the method by analyzing its impact in query performance using benchmark TPC-H executing query work-loads for data integration at various insertion time rates	Tools and algorithm to populate DWH in offline fashion. On the fly computation of queries. Conceptual ETL Modeling ETL Algorithms along with their consequent OLTP and OLAP performance issues. Efficient Techniques for maintaining DWH without disturbing source operations Temporal aggression operations	Develop ETL tool which will integrate this methodology with extraction and transformation routines for the OLTP systems. There is also room for query instructions used for proposed methods.

 Copyright ⓒ 2015 SERSC

		Data currency quality factors Transactions in temporal active databases ARKTOS ETL Tool for modeling and executing practical ETL scenarios Zero Delay DW	
[3]	Peer network environment for supporting OLTP queries, peer OLAP acts as large distributed cache, which implies the benefits of traditional client side caching.	Technique to accelerate OLAP is to pre calculate aggression and store them as materialized views Special case of OLAP semantic cache manager was developed names as Watchman. Dynamat another cache manager for OLAP. Infrastructure for OLAP Cache Servers Piazza is the system to deal issues of database management in P2P systems	Possible extension is to develop algorithm for reconfiguration of network by identifying the neighbor peers and clustering problems
[4]	Multiple view of self-maintenance in data ware housing and run time approach to self-maintenance	Autonomously Computable Updates have the self-maintainability conditions. Single- view strict self-maintenance problem. Compile time self-maintenance	This technique allows generating tests and maintenance expressions in SQL queries.to improve efficiency of DWH to deal with multi-set semantics.
[5]	Survey on the discussion that took place in Dagstuhl in the workshop on “Data ware house at the Crossroads ” by discussing the important issues on modeling and design of data ware housing.	Distinguishing the logical design nad conceptual design. Physical design methods CASE solutions by different vendors of DWH.	Fragmentations of DWH where new data marts are often added
[6]	View maintenance in warehousing and Data ware housing update processing in a single source model.	Incremental view maintenance algorithm. Algorithm to handle views defined by Datalog expression and SQL. Materialized views in distributed systems. Algorithm named snapshot for time stamping. Approaches for the timing of view maintenance.	ECA and its extensions are used for views over multiple sources, handle the set of updates at once in spite of one update at a time, by modifying this algorithm to handle views of more complex relational algebra expression and also how this algorithm is useful for other data models.
[7]	Real time data ware housing with temporal requirements and minimizing the delay between time a web page changes on internet and this change in time has been detected by this system that is proposed in this paper	Write Wrappers and encapsulate it with the access of sources. No related work with temporal consistency.. Data ware house refreshment. ETL tool Arktos for modeling and executing scenarios of ETL. Xyleme a warehouse for monitoring XML data for supporting queries. Tracking static and dynamic monitoring by WebCQ.	Exposed this research to for information retrieval and integration of temporal requirements, and working on incorporate valid time and transaction time.

3. Conclusion

This survey complies the different scenarios of real time data ware housing. Real time methodology for supporting the RTWD for data integration and execution of query on the user end. This methodology simulation is based on TCP benchmarks for data integration on various rates of time. It gives the real time DW performance in maximum time of execution of query that is also effect on cost.

Survey finds the peer caching system for giving the OLAP results in client server environment by maintaining the previous results with new one. And also find out the algorithm to test the views of updates along with the previous stored views and handled the base updates by maintain the updates and also allow the user to test and maintained the SQL expressions. Research finds the algorithm along with its two extensions that are used to maintaining the materialized views in real time DWH environment.

Real time data ware housing achieving the temporal requirements by doing the wrapper technique and get the results such as user and client delivery in real time, saving database

Copyright ⓒ 2015 SERSC

for future use (XML conversion) and its integration and data extraction and finds the design and modeling issues in DWH.

References

[1] J. Vandermay, "Considerations for building a real-time data warehouse", White Pape DataMirror Corporation, (**2001**) November.
[2] R. J. Santos and J. Bernardino, "Real-time data warehouse loading methodology", In Proceedings of the 2008 international symposium on Database engineering & applications, ACM, (**2008**) September, pp. 49-58.
[3] P. Kalnis, W. S. Ng, B. C. Ooi, D. Papadias and K. L. Tan, "An adaptive peer-to-peer network for distributed caching of olap results", InProceedings of the 2002 ACM SIGMOD international conference on Management of data, ACM , (**2002**) June, pp. 25-36.
[4] N. Huyn, "Multiple-view self-maintenance in data warehousing environment", (**1997**).
[5] S. Rizzi, A. Abelló, J. Lechtenbörger and J. Trujillo, "Research in data warehouse modeling and design: dead or alive?", InProceedings of the 9th ACM international workshop on Data warehousing and OLAP, ACM, (**2006**) November, pp. 3-10.
[6] Y. Zhuge, H. Garcia-Molina, J. Hammer and J. Widom, "View maintenance in a warehousing environment", ACM SIGMOD Record, vol. 24, no. 2, (**1995**), pp. 316-327.
[7] F. Araque, "Real-time Data Warehousing with Temporal Requirements", In CAiSE Workshops, (**2003**), June.
[8] R. M. Bruckner, B. List and J. Schiefer, "Striving towards near real-time data integration for data warehouses", Springer Berlin Heidelberg, (**2002**), pp. 317-326.
[9] H. Watson, T. Ariyachandra and R. J. Matyska, "Data warehousing stages of growth", Information Systems Management, vol. 18, no. 3, (**200**1), pp. 42-50.
[10] D. Bhandari, "Extraction Of Web Information Using W4F Wrapper Factory and XML-QL Query Language", (**1999**).
[11] A. Sahuguet and F. Azavant, "Web Ecology: Recycling HTML pages as XML documents using W4F", Database Research Group (CIS), (**1999**), p. 24.
[12] A. Sahuguet and F. Azavant, "Wysiwyg web wrapper factory (w4f)", (**1999**).
[13] M. Arif, K. Amjad Alam, and M. Hussain, "Application of Data Mining Using Artificial Neural Network: Survey", International Journal of Dataabase Theory and Application (IJDTA), vol. 8, (**2015**), February.
[14] M. Arif, K. Amjad Alam and M. Hussain, "Crime Mining: A Comprehensive Survey", International Journal of u- and e- Service, Science and Technology (IJUNESST), vol. 8, (**2015**) February.

Author

Muhammad Arif, he is a PhD student at Faculty of CS and IT, University of Malaya. Currently he is working on Medical image Processing. His research interests include image processing, E learning, Artificial intelligence and data mining. He joined UM as a Bright Spark Scholar in September 2013 for the period of 3 years. Before this he completed masters and bachelor degrees in Pakistan. He received his BS degree in Computer Science from University of Sargodha, Pakistan in 2011. He obtained his MS degree in Computer Science from COMSATS Islamabad 2013 Pakistan.

 Copyright ⓒ 2015 SERSC