# *Data Warehousing: A Journey from Traditional to Active/Real time Data Warehousing*

*Ali Raza*
Department of Computer Science
University of Management and Technology
Lahore, Pakistan
aliirazii@gmail.com

*Dr.Sajid Mahmood*
Department of Computer Science
University of Management and Technology
Lahore, Pakistan
Sajid.mahmood@umt.edu.pk

***Abstract*—*Data warehouses are an important component of decision support systems because they provide adhoc access to the data of interest for analysis and decision support. Data warehouse are used where data grows to be very large overtime and handling such huge amounts of data needs special handling different from traditional management of data, which does not provide support for adhoc access analytics and decision making process. Traditional data warehouses have fixed schemas and structures that are not able to adapt the changing overtime where as active data warehouse is more adaptive to change and it update data frequently. In this paper, chronic shifts in the data warehousing techniques are discussed from traditional to active/real time data warehousing. Different frameworks of traditional and active data warehouse are reviewed with highlighting the aim of each framework, its usefulness and likely drawbacks***

***Keywords—Traditional, active, data warehouse, integrated data, schemas, star schema, snowflake schemas***

## I. INTRODUCTION

Data warehouse is a large collection of data integrated form different homogenous/heterogeneous sources to help knowledge workers in analysing the data by providing adhoc access from business point of view. Data warehouse design and implementation has undergone changes over time making it more advanced and which helps managerial decision making [1].Traditional data warehouses consist of relational database, which do not provide full access to ad hoc queries [35]. Organizational needs for information continue to evolve, that triggered the era of active data warehouse for decision making process. Enterprise application integration becomes an important part of business intelligence framework [2]. This paper, presents a journey of traditional data warehouse that was designed and implemented in the past and how it transferred changed to active data warehouse. While during the implementation and improvement of data warehouse phase, support is required from the end users through the information centers which increase satisfaction level of end users when dealing with data warehouse environment [3].

For each of the techniques discussed in this paper, there is an introduction to the significance of that research contribution towards traditional or active data warehouse, then generally it is explained what was the utilization or drawback of this strategy if there is any.

The rest of the paper is organized as follows. Section 2 present the traditional data warehouse frameworks, Section 3 present the active data warehouse frameworks and Section 4 consist of the conclusion and future work.

## II. TRADITIONAL DATA WAREHOUSING TECHNIQUES

Tryfona et al. [7] introduced starER model for the conceptual modelling in the data warehouse, which is a combination of ER model and data warehouse star structure data. Further this model is been evaluated with other conceptual models and provide efficient results for complex information and semantically starER model is more substantial than the dimensional fact schema. Hammer et al. [4]explained how data warehouse can help in collecting scientific data, maintains history of business, the architecture of project is in figure 1. The drawbacks consist of inconsistencies in data and data being physically copied from original source.

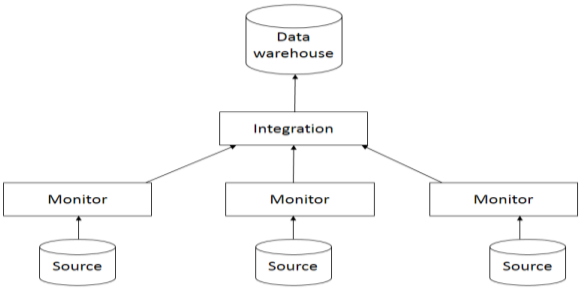


**Figure 1: Data warehouse Architecture WHIPS Project**

Golfarelli & Rizzi [5] explained a framework for the data warehouse on the principles of dimensional Fact modelling.

This methodology of data warehouse design consist of six steps in which include analysis and requirement specification of information system, conceptual design , refinement and finally logical design is converted into physical design. Franconi & Sattler [6]describe that database schemata using the object oriented and semantic data model. Conceptual model is discussed using the aggregated entities and the relevant entities making aggregated entities. It also explains how data warehouse quality architecture matches with the conceptual data model but this framework does not support temporal and spatial dimensions.

Generalized multi-dimensional normal form can give conceptual warehouse schema and how operational database schema can be converted into the conceptual schema of warehouse using the framework designed which follows three steps, recognizing the attribute property, graphical way of conceptual design and last obtaining generalized multi-dimensional normal form (GMNF) [8]. Phipps & Davis [9]created algorithms, first for converting the OLTP schema into the data warehouse conceptual schemas, this algorithm was designed for the data model partitioned textual, numeric and date/time. Second algorithm consists of the evaluation of the conceptual schemas using the user queries. TPC-H benchmark is used to illustrate the algorithms. Lechtenbörger & Vossen [10]multi-dimensional normal forms are defined which tells about the quality of conceptual data warehouse schema, three normal forms have been defined from generalized multi-dimensional normal form. Rao et al. [11]introduced to the process that consist of three steps, use spatial indexed tree on the spatial dimensions, algorithm to calculate the pre aggregation and finally heuristic query method that can improve OLAP query performance. Results showed efficient results and prototype system based on this method showed feasible solutions.

## III. Active Data warehouse

Thalhammer et al. [12] presented the active data warehouse architecture, following a close loopsystem between Data warehouse and OLTP system as in Figure 2 below

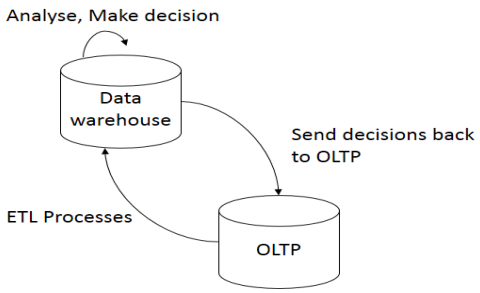


**Figure 2: Active Data warehouse Cycle**

Bruckner [13] introduces data model concerning time delays which include, hyper cube multi-dimensional model and star schema based but it does not provide solution for the history data. Data warehouses have built in time consistency support which provide better view of history of the organization data and this methodology provides time consistent view for analysis. Bebel et al.[14]introduced multi version data warehouse that is able to handle the changes in the schemas which help in dealing with different business scenario. The prototype of this system was made in Visual C++ and data, whereas oracle was used for storing Meta data. Time integrity constraint were used while defining types of versions needed and another feature is defining alternative versions of data warehouse in multi version data warehouse. Karakasidis [15] explained a framework consisting in which ETL activities are implemented on the network queue and prediction of the performance of the system is evaluated using the queue theory. Minimal source and software overhead and smooth running of the system are the aim of this framework and considering data freshness the results were produced satisfactorily. Santos [16] proposed a methodology in which techniques such as table structure replication and query predicate restrictions used for real time data integration and loading without effecting the query execution time and the effectiveness of this method is evaluated using the TPC-H query workload. The methodology for loading data is given in Figure 3:

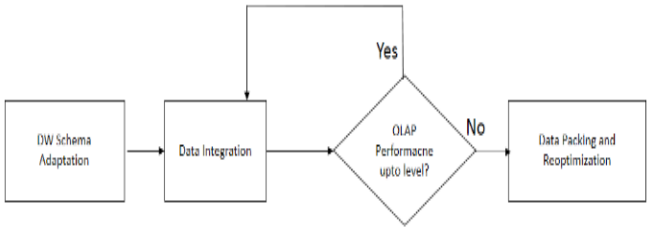


**Figure 3: Continuous Data Warehousing Loading Method**

Thusoo et al. [17] introduce Hive, an open source data warehouse built on the Hadoop platform which is a map reduce implementation. Query language in Hive allows users to use map reduce scripts into queries. It also contains schemas and statistics used for query optimization and query compilation in Metastore (System Catalogue). Leonardi et al. [18] presented the framework for storing the aggregate information on trajectories of changing objects and it also provide Visual analysis for the OLAP operations. This framework deals with the free/constraint objects, and works for both temporal and spatial dimensions. The trajectory data warehouse and Visual notation provides one to view the movement of the objects and analyse them. Salem [19] introduced the active XML (AXML) framework for the integration of data from different heterogeneous system. Active rules are also used to activate integration services and finally Frequency XML based tree is represented for the mining association rules. Finally a prototype of this application is made based on this framework. Boopathy

[20]proposed solution to scheduling of data streams in the data warehouse using the Particle Swarm Algorithm(PSO) and this framework give separate time for the short jobs and using PSO for handling data streams showed improved results in the form of reduced staleness. Sumbaly et al. [21] explains how Hadoop based stack in LinkedIn is used to extract data and know about the insights of the large scale data. It also provides a detail view of how data flows from offline system to online system. Gupta et al. [22]elaborate about the MESA (cloud based data warehouse), which is a geo replicated real time data warehouse, it handles the advertisement information of the Google and it deals with trillions of rows per day, executing operations robustly and providing accurately and scalable data. Finally it defines the architecture and report on the performance of MESA. Mane [23]explains that near real time data warehouses shortens the refreshment time taken by the data and Extract-transform-Load (ETL) systems are used to help these processes. It also highlights the important data that should be refreshed in order to help analyse the activities. Chan [24]explains architecture of real time customer relationship management system (CRM) and it was made using the e-business, Knowledge base system and virtual and real time analytics. It also defines the significance of CRM in the business analytics. Nath [25]introduced programmable semantic ETL framework, to facilitate programmers and then evaluation of SETL is carried out by comparing it with the traditional tools. Results have shown that SETL showed better quality results. Saeed et al. [26] explored the key characteristics of different queries with the comparison of different accelerator hardware for running on the different data warehouses. The findings suggested that multiple integrated cheap GPUs perform better than the expensive GPUs for the complex queries in data warehousing applications if transfer of data is not optimized between devices.

Cao et al. [27] proposed a framework for the data cube model used for the support of systematic spatiotemporal analysis of location based social media data. With the help of particular data cube, one can summarize, query the spatiotemporal distribution, it can also handle location based social media dynamics. Johnson [28] developed data stream warehousing system named Tidalrace system aimed at network monitoring and maintenance application and some of the features like support for temporal consistency, partition re-organization. Real time loading, deep analytics and long histories are provided by the Data stream warehousing. Song [29]indicated the three types of the de-duplication in real time data warehousing which are de-duplication prior scheduling strategy (DPS), ETL Prior scheduling strategy (EPS) Real time scheduling strategy (RS) and then a new scheduling strategy is derived from ETL scheduling strategy (EPS) named Time Triggered scheduling strategy (TTS) and results of the proposed strategy proved that it can be efficiently used in real time data warehouses. Athanassoulis et al. [30]experimented that use of SSDs to cache incoming updates increase the support for the updates in the data warehouses. MaSM algorithm is used to implement this process and proofs regarding the properties of MaSM algorithms are also provided which help in providing correct ACID support. Whereas results showed that MaSM algorithm query response time is same even if the updates are running at the same time. Qu et al. [31] proposed method for the maintenance of real time snapshot in MVCC data warehouses. The consistency models defined proved to provide fair scheduling of the OLAP queries and concurrent maintenance flow. The proposed model was used on incremental ETL pipeline which enabled high query throughput achievement. Average results were reached with this approach when compared to traditional near real time ETL maintaining the query consistency. Baranowski [32] explains how scale out databases are means of solution for the large data warehouses. Hadoop and Cloudera Impala engine architectures are discussed and tested in this paper. The results showed that Hadoop systems showed high throughput for processing and storing large sums of information at CERN. HDFS beneath the Impala engine has shown better results of the tests and by default this approach is fit for many data warehouses as it allow to query data through SQL interface. Hogan & Jovanovic [33] explained the results of offloading the inactive data from Data warehouse relational database management system (RDBMS) to Hadoop distributed file system (HDFS). Extract transform and load workflows are achieved from the tools for the data offloading solution. The results of the experimentation suggested that offloading data to Hadoop using the ETL workflows can be achieved. Yi et al. [34]proposed private data warehouse queries phenomena, which allows client to retrieve cell from the data warehouse without notifying it to the data warehouse operator so that is operator cannot follow the client queries and interest. This is done with the Boneh-Goh-Nissim cryptosystem and the results showed that this solution provide security to both client and server.

## IV. Conclusion

The design and framework of data warehouse evolved with the time due to business needs and analytics required. The change from traditional data warehouse that consist of static requirements to active data warehouse which is dynamic to change happened gradually due to industry demand from business analytics and other fields. In future, an evaluation framework for measuring efficiency of different data warehouses should be emphasized.


## Acknowledgment

I am very thankful to the department of computer science at University of Management and Technology for providing me peaceful and ambient environment. I am also very thankful to all the reviewers who took time out of their busy schedule for reviewing this article.


## References


[1] Sen, A., & Sinha, A. P. (2005). A Comparison of Data Warehousing Methodologies. Communications of The ACM, 48(3), 79-84.

[2] Brobst, S. A. (2002). Enterprise Application Integration and Active Data Warehousing. In Vom Data Warehouse Zum Corporate Knowledge Center (Pp. 15-22). Physica-Verlag Hd.

[3] Soliman, K. S., Mao, E., & Frolick, M. N. (2000). Measuring User Satisfaction with Data Warehouses: An Exploratory Study. Information & Management, 37(3), 103-110.

[4] Hammer, J., Garcia-Molina, H., Widom, J., Labio, W., & Zhuge, Y. (1995). The Stanford Data Warehousing Project.

[5] Golfarelli, M., &Rizzi, S. (1998, November). A Methodological Framework for Data Warehouse Design. In Proceedings Of The 1st Acm International Workshop On Data Warehousing And OLAP (Pp. 3-9). Acm.

[6] Franconi, E., & Sattler, U. (1999). A Data Warehouse Conceptual Data Model for Multidimensional Aggregation: A Preliminary Report. Italian Association for Artificial Intelligence Ia Notizie, 1, 9-21.

[7] Tryfona, N., Busborg, F., &Borch Christiansen, J. G. (1999, November). Starer: A Conceptual Model for Data Warehouse Design. In Proceedings Of The 2nd ACM International Workshop On Data Warehousing And Olap (Pp. 3-8) ACM.

[8] Hüsemann, B., Lechtenbörger, J., &Vossen, G. (2000). Conceptual Data Warehouse Design (Pp. 6-1). Universität münster. Angewandtemathematik Und Informatik.

[9] Phipps, C., & Davis, K. C. (2002, May). Automating Data Warehouse Conceptual Schema Design and Evaluation. In Dmdw (Vol. 2, Pp. 2-2).

[10] Lechtenbörger, J., &Vossen, G. (2003). Multidimensional Normal Forms for Data Warehouse Design. Information Systems, 28(5), 415-434.

[11] Rao, F., Zhang, L., Yu, X. L., Li, Y., & Chen, Y. (2003, November). Spatial Hierarchy and Olap-Favored Search In Spatial Data Warehouse. Inproceedings of The 6th Acm International Workshop On Data Warehousing And OLAP (Pp. 48-55) ACM.

[12] Thalhammer, T., Schrefl, M., & Mohania, M. (2001). Active Data Warehouses: Complementing OLAP with Analysis Rules. Data & Knowledge Engineering, 39(3), 241-269.

[13] Bruckner, R. M., & Tjoa, A. M. (2001). Managing Time Consistency for Active Data Warehouse Environments. In Data Warehousing and Knowledge Discovery (Pp. 254-263). Springer Berlin Heidelberg.

[14] Bębel, B., Eder, J., Koncilia, C., Morzy, T., &Wrembel, R. (2004, March). Creation and Management of Versions In Multiversion Data Warehouse. In proceedings of the 2004 ACM Symposium on Applied Computing (Pp. 717-723) ACM.

[15] Karakasidis, A., Vassiliadis, P., &Pitoura, E. (2005, June). ETL Queues for Active Data Warehousing. In Proceedings of the 2nd International Workshop on Information Quality in Information Systems (Pp. 28-39). ACM.

[16] Santos, R. J., & Bernardino, J. (2008, September). Real-Time Data Warehouse Loading Methodology. In Proceedings of the 2008 International Symposium on Database Engineering & Applications (Pp. 49-58) ACM.

[17] Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Zhang, N., & Murthy, R. (2010, March). Hive-A Petabyte Scale Data Warehouse Using Hadoop. In Data Engineering (Icde), 2010 Ieee 26th International Conference On (Pp. 996-1005). Ieee.

[18] Leonardi, L., Orlando, S., Raffaetà, A., Roncato, A., Silvestri, C., Andrienko, G., &Andrienko, N. (2014). A General Framework For Trajectory Data Warehousing And Visual Olap. Geoinformatica, 18(2), 273-312.

[19] Salem, R., Boussaïd, O., &Darmont, J. (2013). Active Xml-Based Web Data Integration. Information Systems Frontiers, 15(3), 371-398.

[20] Boopathy, M. S., & Subramanian, M. K. Improved Scheduling And Minimized Updates In Data Warehouses.

[21] Sumbaly, R., Kreps, J., & Shah, S. (2013, June). The Big Data Ecosystem At Linkedin. In Proceedings Of The 2013 Acm Sigmod International Conference On Management Of Data (Pp. 1125-1134). Acm.

[22] Gupta, A., Yang, F., Govig, J., Kirsch, A., Chan, K., Lai, K., &Bhansali, S. (2014). Mesa: Geo-Replicated, Near Real-Time, Scalable Data Warehousing.Proceedings Of The Vldb Endowment, 7(12), 1259-1270.

[23] Mane, M. N. G. Near Real-Time Data Warehousing Using Etl (Extract, Transform And Load) Tools.

[24] Chan, J. O. (2015). The Anatomy Of Real-Time Crm. Communications Of The Iima, 6(1), 11.

[25] Deb Nath, R. P., Hose, K., & Pedersen, T. B. (2015, October). Towards A Programmable Semantic Extract-Transform-Load Framework For Semantic Data Warehouses. In Proceedings Of The Acm Eighteenth International Workshop On Data Warehousing And Olap (Pp. 15-24). Acm.

[26] Saeed, I., Young, J., &Yalamanchili, S. (2015, February). A Portable Benchmark Suite For Highly Parallel Data Intensive Query Processing. In Proceedings Of The 2nd Workshop On Parallel Programming For Analytics Applications (Pp. 31-38). Acm.

[27] Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., &Soltani, K. (2015). A Scalable Framework For Spatiotemporal Analysis Of Location-Based Social Media Data. Computers, Environment And Urban Systems, 51, 70-82.

[28] Johnson, T., Shkapenyuk, V., &Hadjieleftheriou, M. (2015). Data Stream Warehousing In Tidalrace. In Cidr.

[29] Song, J., Liu, H., Wu, J., &Bao, Y. B. (2015). De-Duplication Scheduling Strategy in Real-Time Data Warehouse. The Open Cybernetics &Systemics Journal, 9(1).

[30] Athanassoulis, M., Chen, S., Ailamaki, A., Gibbons, P. B., &Stoica, R. (2015). Online Updates On Data Warehouses Via Judicious Use Of Solid-State Storage. Acm Transactions on Database Systems (Tods), 40(1), 6.

[31] Qu, W., Basavaraj, V., Shankar, S., &Dessloch, S. (2015). Real-Time Snapshot Maintenance With Incremental Etl Pipelines In Data Warehouses. In Big Data Analytics And Knowledge Discovery (Pp. 217-228). Springer International Publishing.

[32] Baranowski, Z., Grzybek, M., Canali, L., Garcia, D. L., &Surdy, K. (2015). Scale Out Databases For Cern Use Cases. In Journal Of Physics: Conference Series (Vol. 664, No. 4, P. 042002). Iop Publishing.

[33] Hogan, M. T., & Jovanovic, V. (2015). Etl Workflow Generation For Offloading Dormant Data From The Data Warehouse To Hadoop. Issues In Information Systems, 16(1).

[34] Yi, X., Paulet, R., Bertino, E., &Xu, G. (2013, June). Private Data Warehouse Queries. In Proceedings Of The 18th Acm Symposium On Access Control Models And Technologies (Pp. 25-36). Acm.

[35] Https://Docs.Oracle.Com, Accessed: 7/03/2016