MAGNITUDE™

RapidDecision®

White Paper

The Challenges of Creating a Real Time Data Warehouse

RapidDecision, A Likely Solution

MAGNITUDE™
RapidDecision®

Executive Summary

In today's world, data volumes are growing exponentially. At the same time, businesses need accurate and operational business intelligence more than ever before. To achieve this operational intelligence, the key is to keep the data warehouse in sync with the source systems in real time. The traditional approach of loading the warehouse using periodic batch jobs (hourly, daily, weekly, or monthly) is proving costly; these periodic batch jobs now require more resources because they must process more data. Therefore, a real time data warehouse is a better solution. But there are implementation challenges.

In a perfect world, the data in the warehouse would instantly reflect changes from source databases. In reality, some delay is inevitable. A delay of up to a few minutes for synchronization of a warehouse with its sources is almost always acceptable. Longer delays, however, can create problems. This paper refers to any data warehouse that maintains a short and acceptable delay as a real time data warehouse. This paper also gives a glimpse into how Magnitude RapidDecision can be used as a real time data warehouse to address your BI needs.

Key Challenges

Change Data Capture (CDC):

The process used to achieve real time synchronization is Change Data Capture, or CDC. Many ways to achieve CDC have been tried over the years.

Database Triggers: This DBMS feature invokes a prewritten routine each time a specific set of conditions is met, such as adding or updating of a record in the database. The trigger can write a record of the transaction to a database table and the ETL tool can then poll these tables on a periodic basis. However, the triggers can regularly be deleted or disabled and then added again or reenabled in the normal course of a business operation. Triggers also place a relatively high overhead burden on the source database server.

Message Queues: Middleware products such as IBM MQ Series can capture application changes, and not database changes, and report them to an ETL tool. A disadvantage is the cost of the license for these products. More importantly, message queues are not a completely reliable source of change information because they are only aware about data changes that are sent to them by the applications and not batch routines or manual updates to the database.

Date & Time Stamps: Many ERP applications and other data sources maintain data fields within each record that indicate when the record was last changed. CDC is achieved by reading through these data records and looking for recent changes. The flaw in this approach is that it relies on the programs that change data to unfailingly update this field. This approach can also lose track of deletions because the entire record, including the time stamps, are deleted.

Log Based CDC: Only one approach has stood the test of time. Log-based CDC is an updating technique that ensures that all changes to the source databases are reflected in the data warehouse. It is based on obtaining data from log files created by the database

management system of the source computer. Log files contain a complete record of all changes that were made with a date and time stamp indicating when each change occurred. These log files can be used to repair the source database in the event of failure. A data warehouse that is synchronized with its data sources without fear that changes might be lost can thus be created.

Extract, Transform, and Load (ETL)

Most data warehouses are built using a type of software tool named for what it does: Extract (obtain data from a source server), Transform (reformat this extracted data) and Load (update the data warehouse with this reformatted data) or ETL. Popular tools in this category include Informatica PowerCenter, IBM® InfoSphere® DataStage®, SAP Data Services, and Oracle Data Integrator.

Not all these tools can perform log-based CDC on their own. ETL tools that do not support log-based CDC have to be combined with another software tool that adds this capability. Most major vendors have acquired software firms that offer a log reading capability, and then these vendors have either integrated the log-based reader into their ETL tools or offer this as an optional complementary product. Therefore, it must not be assumed that vendors whose product line includes a log-based CDC capability will automatically include it in their proposed solution.

If you are about to undertake a BI project, you must fully understand the impact that real time updating will have on the data model design. If a warehouse is first implemented using daily batch updates, then a significant redesign might later be necessary to make log-based CDC workable.

Best Practices

An examination of the way log-based CDC works makes it clear why this approach is vastly superior to any alternative. The ETL tool that is used to organize and manage the entire process of building data warehouses becomes the crux of such a solution. Some ETL tools do not have the ability to perform the extract operation directly from database log files; they can only extract information directly from the database. Such tools must be combined with another tool that offers this capability if log-based CDC is going to be implemented. Each DBMS has its own unique way of creating log files. A driver or software interface to each DBMS from which source data must be obtained will thus be needed. Commonly used database management systems include Microsoft SQL Server, Oracle, MySQL and IBM Db2. The version of Db2 used with IBM's iOS differs considerably from the other versions and is not supported by some ETL tools.

When a real time data warehouse is being created, the first step is a one-time process to populate the warehouse initially. This process of extracting every data item from the source databases, processing all the required information, and then loading the data into the data warehouse can take hours, depending on the size of the files involved and the power of the servers. It is important to note that this time-consuming process only needs to happen once in a CDC-fed data warehouse.

With the log-reading drivers in place, the ETL/CDC tool can now be switched on. The log-reading drivers poll the log files and use them to identify all changes that have occurred. These changes are sent to the ETL tool for transformation and loading. When the ETL tool is ready to handle the next

wave of update transactions, another polling cycle is initiated. The polling cycle takes place every few seconds and can be changed as required.

The log reading logic thus handles the extraction phase of the ETL process. It feeds the change transactions to the transformation function. The source data can be made more usable by BI analysis tools. The names of data fields are often changed to make them both understandable and consistent.

The final function is loading the data warehouse. The transformed data is normally organized differently than it was in source databases. Combinations (de-normalization) of tables are frequently performed to make access faster and easier.

Design Considerations

Transformation and loading are complex procedures and there can be occasions where an unusually heavy volume of change transactions can cause the ETL process to fall behind the log file. In these cases, update transactions are not lost, just delayed.

Designers of log-based CDC data warehouses must determine how much server capacity is required to handle such intermittent bursts of change traffic to limit the duration and frequency of delays. Real time data warehouses must be designed to continuously monitor their own level of synchronization and alert administrators and users if pre-set service levels are not met.

A well-designed log-based CDC system will have no noticeable impact on performance of the source servers from which data is being extracted. In most cases it is very important to avoid slowing down the process of updating source servers because the transactions they handle are often vital to the ongoing operation of the entity. Log-based CDC minimizes impact on source transaction servers.

Magnitude Solution: RapidDecision

RapidDecision is a packaged software offering from Magnitude Software that addresses your real time data warehouse needs. RapidDecision evolved out of custom data warehouses that were built for a variety of businesses in varied industries. This product is a pre-built real time data warehouse built around a highly flexible and expandable data model. Packaged versions of it are sold for Oracle JD Edwards and Oracle E-Business Suite.

RapidDecision uses SAP Data Services to perform ETL Operations and this tool can handle the log reading for Microsoft SQL Server and Oracle DBMS sources. When source data resides on an IBM server using the iOS version of Db2, then a third-party log reader is included as well.

Every implementation of RapidDecision uses log-based CDC as its synchronization approach, and therefore RapidDecision is less expensive and simpler to install and operate than warehouses that use other approaches. In several cases, RapidDecision has replaced failed data warehouse efforts. The real time capability of RapidDecision was often the key reason it was selected. In every case, performance has improved dramatically.

A unique feature of RapidDecision is called its heartbeat. This feature makes it possible for RapidDecision to continuously monitor the lag time between posting of changes to source logs and the completion of the resulting updates to the data warehouse. Alerts can be sent to administrators when pre-set limits are exceeded or when the updating process is interrupted for any reason, and immediate action can then be taken. You can have the monitoring and response function managed by the RapidDecision support staff.

RapidDecision typically can become fully operational within a few weeks, even for buyers implementing large numbers of RapidDecision data marts. More information about RapidDecision and data warehouse design services can be found at www.magnitude.com.

MAGNITUDE™
RapidDecision®

Dive deeper into this topic or contact us today at info@magnitude.com or **1.866.466.3849.** Visit our website: **www.magnitude.com**

About Magnitude

Magnitude's transformative approach to unified application data management delivers vast operational efficiencies to business application data access, management, analytics and reporting for the modern enterprise. Magnitude's portfolio of products includes: simplified application data access to any data source; data management solutions for the SAP and commerce verticals; simplified master data harmonization and governance; and packaged application analytics and reporting solutions for SAP and Oracle. The company helps thousands of business users simplify management of their data and deliver on the substantial productivity gains these applications originally promised.