Survey on Data Warehouse from Traditional to Realtime and Society Impact of Real Time Data

Farhad Alam
Analyst, eBay (Shanghai), China

Neel Kamal
Research Scholar, Himalayan University
Arunachal Pradesh, India

ABSTRACT
The traditional data-warehouse don't have real-data or near to real-data or today-data. Generally, the data loading into traditional-data warehouse from operation sources whether its single or multiple and its scheduled on weekly bases or nightly bases. And these kinds of data is hard to think to make some decision and to do some prediction and treat them. As todays to make some conclusions in the corporate world become more and more real-time or near to real-time systems for end-users. It is only natural that data-warehouse business intelligence bi decision support and olap systems rapidly start to incorporate real-time data. in this paper we are interested in giving a survey on data-warehousing starting from a traditional data-warehouse to a real-time data-warehouse and what is the society impact of real-time data. this survey is also focus on data-warehouse architecture. It details the changes in the extract-transform-load process to deal with real time data-warehousing. it sketches the integration data in the real time data warehouse. Finally, a comparative study concerning the real data warehouse approaching is also presented in this paper.

Keywords
Traditional Data Warehouse (DWH) OR Enterprise data-warehouse (EDW), Real Time Data Warehous（RTDW,Near to Real Time Extract-Transform-Load, Real Time Data Society Impact.

1. INTRODUCTION
A DWH or EDW which the old integrated data from different sources. And its central repositories for combined data from different sources and providing the information to end-user to generate there. report for business organizations to make some decision to do some prediction to enhance the business to make more values in market. And these data in DWH are turned into structured forms that can simply be handled by decision-making procedures used to make strategic decisions by means of online systematic processing OLAP techniques.Because of the new requirements of business forever up-to-second updated data and since timely data ensure better-informed decisions real-time DWH is one of the savior trends that make available an entree to accurate integrated consolidated view of the information of organizations in real-time.

A real-time DWH have no official definition currently exists in the literature. A RTDW can be de need as a system that signifies the characteristics and the actual situation of the organization. for instance, if we have a request to investigate a actual fact of the organization embarked on an RTDW the answer will be represented in the real state of the organization at the time of the request sending. Unlike most traditional data-warehouses an RTDW is seen to contain recent data real-time of the organization. thus, the refreshing frequency plays a predominant role in RTDW. generally based on the literature a RTDW undergoes refreshments several times a day which enables the decision makers to have access to the current data of the organization.

To make available data freshness performance accessibility prediction capabilities and to offer an integrated information repository to drive and tactical decisions many tactics are proposed to treat the architectures RTDW and the data integration in the RTDW.

2. NEED FOR THE REAL TIME DATA WAREHOUSE
The DWH can be updated in two ways traditionally and also in real-time but traditional method drawback is contents of data is not updated and then hard to make a good decisions and end-users is not able to generate the report having fresh data. so, company need a near to real-time DWH where some information are updated in real-time and the remaining are traditionally refreshed. thus, the source system will be over load because only critical data will be frequently extracted from the source or strategic decisions are made using old data. storing data in near-real time will reduce the latency between business transactions to operational sources and their appearance in the DWH. the facilitates of the investigation of more recent data and makes the decision faster.

However, DWH and business-intelligence real-time applications have been proposed to answer accurately the types of questions that end-users would like to ask for real-time data. a real-time DWH enables data to store the data at a time when they are produced and immediately captured cleaned and stored within the structure of the DWH. then traditional refreshment cycles are no longer valid. the DWH must be able to read the same data that move around the operating systems at the same time of its generation.

The purpose of RTDW is to enable enterprises to rapidly access information and notify the user or decision-making system to react almost immediately to the information. There are two arguments that justify the use of real-time data-warehouse. (1) The traditional data-warehouse which provides the state of the company at a specific time in the past (every day, every week or every month). However, some applications require a smaller temporal resolution (every hour). The real-time warehouse enables the organization to find variants of its state even within a day. (2) The second argument focuses on the fact that when a large volume of data is entered into the transaction systems in a single day (financial sector); processing ETL sometimes causes problems [1]. The operation of the real-time warehouse suggests rather frequent and regular additions in a single day. Table1 presents a detailed comparison between the traditional data-warehouse environment and the real time data-warehouse environment.

In the following sections, we will present some approaches that can move closer

Table 1. Major Differences Between A Traditional Dw And An Rtdw

Traditional Data Warehousing	Real Time Data Warehousing
Updating data Historical data periodically	Updating Real-time data
For strategic decisions only	For strategic and tactical decisions
Results hard to measure	Results measured with operations
Highly restrictive reporting used to check the pattern to make some prediction	Flexible ad hoc reporting and machine
Moderate user currency	Results measured with operations
Daily, weekly, monthly data concurrency is acceptable	Only data available within minutes is acceptable

3. DISCUSSED RTDW AND OTHER APPROACHES WHICH REPRESENT AN RTDW ARCHITECTURE AND INTEGRATION OF REAL-TIME DATA. CLOSER TO A REAL TIME DATA WAREHOUSE

In most cases, updating a DW also means putting it out of use (shutdown) during the updating or at least this one makes its use much more di cult and with poor performances. Using a warehouse for an update may also cause some inconsistencies in the results which are returned by the query execution. They will be rarely logical and correct if interviewed data are updated at the same time. The criteria required for continuous updates without involving a shutdown are generally unpredictable with traditional ETL tools. To address this problem, new solutions specialize in ETL real-time and data loading. There are also so-lutions modifying conventional ETL systems in order to load a warehouse on a frequency approximating the real-time.

Existing ETL system can be modified to perform real-time or near real-time data warehouse loading. Some of these solutions are described in [3]:

Near Real-Time ETL: This is first solution consists in eliminating the real-time reality of the choice if the need is absent. We could simply increase the frequency of loading new data into the warehouse. For example, a load that is usually done on a weekly basis could be executed once a day. This approach may, however, involve shutdown problems. This change would enable the users of the DW to have access to more recent data without having to make major modifications to the loading process or data model. Not being the real-time reality, the near-real time can be a good first low-cost solution.

Direct Trickle: If an application requires real-time, the simplest approach would be to continuously load the DW with the new data. In this way, we would eliminate any intermediate storage step. However, this solution can cause a loss of performance when the exploitation of the DW by one or more user(s) since the update of a warehouse itself can require a lot of work from the machine that hosts it.

Trickle and Flip: This approach involves making the second partition of the fact table of the data warehouse on which we make the load. On a periodic basis, we replace the fact table of the warehouse with the second partition, which is responsible for new data. In this way, the searchable fact table is updated at a low frequency to limit

the performance loss during its operation. It also contains all the new data since the last update. This solution can be used according to a cycle, which can vary from hours to minutes.

External Real-Time Data Cache This tactic consists in Loading the data in a real-time memory (Real Time Data Cache) and external to the data-warehouse by fully avoiding any possible problem of performance issue. The Real Time Data Cache can be simply another dedicated data-warehouse loading of the storage and data processing. The applications which handle a large volume of data or which ask for a very short processing time could bene t from this solution. The major disadvantage of this solution is that it involves an extra database that must be installed and maintained. The work done to realize this approach is higher, but the costs that would be spent to buy more e cient equipment or additional memory are justified in many cases using this method.

The authors proposed some challenges and possible solutions for near real-time ETL. They identified two problems for each extraction, transformation and loading phase. In the extraction phase, there are problems, such as the integration of multiple heterogeneous data sources solved using change data capture with the stream processor and data integration tools.

The second prob-lem of this phase is the overload data source using a service of updating. In the transformation phase, the authors identified two problems, the first is master data overhead and the second focuses on the need immediate server to aggre-gate data. The solution for the first is to maintain a master cache of data and database queue. However, instead of transforming and loading the data, the so-lution to the second problem consists in switching both tasks, in other words, loading data first, and then transforming them, which could possibly reduce the time consumed for aggregation. During the loading phase, problems are: per-formance degradation and OLAP internal inconsistencies.

Among the solutions who allows to solve these problems are: the staging table (Trickle and Flip), a multiple stage trickle and ip, organizing outside the DW update period, snap-shot data, Real Time Data Cache and layer-based view have been proposed. Apart from this work not much work has been done on the investigation ETL challenges in near real time.

4. REAL TIME DATA WAREHOUSE ARCHITECTURES

4.1 RTDW Architecture Proposed by Ricardo Jorge Santos

The conceptual architecture of real-time centralized data-warehouse. in real time data-warehouse dba are able to manage real time database and could monitor the data loading execution to provide the fresh data to business and where user can see the fresh data to make the decision. and able to build the original dw-schemas and will able to set up all the values when tool is used first time as properly as reoptimizing any database. the RTDW database having the information data warehouse tables and different information structures indexes materialized views etc definitions of all the related transactions during the business process how frequently is new statistics to be loaded which is the current accessible database for users boolean flags indicating if any database is being reoptimized queried or updated etc. the RTDW tool loader executes dw refreshment procedures loading new information into its databases. the RTDW tool query executor handles consumer queries deciding on which replicated database to use. it without a doubt redirects the requested queries according to which is the accessible database for querying defined in the tool's database providing the results to the query's requested by user to see the information.

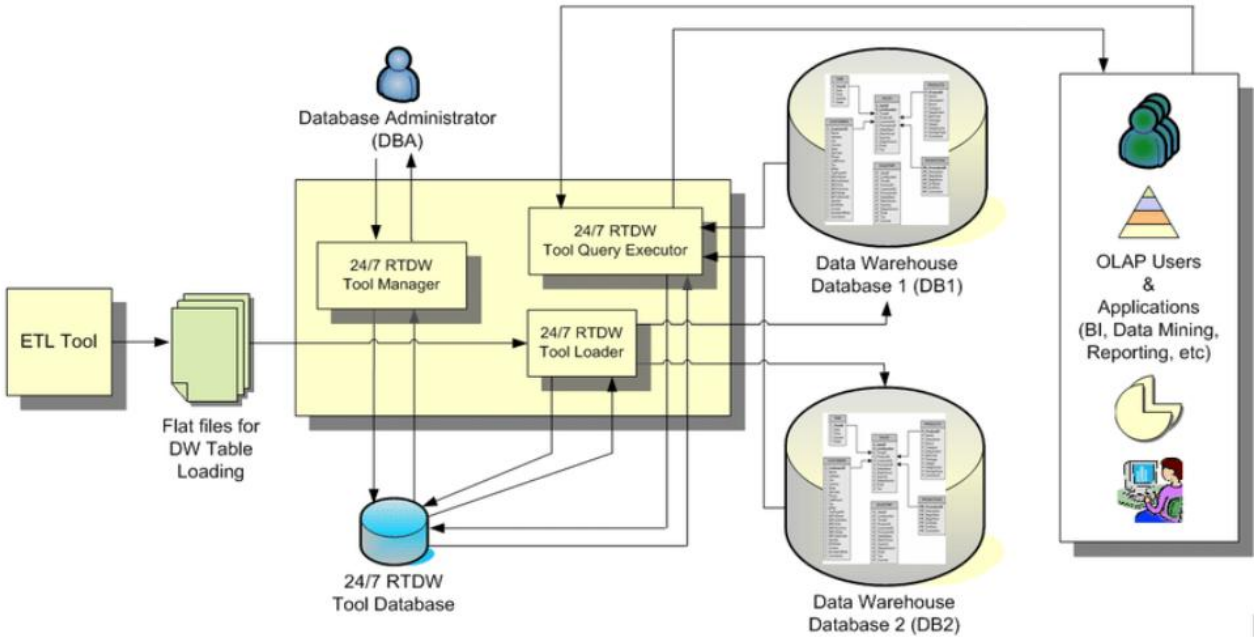

Fig 1:**The Proposed Architecture of real time data warehouse[4].**

4.2 Agile Data Warehouse Automation RTDW Architecture:

The pattern of real-time-data need the data-warehouse having timely records of required data based totally on a non-stop efficient information to gain the process. implementing such a procedure with some domestics and usual etl-software could be complicated high priced and not efficient. this architecture is an progressive data-warehouse software platform that manage more efficiently of full records warehouse-lifecycle to aid real-time records in data-warehouse. it dramatically reduces the time-costs and risks of statistics. read the antiunity compose statistics sheet. antiunity replicate provides: heterogeneous coverage. support many sources inclusive of all major rdbms sap and mainframe platforms high overall performance and low risk. optimized and licensed integration for all records in warehouse-platforms such as amazon rdbms redshift azure sql dw snowflake google big query and many others.

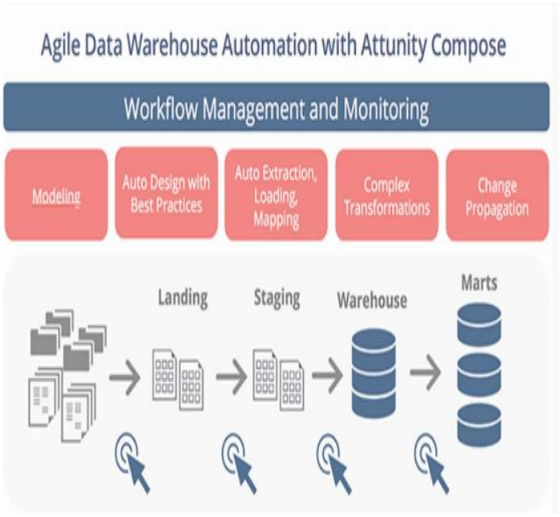

Fig 2. Architecture of Real time DWH[5].

4.3 Federated Data Warehouse (FDW) Architecture:

The federated-word means merge different sources-data into the data-storage. and this word refers to no-of-states or organizations formed in a single centralized unit in which each state or organization has some internal freedom. fdw-architecture most likely used to improve a data-warehouse

primarily based on business requirements. the federated records data store is more based on a structure that is used to combine different archive stores to provide the separate version of reality throughout the organization. the federated records data-store must be assigned to different-platforms. the basis of a federated data-warehouse is no longer the platform on which it is built but the common business model on which it depends to have a unique information structure at all integrated disk stores. because this is the key to integrate the data of different-sources. Most federal statistical stores are used in large enterprises where their industrial enterprise extends over a large number of areas and regions of the world. these types of companies can also have it information technology divisions across different countries and so on.

Strengthen the systems. in such cases each exchange team can expand its stock of highly personal statistics to present operations and actual analysis in the unite states or their region. this type of data store is used in large organizations.

use of federated statistical store is much faster and more expensive method than the price in contrast to the architecture of the company data-warehouse. unlike the company data-storage architecture a federated data-storage architecture can be developed more flexibly based on information needs geographical time requirements and so on. the federated data-storage structure must contain some extraction techniques that are completely based on our requirements. below figure

shows a template for information storage models. in below figure having many different sources like ERP source app and E-com and then data transfer to staging area and then into federated-data-warehouse and from e-com data-source data also transferring directly into real-time data-mart and then sending to real-time reports where end-users can use it for business purposes. And from other sources data sending to data-mart from federated-data-warehouse and then sending to end-users to use it for business purpose.

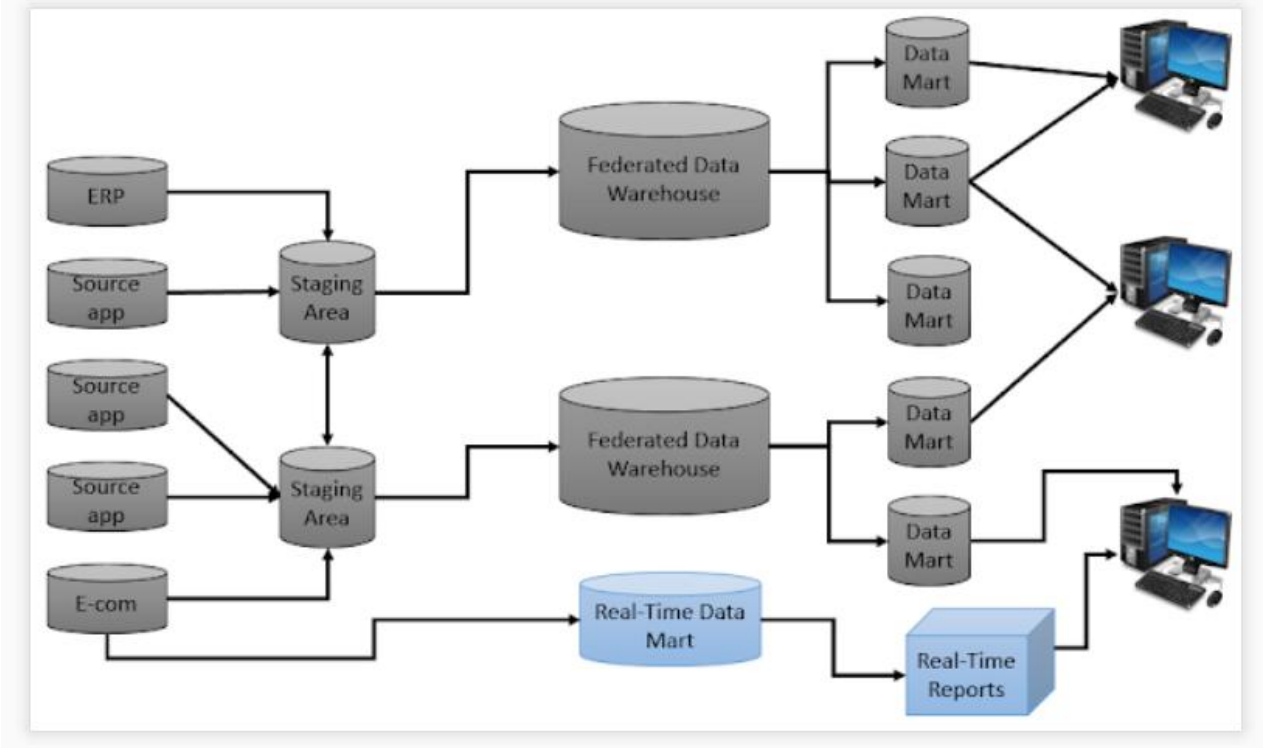

Fig 3:The proposed Architecture of Federated Data Warehouse[6].

5. DISCUSSION

Here went through in details and described Traditional-Datawarehouse different methods and ETL process to load the data from different sources into target (DWH) and went through architecture of DWH and did comparison with RTDW (Real-Time Data-Warehouse). Like many approaches Direct-Trickle, Trickle-and-Flip and External-real-time-data-cache. And many different architectures of RTDW like proposed by Ricardo Jorge Santos, Agile Data Warehouse Automation RTDW Architecture and Federated Data Warehouse (FDW) Architecture.

6. SOCIETY IMPACT OF REAL TIME DATA

Real-time information or data use up-to-date and reliable information in systematic and transactional systems to achieve a comprehensive view for end-users in business using all the data-sources together with log files databases sensors and messaging systems. it accelerates building flowing data pipelines using pre-built combination and wizards-based development. allow timely insight for better

operational decision-making. stream combines non-intrusive real-time change data capture abilities with in-flight data processing to deliver timely and enhanced data to the rest of the enterprise. it is an end-to-end enterprise-grade platform with built-in stream analytics and data visualization and delivers real-time insights while moving the data with sub-second latency. stream provides an intuitive development experience with wizard-based user interface and speeds time-to-deployment with pre-built data pipelines. using an sql-like language it is familiar to both business analysts and developers. with stream you adopt a future-proof smart data architecture for accelerated innovation. real-time analytics is the ability of a business enterprise to use all available enterprise data when needed. a crucial feature of real-time analytics is that the

available systems and setup should be able to quickly generate analytics based on the data received ideally within a minute of the data being generated. a big advantage of real-time analytics is the freshness and the context of the data. organizations can reap a lot of benefits by accessing real-time analytics purely because of their close relevance to market realities.

7. CONCLUSION

In this paper went through about tradition alandreal-time and also federated data-warehouse architecture by indicating how can find the way to getthe real-time data and review some tacticsof combination of real time data to move closer to this RTDL and did survey to review literature of existing architecture of DWH and introduced the variations in the ETL procedure to deal with real-time data-warehouse approaches and treats the combination data in the

RTDL and described the real-time data impact on society and benefits. as future work we intend to develop an tool as real time data loader RTDL which will integrate this methodology with extraction transformation and load routines for the OLTP approaches. there is also room for optimizing the query to improve the performance instructions used for our methods. we will think about a method to load the data in to RTDL from source system and real-time ETL method considering and data collection features in real time. now a days mostly company need real-time data to make some decision and and to make some prediction in warehouse and like how we can manage in best way to make it more useful.

8. REFERENCES

[1] Ravi, J., 2007. Real-Time Data Streaming Tools and Technologies – An Overview.

[2] Xenon StackInnovator:www.xenonstack.com/insights/real-time-data-streaming.

[3] Isaac, S., 2018. Real-time data processing with data streaming: new tools for a new era.

[4] Ricardo, J., Santos, J. B. and Marco, V., 2012. Leveraging 24/7 Availability and Performance for Distributed Real-Time Data Warehouses. Conference: COMPSAC 2012 - IEEE Signature Conference on Computer Software & Applications and DOI: 10.1109/COMPSAC.2012.92

[5] AttunityADivisionofQlik:www.attunity.com/solutions/data-warehousing/real-time-data-warehousing.

[6] DatawarehouseBlogs:http://blogsofdatawarehousing.blogspot.com/2017/02/federated-data-warehouse-architecture.html.

[7] Isaac, S., 2018. Real-time data processing with data streaming: new tools for a new era.

[8] Babak,Y., Seyedfaraz, Y., Nasseh, T., 2017. Developing a Real-Time Data Analytics Framework for Twitter Streaming Data.Published in IEEE International Congress on Big Data and DOI:10.1109/bigdatacongress.2017.49.

[9] Jukic, N., 2006. Modeling strategies and alternatives for data warehousing projects. Communications of the ACM, 49(4), pp.83-88.

[10] Kimball, R. and Ross, M., 2011. The data warehouse toolkit: the complete guide to dimensional modeling. John Wiley & Sons.

[11] Cuzzocrea, A., Song, I.Y. and Davis, K.C., 2011, October. Analytics over large-scale multidimensional data: the big data revolution!. In Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP (pp. 101-104). ACM.

[12] Golfarelli, M., Rizzi, S. and Cella, I., 2004, November. Beyond data warehousing: what's next in business intelligence?. In Proceedings of the 7th ACM international workshop on Data warehousing and OLAP (pp. 1-6). ACM.

[13] Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P., 2003. Fundamentals of Data Warehouses, second ed. Springer-Verlag.

[14] Felix, G. and Norbert, R., 2016. Scalable data management: NoSQL data stores in research and practice. In Conference: 2016 IEEE 32nd International Conference on Data Engineering (ICDE) and DOI: 10.1109/ICDE.2016.7498360.

[15] Wolfram, W., 2017. A Real-Time Database Survey: The Architecture of Meteor, RethinkDB, Parse & Firebase.