Change Data Capture Efficient ETL for Real-Time BI

Ankorion, Itamar . DM Review ; New York Vol. 15, Iss. 1, (Jan 2005): 36.

ABSTRACT (ABSTRACT)

Cost reduction. Bulk ETL operations are costly and inefficient, as they require more processing power, more memory and more network bandwidth. In addition, as bulk ETL processes run for long periods of time, they also require more administration and IT resources to manage. To stay ahead of these changing business conditions and increase the value of real-time BI implementations, a new and intelligent approach for ETL is required. The power behind it is change data capture.

FULL TEXT

Business intelligence (BI) is at the heart of the best global organizations, enabling them to understand business trends, improve decisions and support day-to-day operations. ETL (extract, transform and load) is the process that enterprises use to build the consolidated data stores (e.g., data warehouses and data marts) required for effective BI. Traditionally, ETL processes have been run periodically, on a monthly or weekly basis, and use a bulk approach that moves and integrates the entire data set from the operational source systems to the target data warehouse. While this approach was acceptable for enterprises over the years, current business conditions require a new way of integrating data - in real time and in an efficient manner.

Changing Business Conditions Demand a Change of Approach

* Business globalization and 24x7 operations. In the past, enterprises could stop online systems during the night or weekend, to provide a window of time for running bulk ETL processes. Today, running a global business with 24x7 operations means smaller or no downtime windows. * Need for up-to-date, current data. Customer demand, competitive pressure and improved decisions require timely information. To make the most of BI in today's ever-accelerating business climate, managers should not be working with last week's or yesterday's data. Today, decision-makers need data that is updated a few times a day or even in real time.

* Data volumes are increasing. As time passes and the business grows, data volumes in operational data stores become larger. The larger the data volumes become, the more resources and time are required by the ETL processes. This trend challenges the bulk extract windows that are getting smaller and smaller.

* Cost reduction. Bulk ETL operations are costly and inefficient, as they require more processing power, more memory and more network bandwidth. In addition, as bulk ETL processes run for long periods of time, they also require more administration and IT resources to manage. To stay ahead of these changing business conditions and increase the value of real-time BI implementations, a new and intelligent approach for ETL is required. The power behind it is change data capture. Change data capture (CDC) is an innovative approach to data integration that is based on the identification, capture and delivery of the changes made to enterprise data sources.

Change Data Capture Meets Business Demands

CDC is changing data integration as we know it today, making integration efficient and real-time. CDC means:

* No downtime window for ETL. CDC enables organizations to move the changes made to enterprise data while the operational systems are running, without the need for a downtime window. This also means minimal to no intrusion to operational systems and no degradation in their performance and service levels.

* Current, up-to-date data. By constantly identifying changes, CDC delivers new data more frequently - even in real

time - providing timely information for enterprise users and decision- makers.
* Reduced cost. By moving only the changed data, CDC requires significantly fewer resources, a fraction of what bulk ETL typically requires. Cost is reduced in hardware, software and human resources.
Solution Components CDC solutions are designed to maximize the efficiency of ETL processes, minimize resource usage by replicating/moving only changes to the data (i.e., the deltas) and minimize the latency in the delivery of timely business information to the potential consumers. A CDC solution is based on the following key components:

* Change Capture Agents * Changed Data Services
* Change Delivery

Change Capture Agents
Change capture agents are the software components that are responsible for the identification and capture of changes to the source operational data store. Change capture agents can be optimized and dedicated to the source system (i.e., typically using database journals, triggers or exit hooks) or by using generic methods such as data log comparison.
Change Data Services
Change data services provide a set of functions critical to achieving successful CDC, including but not limited to: filtering (e.g., receiving only committed changes), sequencing (e.g., receiving changes based on transaction/unit of work boundaries, by table or by timestamp), change data enrichment (e.g., add reference data to the delivered change for further processing purposes), life cycle management (i.e., how long will the changes be available for consuming applications) and auditing that enables monitoring of the system's end-to-end behavior, as well as the examination of trends over time.
Change Delivery
Change delivery mechanisms are responsible for the reliable delivery of changed data to change consumers - typically an ETL program. Change delivery mechanisms can support one or more consumers and provide flexible ways by which the changes can be delivered including push and pull models. A pull model means that the change consumer asks for the changes on a periodic basis (as frequently as needed, typically every few minutes or hours), preferably using a standard interface such as ODBC or JDBC. A push model means that the change consumer listens and waits for changes, and those are delivered as soon as they are captured, typically using some messaging middleware. Another important function of change delivery is the ability to dynamically go back and ask for older changes for repeated, additional or recovery processing.
CDC Components in Action CDC offers companies the flexibility to handle the data integration challenges particular to its business needs. Following are two sample scenarios that highlight how organizations can leverage CDC.
Sample Scenario 1: Batch-Oriented CDC (pull CDC)
In this scenario, an ETL tool periodically requests the changes, each time receiving a batch of records that represent all the changes that were captured since the last request cycle. Change delivery requests can be done in low or high frequencies (e.g., twice a day or every 15 minutes). For many organizations, the preferred method of providing extracted changes is to expose them as records of a data source table. This approach enables the ETL tool to seamlessly access the changed records using standard interfaces such as ODBC. The CDC solution needs to take care of maintaining the position of the last change delivery and deliver new changes every time.
This scenario is very similar to traditional bulk ETL, except that it processes only the changes to the data instead of the entire source data store. This approach greatly reduces the required resources and eliminates the need for a downtime window for ETL operations.
When should organizations use this approach? This batch-oriented approach is very easy to implement, as it is

similar to traditional ETL processes and capitalizes on existing skill sets. Organizations should use this method when their latency requirements are measured in hours or minutes.

Sample Scenario 2: Live/Real-Time CDC (push CDC)

In this scenario, which accommodates near real-time or real-time latency requirements, the change delivery mechanism pushes the changes to the ETL program as soon as changes are captured. This is typically done using a reliable transport such as an event-delivery mechanism or messaging middleware. Some CDC solutions use proprietary event delivery mechanisms, and some support standard messaging middleware (e.g., MQ Series). Note that while message- oriented or event-driven integration is more common in EAI products (i.e., using tools such as Integration Brokers), many of the leading ETL tool vendors are offering such capabilities in their solutions to accommodate the demands of high-end, real-time BI applications. This real-time approach is required when the BI applications demand zero latency and the most up-to-date data.

Change Data Capture Technical Considerations

While CDC seems to offer significant advantages, there are several factors that need to be considered and evaluated, including:

Change Capture Technique. Change capture methods vary, and each has different implications on the overall solution latency, scalability and level of intrusion. Common techniques for capturing changes include reading database journals or log files, usage of database triggers or exit hooks, data comparison and programming custom event notifications within enterprise programs.

Level of Intrusion. All CDC solutions have a certain degree of system impact, making intrusion a critical evaluation factor. The highest degree of intrusion is "source code" intrusion that requires changes to be made to the enterprise applications that make the changes to the data stores. A lesser degree of intrusion is "in- process" or "address space" intrusion, which means that the change capture solution affects the operational system resources. This is the case when using database triggers and exit hooks because they run as part of the operational system and share its resources. The least intrusive solution does not affect the operational data sources of applications. Using database journals as the source of change capture is an example of such a case.

Capture Latency. This factor is a key driver for choosing CDC in the first place. Latency is affected by the change capture method, the processing done to the changes and the choice of change delivery mechanism. As a result, changes can be streamed periodically, in high frequency or in real time. One should note that the more real- time the solution is, the more intrusive it typically is as well. Yet another point to consider is that different BI applications will have different latency requirements, and thus enterprises should look for CDC solutions that support a wide range of configurations.

Filtering and Sequencing Services. CDC solutions should provide various services to facilitate the filtering and sequencing of delivered changes. Filtering helps to guarantee that only the needed changes are indeed delivered, for example: an ETL process will typically need only the committed changes. Another example is the ability to discard redundant changes and deliver the last change to further reduce processing. Sequencing defines the order by which changes are delivered. For example, some ETL applications may need changes on a table by table basis, while others may want the changes based on units of work (i.e., across multiple tables).

Supporting Multiple Con-sumers. Captured changes may need to be delivered to more than one consumer, such as multiple ETL processes, data synchronization applications and business activity monitoring. CDC solutions need to support multiple consumers, each of which may have different latency requirements.

Failover and Recoverability. CDC solutions need to guarantee that changes will be delivered correctly, even when system, network or process failures occur. Recovery means that a change delivery stream can continue from its last position and that the solution keeps transactional integrity to the changes throughout the delivery cycle.

Mainframe and Legacy Data Sources. BI is only as good as the data it relies on. Analysts estimate that mainframe systems still store approximately 70 percent of corporate business information, and mainframes still process most of the business transactions in the world. Mainframe data sources also typically store higher volumes of data, further increasing the need for a more efficient approach to moving data such as change data capture. In addition,

popular mainframe data sources such as VSAM, which are non-relational, present additional challenges when incorporating that data into BI solutions. As ETL and DW tools expect relational data, the non- relational data needs to somehow be mapped to a relational data model.

Seamless integration with ETL tools. When choosing a standalone CDC solution, enterprises should consider the ease of interoperability with its ETL program (off-the-shelf or homegrown). Standard interfaces and plug-ins can reduce risk and speed the data integration project.

While change data capture can be implemented in house, it is a complex solution with many considerations that must be addressed. Off-the-shelf change data capture software is available today and can speed project delivery and reduce its cost and risk.

Changing business requirements demand that IT organizations deliver real-time business intelligence based on timely data while, at the same time, reducing the cost of data integration. Supporting both current and future initiatives, change data capture is an approach to data integration that delivers such a solution by moving only the changes to enterprise data sources, dramatically reducing the required resources and its associated costs, and increasing the timeliness of the data in the data warehouse. For organizations looking for ways to meet these demanding business needs, create an event-driven enterprise and provide real-time business intelligence, change data capture is a key component in the data integration architecture. Itamar Ankorion is the director of product management and marketing at Attunity, a provider of enterprise data integration software. He has extensive experience in enterprise integration including data integration, application integration and service-oriented technologies. He can be reached at (781) 213 5220 or via e-mail at itamar.ankorion@attunity.com.

DETAILS

Company / organization:	Name: Centers for Disease Control; NAICS: 923120; SIC: 9400
Publication title:	DM Review; New York
Volume:	15
Issue:	1
Pages:	36
Number of pages:	0
Publication year:	2005
Publication date:	Jan 2005
Section:	Change Data Capture
Publisher:	SourceMedia
Place of publication:	New York
Country of publication:	United States, New York

Publication subject:	Computers--Data Base Management
ISSN:	15212912
Source type:	Scholarly Journals
Language of publication:	English
Document type:	PERIODICAL
ProQuest document ID:	214690875
Document URL:	https://search.proquest.com/docview/214690875?accountid=10472
Copyright:	(Copyright c 2005 Thomson Media. All Rights Reserved.)
Last updated:	2011-09-01
Database:	Business Premium Collection

LINKS

Check SFX for Availability