From Data Warehouses to Streaming Warehouses: A Survey on the Challenges for Real-Time Data Warehousing and Available Solutions

Revathy S
PG Scholar
Sri Ramakrishna Engg College
Coimbatore, Tamil Nadu, India

Saravana Balaji.B
Assistant Professor
Sri Ramakrishna Engg College
Coimbatore, Tamil Nadu, India

N.K.Karthikeyan, Ph.D
Professor & Head - IT
Sri Krishna Coll of Engg & Tech
Coimbatore, Tamil Nadu, India

ABSTRACT

Data Warehouses usually work on history data. In most cases, the Data Warehouse is loaded with data from operational or transactional systems on a weekly or nightly basis. As today's decisions in the business world are becoming real-time, it is only natural that Data Warehouse, Business Intelligence, Decision Support and OLAP systems must quickly begin incorporating real-time data. When shifting from a traditional offline and time-consuming data warehousing system to a real-time system, two important considerations are speeding up the ETL and the OLAP process. This survey looks into the various challenges involved in building a real-time Data Warehouse and some of the solutions available to overcome them.

Keywords

Real-Time Data Warehousing, Real-Time ETL, Data Stream Management Systems

1. INTRODUCTION

The advent of the decision support systems came in the early 1970's. A decision support system (DSS) is a computer based information system that supports organizational decision making activities. Decision-making has become one of the criteria's for a successful and competitive business in today's world. Sound business decisions are based on data that is analyzed according to pre-defined business criteria. Such data, used for performing statistical and analytical processing efficiently, resides within a Data Warehouse. The Warehouse stores data for a Data-driven DSS.

Beginning in about 1990, Data Warehousing and on-line analytical processing (OLAP) began expanding the horizon of DSS. According to Bill Inmon, who is often called the Father of Data Warehousing, "A Data Warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data". Another definition of a Data Warehouse is "A Data Warehouse is a copy of transaction data specifically structured for query and analysis"[1].

A data warehouse is a database, most often a relational database. It is a central repository of data created by integrating data from one or more different sources. It usually contains historical data that is derived from transactional data. It enables an organization to separate the analytical workload from the transactional workload. Fig 1 gives the various architectural components of a general Data Warehouse system [1].The data available within the data warehouse can be queried by users to perform analysis on the data.

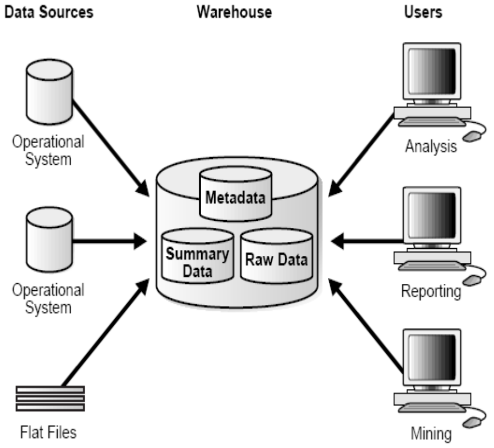

Fig:1: A General Data Warehouse Architecture

The business markets around the globe are weighed down with rapid changes. To cope up with such a business market, business level decision making must also be quick. Decision makers must adjust operational processes, corporate strategies and business models at lightning speed and must be able to leverage business intelligence instantly and take immediate action [3].

A data stream is a real-time, continuous and ordered sequence of items. The growing need for real-time data has caused a change in the data driven decision making process. Data Driven DSS is shifting from Database Management Systems to Data Stream Management Systems (DSMSs). A Data Stream Management System is a program to manage a continuous stream of data. It is similar to a Database Management System, which is, however, designed for static data . Traditional Data Warehouses employ a store-and-then-query data processing model, where data is stored fully in the database and results are provided to queries based on the data currently available within the database. In contrast, in DSMSs, monitoring applications register Continuous Queries which continuously process unbounded data streams looking for data items that represent events of interest to the end-user [4].

Table 1: Difference between a DBMS and a DSMS

Parameters	**Database management system**	**Data stream management system**
Nature of Data	Non-Volatile	Volatile data streams
Data Access	Random	Sequential
Query Nature	One-time queries	Continuous queries
Storage Capacity	unlimited secondary storage	limited main memory
Rate of Change	Low update rate	High / Continuous update rate.
Time Requirements	Little or no time requirements	Real-time requirements

A Stream Warehouse is a Data Stream Management System (DSMS) that stores a very long history, e.g. years or decades; or equivalently a data warehouse that is continuously loaded [6]. It can also be called as a data warehouse which is updated as and when data arrives. Table 1 gives the difference between a DBMS and a DSMS. Many systems generate data in streams. A few examples of such streaming systems are stock trading companies, network monitoring systems and traffic monitoring systems. These systems are hugely dependent on the results obtained from processing the data streams. The faster the result is obtained, the more efficient is the system performance.

Data Stream Management Systems provide real-time analysis by processing the events currently in-memory over a short time frame. Users would also require long-term analysis of data over large time frames. A Stream Warehouse bridges the short-term vs. long-term gap by loading data continuously in a streaming fashion and warehousing them over a long time period [8].

2. REAL – TIME DATA WAREHOUSING CHALLENGES

A typical data warehouse system consists of a staging area, where the data from the different sources are brought together. The staging area is often contained in a different database, most commonly referred to as the staging database. The data from the staging area is integrated, cleansed and transformed. The transformed data is then moved to another database, often called the warehouse database, where the data is modeled into tables called facts and dimensions.

The data from the staging area is processed using the Extraction, Transformation and Load (ETL) process. The most recent data is pushed into the data warehouse database by the ETL process. Almost all the ETL tools and systems, available in the market today, operate in batch mode. For batch mode operation, the data should be available in an extract file in a specific format. This data will be taken up by the ETL process and transformed according to the specified business rules and loaded into the Warehouse database. This operation usually requires the Data Warehouse to be offline. In a real time system, even a small amount of down time is not acceptable. The more the amount of incoming data, the more will be the usage of the warehouse data.

2.1 Enabling Real-Time ETL

Existing ETL systems can be modified to perform real-time or near real-time warehouse loading. Some of the techniques described in [5] are :

2.1.1 Near Real-time ETL

The simplest way to solve the real-time ETL problem is to not perform ETL. The system should be analyzed thoroughly to check whether the cost of performing ETL for the system can be justified. Some systems will only require that the data available in the Warehouse be the most recent data. For such systems, the Warehouse refresh period can be changed from a weekly basis to a daily or hourly basis. This will provide users access to fresher data without having to bear the brunt of the ETL or the reporting processes

2.1.2 Real Time Facts

The Data Warehouse maintains two types of tables, Facts and Dimensions. In most of the cases, the actual querying happens on the fact tables. The second approach is to create a separate partition for the fact tables corresponding to real-time data. The fact tables in the real-time partition can then be directly inserted or updated with data from the source system.

Real-time data loading packages that are specifically designed for this approach are available. Packages are available from DataMirror and MetaMatrix. Tibco provide solutions for real-time data transport. For systems based on the latest Java technologies, Java Messaging Service (JMS) can be used to transmit each new data element from the source system to a lightweight listener application that in turn inserts the new data into the warehouse tables. For data that is received over the Internet, the data can be transmitted in XML via HTTP using the SOAP standard.

2.1.3 Trickle & Flip

In the Trickle and Flip approach staging tables that are an exact replica of the warehouse tables are created. Data is continuously loaded into these staging tables. On a periodic basis, the actual warehouse tables are swapped with these staging tables thus bringing the Data Warehouse instantly up-to-date

2.1.4 External Real-time Data Cache

Another approach is to store the real time data outside of the Data Warehouse in an external real-time data cache (RTDC). This approach will avoid any performance problems to the existing warehouse. The RTDC can simply be another dedicated database server or a separate instance of a database system. Data Warehouse systems dealing with large volumes of real-time data or those that require extremely fast query performance can benefit from using an in-memory database (IMDB). Such IMDBs are provided by companies such as Angara, Cacheflow, Kx, TimesTen, and InfoCruiser.

2.2 Enabling Real-Time Business Intelligence

The analysis of the data that has been extracted, transformed and loaded into the warehouse tables can be done by using Data Mining or Online Analytical Processing (OLAP) tools. OLAP can be defined as the general activity of querying and presenting text and number data from data warehouses [2]. OLAP consists of a set of tools and techniques that allows users to query databases and analyze the data. They aid in preparing reports based on the data in the Data Warehouse. These reports are used by managers and business analysts for making business decisions.

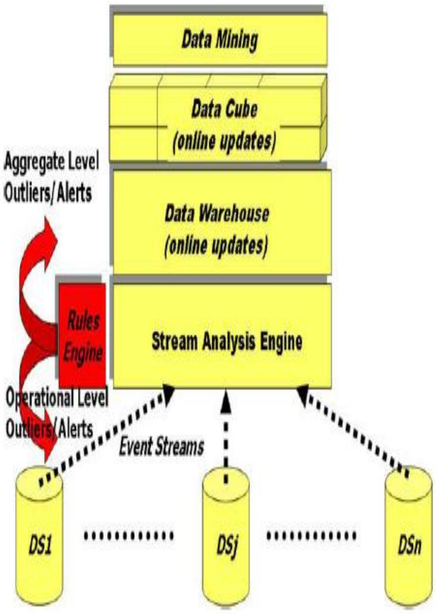
Data Mining
Data Cube (online updates)
Data Warehouse (online updates)
Stream Analysis Engine
Rules Engine
Aggregate Level Outliers/Alerts
Operational Level Outliers/Alerts
Event Streams
DS1
DSj
DSn

Fig 2: Architecture for a Real Time Business Intelligence System.

One of the methods proposed for real-time business intelligence to stream data in real-time from the source systems to the data warehouse is [9] where a component referred to as the stream analysis engine. The goal of the stream analysis process is to extract crucial information in real-time and then have it delivered to appropriate action points which could be either tactical or strategic .Fig 2 gives an overview of the Stream Analysis Engine[7]. This stream analysis engine performs in-depth analysis on the incoming data to identify interesting patterns.

2.2.1 Stream Analysis Engine

In this approach, query processing is done on continuously arriving data streams or event streams. The arrival of streams triggers the query processing. Fig 3 gives an overview of a complex event processing server architecture [9]. The drawback in this architecture is again the performance of the Data Warehouse. Continuous running queries may reference data in the database and have an impact on the near real-time requirements

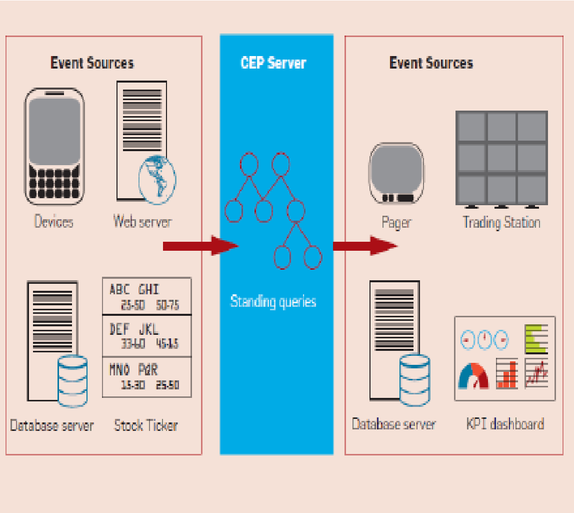

Fig 3: Complex Event Processing Server Architecture

3. DATA STREAM MANAGEMENT SYSTEMS

Customary Data Warehouses are refreshed with the latest available data only on a periodic basis. The period of refresh ranges from weekly to daily. Analysis happens on terabytes of data; but the most recently arrived data will not be available within the Warehouse. Data Stream Management Systems, however, support analysis on the latest available data in real-time. But the amount of data on which the analysis can be done is very limited.

A Streaming Warehouse combines the historical analysis of the Data Warehouse and the real-time analysis of the Data Stream Management Systems. The goal of a streaming warehouse is to propagate new data across all the relevant tables and views as quickly as possible [6]. If the most recent data is available in the Warehouse immediately, then analysis can be done on the latest data and business decisions can be taken accordingly in real-time.

4. CONCLUSION

As enterprises, such as E-commerce sites, are increasingly facing the need for real-time Business Intelligence and Predictive Analytics, the need of the hour is to build ETL tools, which will provide real-time data into Data Warehouses.

The trade-off between the cost of real-time Data Warehousing and the actual requirement for such an analysis calls for serious research and consideration. Otherwise the resulting system may have prohibited costs associated with it [5]

The underlying technology components and custom solutions for real-time data warehousing are excessively expensive. The importance, complexity and criticality of such an environment make real-time BI and DW a significant topic of research and practice. Therefore, these issues need to be addressed in the future by both the industry and the academia [9].

5. REFERENCES

[1] Oracle Data Warehousing Guide – Oracle Documentation,docs.oracle.com/cd/B28359_01/server.111/b28313.pdf by P Lane.

[2] The Data Warehouse Lifecycle Toolkit ,Ralph Kimball, Margy Ross ,Warren Thornthwaite ,Joy Mundy , Bob Becker , John Wiley & Sons; 2nd Edition.

[3] Dr. Kamal Kakish Dr.Theresa A.kraft , ETL Evolution for Real-Time Data Warehousing,2012, Proceedings of the Conference on Information Systems Applied Research ,New Orleans Louisiana, USA , ISSN:2167-1508 , v5 n2214.

[4] Mohamed A. Sharaf, Alexandros Labrinidis, Panos K. Chrysanthis , ETL Scheduling Continuous Queries in Data Stream Management Systems, ACM 978-1-60558-306-8/08/08.

[5] Langseth, J., "Real-Time Data Warehousing: Challenges and Solutions", http://dssresources.com/papers/features/langseth/langseth02082004.html

[6] Lucas Golab, Theodere Johnson , Vladislav Shkapenyuk Scalable, Scheduling of Updates in Streaming Data Warehouse, IEEE Transactions on knowledge and data engineering ,Vol. 24, N0. 6, JUNE 2012.

[7] Agrawal , D., The Reality of Real-Time Business Intelligence, Proceedings of the 2nd International Workshop on Business Intelligence For the Real Time Enterprise (BIRTE 2008), Springer , LNBIP 27 , 75-88.

[8] Lukasz Golab and Theodore Johnson, Consistency in a Stream Warehouse, 5th Biennial Conference on Innovative Data Systems Research (CIDR '11) January 9-12, 2011, Asilomar, California, USA.

[9] Chaudhuri, S., Dayal, U., Narasayya, V., (2011) An overview of Business Intelligence Technology, Communications of the ACM, 54(8), 88-98.