Real Time Data Integration

Using Change Data Capture Technology with Microsoft SSIS

An Attunity White Paper

Change data capture (CDC) technology has become a strategic component of many data warehouse and business intelligence (BI) architectures. Given today's ever increasing struggle for more immediate access to real time data and information, constant pressure on efficiency costs, and the exponential growth of underlying data volumes, CDC is now a 'must have' for the modern BI and data warehouse project. Furthermore, when combined with the very comprehensive, powerful and competitive ETL offerings of Microsoft SQL Server Integration Services (SSIS), the cost-benefits of an architecture based on a combination of CDC and SSIS are now hard to ignore.

Real Time Data Integration (using SSIS and CDC)
White Paper – July 2008

Attunity Ltd. follows a policy of continuous development and reserves the right to alter, without prior notice, the specifications and descriptions outlined in this document. No part of this document shall be deemed to be part of any contract or warranty. Attunity Ltd. retains the sole proprietary rights to all information contained in this document. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopy, recording, or otherwise, without prior written permission of Attunity Ltd. or its duly appointed authorized representatives.

Table of Contents

ETL Tools and Microsoft SQL Server Integration Services

Business Intelligence (BI) systems are now a core part of any corporate IT department, providing the business with key information, improving decision making and supporting day-to-day operations. Underpinning such applications are information systems and databases organized into the well-defined structures that the BI applications need in order to run. These underlying database structures are typically and collectively known as data warehouse (DW) systems (although from a purist's perspective only a subset probably is).

There are many reasons that data is stored in these separate data warehouses, such as:

- Performance – having a separate dedicated database structured to better support known database queries typically means better performance of those queries. For example it is common practice to create aggregates and summaries of data within the data warehouse structure itself

- Impact on transactional systems – by separating BI requirements and decisions support away from the business's operational/transactional systems means those systems are isolated from any resource demands required by the BI systems

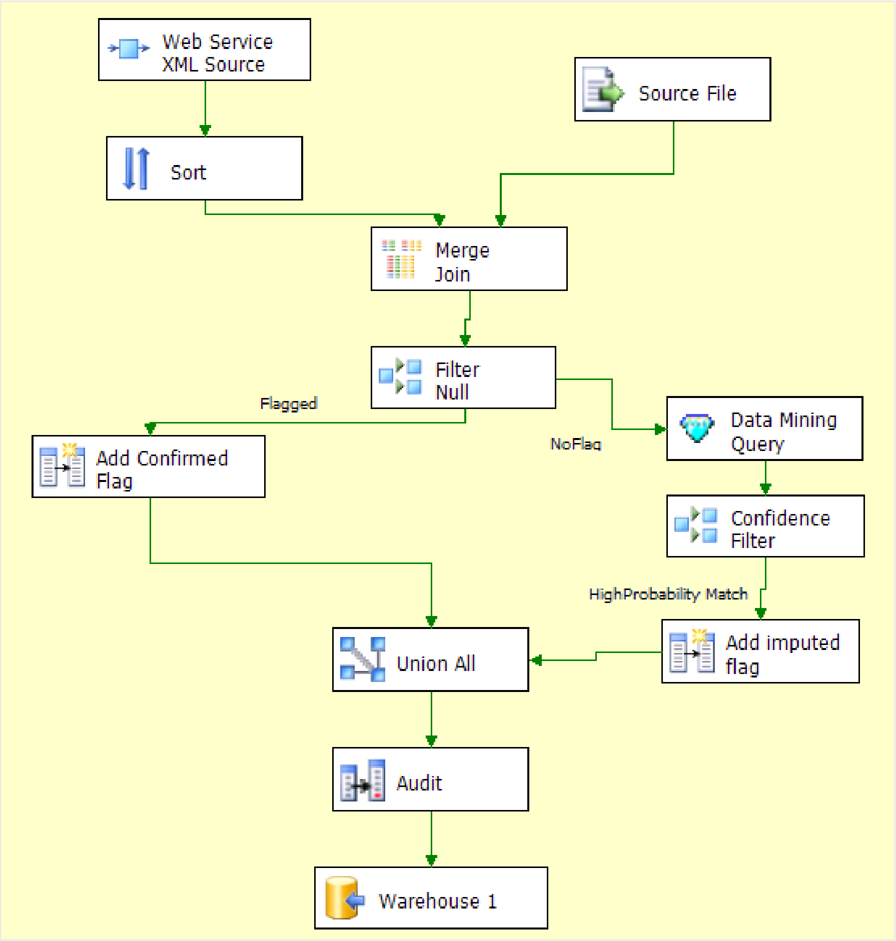
- Collection of historical data – transaction systems typically just have current data. By storing 'snapshots' of this data collectively over a period of time allows historical data to be built

What all data warehouse projects have in common however is the challenge of getting data from the source (transactional) systems of the business into the structures of the data warehouse. Depending on the specific environment, as well as the complexity and number of source systems this process, known as Extract, Transform and Load (ETL), can be time-consuming, costly and very difficult. As a result an entire market has grown around the software companies that build and supply the tools to attempt to make the process easier (the ETL tools market).

ETL tools have now become a critical component in overall integration architectures, especially for business intelligence and data warehouse projects. The last few years have seen a number of software vendors emerge and grow on the back of the ETL market, typically with software license price tags running to the tens of thousands and often hundreds of thousands of dollars.

However, when Microsoft announced SQL Server 2005, they introduced a free new product component as part of the server suite, called SQL Server Integration Services (or SSIS for short). This announcement heralded Microsoft's first serious entry into the ETL market. It was a bold statement of their intentions, and overnight changed the dynamics of the previously unchallenged ETL vendors. As a new Microsoft SQL Server Business Intelligence application, and the successor to Microsoft's previous Data Transformation Services product, it has become a platform for a new generation of high-performance data integration solutions.

Now with on its second major release, in conjunction with SQL Server 2008, SSIS is a serious contender in any ETL and data warehouse project design, especially when the (free) cost is brought into consideration.



The Need for Change Data Capture

Traditionally, ETL processes run on a periodic basis (weekly, daily), and bulk transfer source data to a target data source (the data warehouse). However, in today's fast moving world, where immediacy of information is no longer nice-to-have but is becoming absolutely critical for business competitiveness and survival, historical data is no longer enough. Furthermore, with the exponential growth of raw data volumes over the past years (some estimate doubling every 18 months on average), the previously simple process of transferring in bulk the data from source to destination, has now become unwieldy and far too time consuming.

A new approach is required; one that provides immediate transfer, or 'streaming', of data from operational systems through the transformation and load processes of ETL to the destination databases. And one that minimizes bulk transfer, instead just identifying the 'differences' or changes from before and transferring them only. This approach is called change data capture, and represents an exciting new development in the world of information management and 'real time BI'.

What is CDC?

Change data capture is an approach to data integration, based on the identification, capture, and delivery of only the changes made to operational/transactional data systems. By processing only the changes, CDC makes the data integration, and more specifically the 'Extract' part of the ETL process more efficient. When done correctly, it also reduces the 'latency' between the time a change occurs in the source systems and the time the same change is made available to the business user in the data warehouse. This latency, although strictly speaking never zero (i.e. true real time), can typically be configured to be as near to real time as is practical, opening up new business opportunities of faster decisions, faster reaction times, and faster business execution.

Some companies have developed their own home-grown CDC solutions, but these are typically limited in scope, costly and difficult to maintain. Today, a new breed of pure-play CDC software is available that supports many different business intelligence initiatives and data integration processes and is a strategic component of the data integration infrastructure and compliments other technologies such as ETL and EAI.

The Value of CDC

The value benefits of using CDC technology can be many and varied. However, when used in conjunction with ETL tools such as Microsoft's SSIS it is normally adopted for one of the following key reasons:

1. ***Delivering data on-demand and in near real time*** – when business requirements demand much more current and/or real time access to information from within their BI systems, CDC provides the means to deliver it.

2. ***Dramatically increasing efficiencies –*** when processing and/or network resources are at a premium, using CDC to move only the data that has changed rather than periodical bulk data transfers, is a much more efficient way of addressing the growth in data volumes head-on.

3. ***Eliminating the need for batch windows -*** data is captured and processed while the underlying systems keep working. The continuous data stream of a CDC solution can virtually eliminate the "batch window" bottleneck.

4. ***A strategic solution*** - a comprehensive platform that works with other integration products and can be used for many initiatives, including BI/DW, CPM, BAM, migrations, consolidations, and more.

5. ***Cost savings -*** CDC can substantially reduce IT operational costs in terms of human resources required for a given integration project, as well as on-going costs related to system and storage requirements.

Use Cases for CDC and ETL

Strategic CDC technology should be a comprehensive solution that addresses many business requirements and can be applied to solve a variety of real time and on-demand initiatives that can substantially improve return on investment. The following are just some examples of how a CDC/ETL combination can be implemented to solve real business problems:

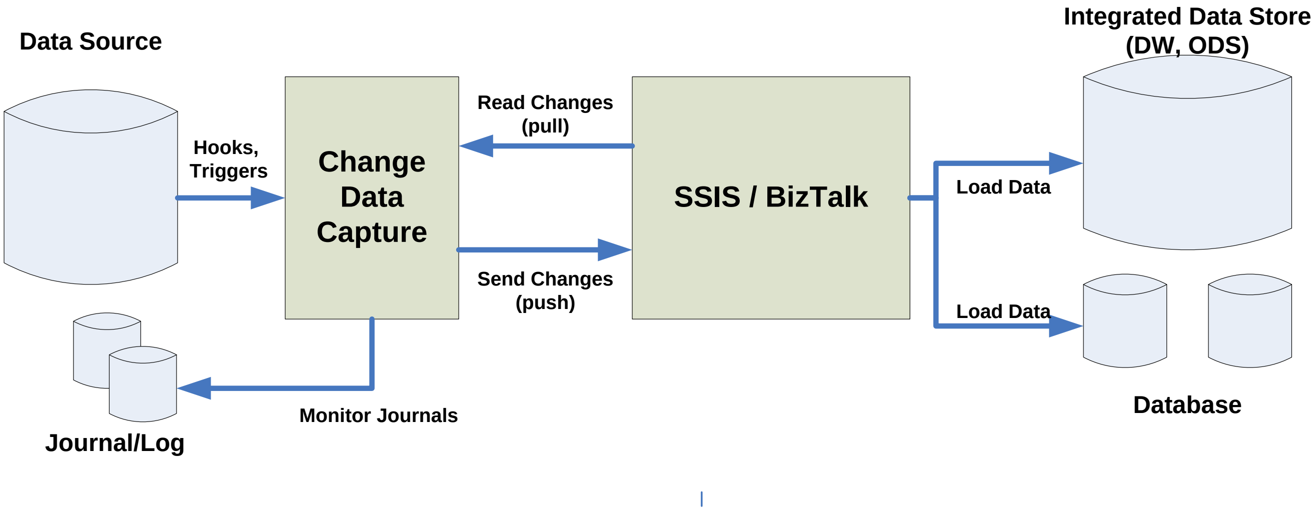
1. **Business Intelligence and Data Warehousing**
2. **Building Operational Data Stores (ODS)**
3. **Real time Dashboards**
4. **Data Propagation**
5. **Data Synchronization**
6. **Data Quality**

Business Intelligence and Data Warehousing

As previously described, a common case for using CDC is in conjunction with ETL tools such as SSIS for faster more efficient data extract in data warehouse implementations. Traditional ETL processes require that the operational system(s) be taken off line for a given period of time. This period of time is referred to as a "batch window", typically measured in hours and sometimes days, during which the system is busy with moving the data and cannot perform operational and other mission critical functions. Given the limitation of this 'bulk' approach, most IT shops update their DW only daily, and often on a weekly basis, meaning the data on which business decisions are being made can be up to a day or a week old.

Competitive pressures nowadays demand much more up-to-the-minute information, so companies now look for ways to update their DW faster, and reduce transfer latency as much as possible. CDC is one such approach, and the leading vendors of the technology work seamlessly with ETL and EAI tools like SSIS or BizTalk Server respectively.

The following diagram illustrates what a CDC-based solution used in conjunction with SSIS (or BizTalk) looks like:



CDC delivers changes to an ETL or EAI tool in batch or real time. This allows dramatic improvements in the efficiency of the entire process, reduces or totally eliminates batch windows, delivers information with low latency, and reduces the associated costs including CPU cycles, storage, network bandwidth and human resources.

Building Operational Data Stores (ODS)

An ODS is an integrated data repository addressing a specific business area (marketing, finance, support and maintenance, etc.) and providing complete and current information that can be used by business users and BI applications. The ODS is stored in a relational database and receives updates from a dedicated set of programs, ETL or EAI tools. Given the business need fulfilled by an operational data store, it requires timely updates.

CDC provides an efficient mechanism to keep an ODS up-to-date, by identifying and delivering the changes to the ETL or EAI tool on a continuous basis, rather than periodically querying the entire database for changes. In addition, CDC can push changes in near real time to support ODS applications that have very low latency requirements.

Real Time Dashboards

Dashboards provide managers with selected metrics, known as Key Performance Indicators (KPI), that measure the performance of various business operations. These metrics can represent sales trends, margin tracking, financial triggers and others that alert the business user to a business condition requiring attention. Dashboards however, are only as good as the metrics they provide and the process of updating these metrics is typically done at the data integration level. Metrics that are updated infrequently mean that the user might become aware of a business problem too late. How far would you get with a fuel gauge that is updated every 2 hours?

Implementing CDC provides a method of identifying changes to different data sources that are required in order to measure a certain KPI. Ideally the CDC solution should come with data filters that can process data changes in a given order and to re-calculate new values of the KPI close to the time of the change. This in turn would provide greater, more timely visibility to the business user resulting in much faster and more effective response to changing business conditions.

Data Propagation

Data propagation addresses the need to have one or more copies of the data from a given data source. Common examples include, making production data available for reporting purposes and accessible by various departments; distributing data from a source system to multiple data centers, and often to multiple ODS to improve response time.

With CDC, the process of propagating data can be made much more efficient and reduce the latency in making new data available.

Data Synchronization

Data synchronization is typically required in order to keep two or more systems in sync and up to date. A common reason for data synchronization is system migration (as a result of business consolidation; downsizing, etc…) where running two systems in parallel for a period of time, sometimes for several years, is required. During that period, transactions that are captured by one system need to be updated in the other as well. Another common reason for data synchronization is a merger or acquisition where systems used by one company overlap with others, and both needs to be kept up to date.

CDC captures changes as they occur and processes them using tools such as SSIS and BizTalk Server. By incorporating CDC, data synchronization can be made efficient and real time.

Data Quality

Many BI initiatives fail because of poor data quality. As a result, organizations are looking for ways to improve the quality of data available to business users. In some cases, data is cleansed as part of the process of creating a new data store (e.g. DW). In others, processes are put in place to clean up the source data.

With CDC, companies can identify and capture changes to source systems as they happen and immediately feed them into data quality processes. Such processes can then apply quality rules and recommend whether clean-up activities are required. This shortens the time for ‘bad’ data to reside in the system.

Evaluating CDC Solutions?

As the previous section demonstrates, CDC can be used for many types of initiatives, and enterprises should take a strategic view of CDC to make sure that it can address current and future requirements. This section provides an overview of key capabilities and functions required in a strategic CDC solution, and ones that you should look for when evaluating any CDC technology. These include:

1. Broad platform and data source support
2. Non-intrusive architecture for change capture
3. Change filtering
4. Both batch and near real time delivery
5. Support for non-relational data (e.g. Mainframe)
6. Resilience and guaranteed delivery
7. Recoverability
8. Performance and throughput
9. Ease of use

Broad Platform and Data Source Support

The IT landscape in medium and large organizations is typically complex. There are often many different server platforms, operating systems and database management systems. Any strategic choice of CDC software should support any and all of the underlying system combinations of the organization, whether mainframe, Unix or Windows, DB2, Oracle, Adabas or SQL Server, and whether Sun, IBM or HP. By adopting and standardizing on a single strategic CDC platform rather than point products, organizations can minimize the number of integration products they use, standardize on architecture, shorten learning curves, reduce required skill sets, reuse trained personnel and minimize total cost of ownership.

Non-intrusive Architecture for Change Capture

A critical factor in the choice of any CDC product is the impact it may have on the source systems. There is no such thing as zero impact, however the different CDC architectures on offer can impact source systems to a greater or lesser extent, from raw performance to maintenance, network demands, operations and cost.

The most invasive approach is one where changes to the underlying transactional application code are made, capturing the change ‘within transaction’ and writing that change elsewhere as part of that transaction. This can impact those systems to a huge degree, from processing requirements to development demands and all of the risks associated with changing application code. For any application other than the simplest, the costs and risks associated with this approach typically make it prohibitive, and should be avoided.

A slightly less invasive approach requires changes to be made to the schemas of the source tables, typically by adding timestamp fields that are updated by the underlying database as a change to the record occurs. The CDC program identifies changed records using the timestamp. However, this approach still modifies the source system in terms of its storage requirements and the processing required to maintain the timestamps.

A third approach uses database triggers within the source system to capture and process changes. While superficially these appear to have no impact on the applications that make the changes, they do have a significant impact on the system. This is due to the additional processing that has been pushed down into the database management system (rather than in the application code), taking away from the resources available to the operational system. Additionally triggers are new programs in their own right, albeit contained within the database system, and as such can introduce error and maintenance costs/overhead, all of which need to be assessed.

The fourth and certainly the most efficient, least resource-intensive and non-invasive approach uses system or database logs or journals. The CDC mechanism in this approach identifies the changes to the source system by monitoring the database logs/journals, which are created as part of the normal operation of the database system. The solution can therefore be managed separately and it does not need to share resources with the operational system, minimizing any impact on those critical operational source systems.

Change Filtering

A key goal of CDC is to improve efficiency by reducing the amount of data that needs to be processed to a minimum. Therefore if the business requirements are for only certain changes to be captured, then it would be wasteful to transfer all changes. The most advanced CDC solutions therefore provide filters that reduce the amount of information transferred, again minimizing resource requirements and maximizing speed and efficiency.

Standard filtering allows specific types of changes to be identified and filtered out (i.e. inserts, updates). More advanced filtering provides for field-level and field-subset-level filtering.

Batch and Near Real Time Delivery

Different applications that use CDC may have different latency requirements. A robust CDC solution provides the ability to work efficiently with different data delivery models supporting the different latency requirements of such applications.

For example, a DW that needs to be updated once a day or every four hours would typically use an ETL tool in conjunction with CDC to read many changes at once and process them in batch.

In another case, two systems need to be synchronized in near real time. The coordination of the process may be done with a message-driven EAI tool, where messages are handled individually, as soon as possible. This case calls for a real time change or business event capture and delivery to the EAI or messaging infrastructure immediately.

Support for Non-Relational Data (e.g. Mainframe)

Information residing in mainframe systems such as VSAM, IMS or Adabas, is usually mission critical. Furthermore, this data is more difficult to manage because of its non-relational structure. Companies that have these types of data sources should look for CDC solutions that are able to deal with non-relational sources effectively.

A CDC solution that can capture non-relational data should be able to deliver the information in a way that can be easily processed by other tools such as ETL or EAI. If changes are processed by an ETL tool that uses SQL, look for a solution that can normalize the non-relational data and provide a relational metadata model. If changes are processed by an EAI tool, typically in XML, look for a solution that can map the legacy data source into an XML document with a corresponding XML schema that represents the original record hierarchy. A robust CDC solution should support both.

Resilience and Guaranteed Delivery

All integration processes require a certain level of guaranteed delivery. While some can use ‘at least once’ guarantee other may require ‘once and only once’. For example, ‘at least once’ guarantee may be acceptable for loading an ODS since processing a record twice does not cause a problem. However, a process that synchronizes applications and updates sales orders must guarantee the processing of a new order ‘once and only once’. There is usually a tradeoff, since a stricter guarantee requires more resources.

A CDC solution needs to provide a robust architecture where the desired level of guarantee can be achieved.

Recoverability

Data integration takes place over a network and over a certain period of time. As machine and network outages can happen, it is important that the CDC solution is fault-resilient and able to recover. CDC solutions should provide support for recovery procedures, including automated restart as well as manual reset to a specific point in time where the system stopped processing.

Performance and Throughput

Always an important factor, a CDC solution needs to support the performance and throughput capabilities that will allow the processing of the accumulated changes in the desired timeframe. There are many factors that impact performance and throughput, including network traffic, storage for staging the changes, communication protocols, etc. When considering CDC solutions, evaluate the technologies and architectures that the vendor puts in place to improve performance and increase potential throughput.

Ease of Use

Integration in general and data integration in particular is not an easy task. It requires domain expertise, technology know-how and the ability to map between systems that were not designed to work together. By making a solution easy to use, a CDC solution can reduce required skill set and accelerate implementation timeframes. These in turn impact the total cost of ownership.

When evaluating ease of use, look for intuitive solutions that assist the user with task-oriented guides, and simplify configuration with wizards and utilities.

Evaluating a CDC Solution for use with SSIS

Data integration is always a complex task, made even more so when dealing with legacy or mainframe data sources. Using SSIS goes some way to simplifying the task; however SSIS does not do everything. We have seen for example that when used in conjunction with a strategic CDC solution many additional benefits can be gleaned. What additional criteria should therefore be evaluated when choosing a CDC solution for use with SSIS?

Certainly all of the previous criteria described should be examined in detail. However, the Microsoft environment demands additional requirements, and can be summarized into the following:

1. Tight integration with SQL Server
2. Easy to use by Microsoft developers
3. Cost-effective

Tight Integration with SQL Server

Using SSIS implies that the destination database is SQL Server, although not necessarily the source database. For simple SQL Server to SQL Server CDC, Microsoft has now introduced its own offering with the release of SQL Server 2008. However, for more complex environments and certainly those with other data sources than SQL Server, a robust CDC solution with tight integration with the entire SQL Server and SSIS environment, is preferable. Given that ETL vendors want to sell ETL, and are therefore in competition to Microsoft's SSIS offerings, their 'independence' and therefore tight working relationship with Microsoft and SSIS has to be questioned.

A specialist data integration and CDC vendor with a close working relationship with Microsoft, is therefore the preferred option of most evaluation and selection projects.

Easy-to-use by Microsoft Developers

The Microsoft Developer is used to a standard, consistent development environment. The CDC solution should integrate into that environment if it is to represent a low-cost, standard-skillset, easy-to-learn choice. Accelerators and Wizards are common in this environment and the CDC solution should provide these to make the creation of SSIS packages using CDC as simple and error-free as possible.

Cost-effective

Microsoft has dramatically lowered the cost price-point of ETL with its SSIS offering with SQL Server. Traditional ETL vendors with their entry-level license costs of tens and often hundreds of thousands of dollars now need to compete at price-levels a tenth of what they were. Any CDC solution used in conjunction with SSIS should therefore fit within the same price-expectations, and have broadly similar pricing models.

Summary

Change data capture is an innovative new software technology that is changing the data integration landscape.

With today's organizations striving for ever more immediate access to key business information in order to remain competitive, and IT departments struggling to maintain the exponentially growing data volumes being generated within their data based information systems, CDC technology can represent a very strategic solution to solving the problems.

Data integration is never simple, especially when legacy and core transactional systems are involved. However, Microsoft's SSIS products go a long way to simplifying and lowering the costs of what has traditionally been a very costly exercise. When combined, SSIS with enterprise-class CDC technologies from an appropriate vendor can represent an extremely compelling and very cost-effective solution to an enterprises data and integration challenges.

About Attunity

Attunity is a leading provider of real time event capture and data integration software, which includes Attunity Stream™, the first and most advanced CDC product on the market. As a specialist in data and application integration for nearly two decades, Attunity's customers enjoy dramatic business benefits by driving down the cost of managing their operational systems, creating flexible, service-based architectures for increased business agility, and by detecting critical actionable business events, as they happen, for faster business execution.

With thousands of successful deployments of its software in organizations worldwide, Attunity provides software directly and indirectly through a number of strategic and OEM agreements with partners such as HP, IBM, Microsoft, Oracle, and Business Objects/SAP. Headquartered in Boston, Attunity serves its customers via offices in North America, Europe, and Asia Pacific and through a network of local partners.

For more information on Attunity, and Attunity Stream (CDC), please visit us at **www.attunity.com**