Paper 303-2009

Change Data Capture and the Benefits to the Modern Enterprise Data Warehouse

David Barkaway, SAS Institute Inc., Cary, NC

ABSTRACT

Historically, organizations built a data warehouse that was updated on a yearly, monthly, or weekly basis. Today in the current business environment, decisions by frontline business users need to be based on information that is refreshed hourly, every minute, or even every second. This session investigates how the data warehouse has evolved from being a departmental reporting solution to a central repository of information that is key to active decision making for front-line users. Underpinning this change is the evolution of the data integration technology that incorporates both batch and real-time technologies, this session covers the new features in SAS® Data Integration Studio 4.2 and how it supports the real-time enterprise.

INTRODUCTION

This year is the 150th anniversary of the publication of *On the Origin of Species*, by Charles Darwin, which introduced the concept that populations evolve over time. Darwin's book contains a wealth of evidence that the diversity of life arose through a branching of evolution and common descent.

Unlike the millions of years that we have seen in the evolution of plants and animals on this planet, the computer has been evolving only since Charles Babbage's "difference engine" of 1822. It was only fifty years ago that we saw the first use of magnetic ink character recognition being introduced with computers in financial institutions, with Bank of America. If we focus specifically on the area of computing that we are interested in, data warehousing, we are now down to an evolution of almost exactly 40 years, when the early pioneers in information management identified that there was a separation between operational systems and analytical systems.

Arguably, technology is evolving at a much faster speed than biological life, possibly with the exclusion of microscopic organisms. In the forty years that the data warehouse has been evolving, we have seen it change from a basic reporting platform (where the pioneers recognized that separating data into an environment that supported reporting and analysis was more efficient than running reports on an operational system) to an "active" data warehouse that aligns Business Intelligence, "What has happened," and Operational Intelligence "What is happening," creating active intelligence that allows actionable information, enabling a company to decide "What is going to happen."

BUSINESS EVOLUTION

There have been two major influences in the evolution of the data warehouse. First, there has been a change in business velocity. Forty years ago you might have visited your bank on a weekly basis. You received a monthly current account statement, an annual investment statement, and had a chat with the branch manager when you wanted a loan. Today, you can call your bank 24 hours a day, and you can go online to transfer funds or to apply for a mortgage or loan. From an investment perspective, you can receive real-time feeds of your investment status directly to your iPhone.

All of this means that organizations have changed from working on a monthly, weekly, or daily basis to an environment where their systems are online 24*366. We have entered the era where companies have become global entities that are "live" twenty-four hours a day. In order to survive, they need to actively use the information that they capture to intelligently make adjustments to the business to be ahead of the competition. To support this requirement, they need a data warehouse that can provide the analytical foundation to provide this intelligence—one that is based on both historical and "live" operational data, hence the concept of what is now seen as an "active" data warehouse.

TECHNOLOGICAL EVOLUTION

In parallel to the requirements of business as the primary influence, there have been improvements in technology over the last forty years. Initially, scripts would have been written to extract data from the operational systems and populate the data warehouse. As more source systems come online and complexity increases, we see the introduction of Extract, Transform, and Load (ETL) technologies. First-generation tools such as ETI and Constellar Hub, which automate the generation of the integration code, make an appearance. ETL is mainly batch oriented, with complex transformation capabilities, ideally suited for daily updates to a data warehouse. As business requires more real-time information—especially from operational systems—we see the emergence of Enterprise Information Integration technologies, which are ideally suited for small, federated, real-time queries. Both of these technologies have the potential to impact the source systems when extracting data, so from a date warehousing perspective, we

are now seeing the adoption of change data capture (CDC) technologies to complement the existing data integration solutions.

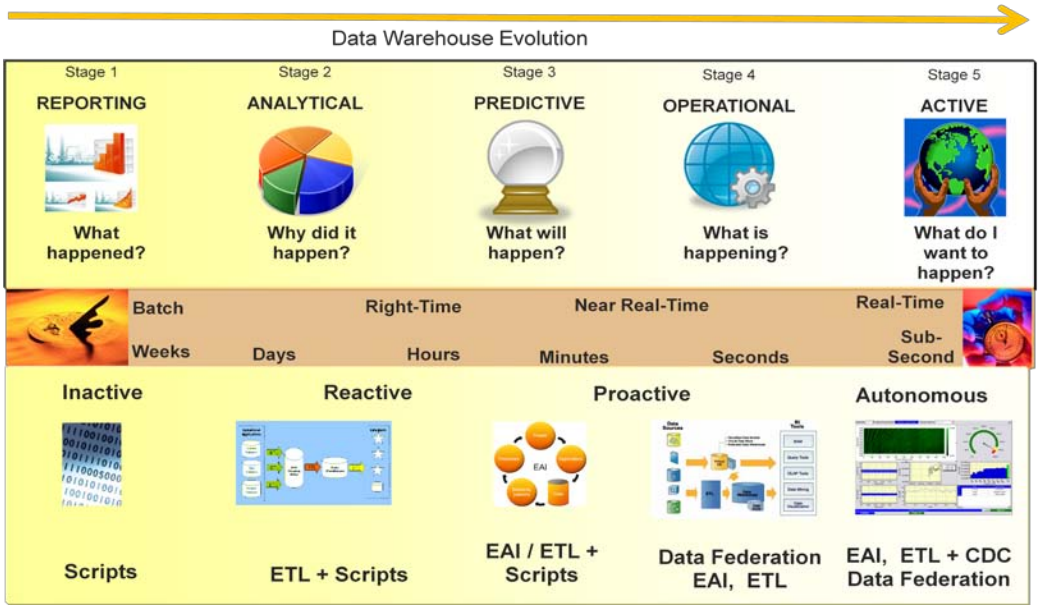

Figure 1. The Data Warehouse Evolution

Combining ETL and CDC in a data warehousing scenario provides a data integration solution to the business and addresses the technical requirements of the organization.

From a business perspective, CDC provides a real-time environment for the business to make rapid decisions and actions and to keep pace with the ever-changing environment. From a technical perspective, it provides a mechanism to capture and propagate the inevitable high volumes of data, with the major benefit of having minimal impact on the operational systems.

In SAS Data Integration Studio 4.2 we have delivered new functionality to integrate with multiple CDC vendor technologies. We have the ability to use the built-in CDC capabilities of DB2 and Oracle, and we have specific transformations to support Attunity. We also have a generic CDC transformation to support other CDC technologies such as GoldenGate.

THE DATA WAREHOUSE LANDSCAPE

If we examine how the data warehouse landscape has changed, we can see several trends:

THE MOVE TO REAL-TIME DATA

A traditional data warehouse was, and in a lot of cases still is, designed to deliver reporting, analytics, and forecasting. The data is loaded on a monthly, weekly, or daily basis, and data is delivered in batch.

If we look at the more advanced environments, companies are working hard to make their data warehouses operational and active—and thus, more critical to the business.

An important requirement is to enable lowest possible latency in which new data is delivered to the data warehouse, ideally in real time. And this new requirement for “fresher data” must often be given careful consideration when investments in batch-oriented ETL technologies already exist.

Why is fresher, real-time data so crucial? In order to determine and influence what should happen next, the enterprise data warehouse needs to know what is happening right now. There are many data integration technologies that serve the data acquisition needs of a data warehouse, and the demand for low-latency data is causing the IT department to evaluate a range of approaches: intra-day ETL batches, mini-batches, EAI, and real-time change data capture techniques. The challenge is to determine which solution or combination of solutions will meet the trend for faster data delivery needs, which will propel the move to operational and active data warehousing.

"We believe that the world of traditional warehousing is dead, really; the world of just looking in the rearview mirror," said Karen Parrish, vice-president of BI solutions with IBM, in an April 2007 interview with TDWI Radio News.[1] "The days have gone by where we looked at yesterday's sales and compared them to same sales of last year and the year prior—don't get me wrong, we still have to do that, but the world is an ever-changing set of transactions that cause us not only to have to look at what happened yesterday, but to have to determine based on what we know about what happened yesterday and the year prior, what exactly is going on today."

THE APPLE SCENARIO

To give you an example of a real-world scenario, if I go to the Apple Web site and configure a new laptop or desktop computer, rather than ordering it there and then, I store the basket by using my unique Apple ID, which identifies me as an existing Apple customer. I can guarantee that within an hour or, at a push, two, I will receive a phone call from their European call center. The sales person will confirm my order and will then offer either additional software or maybe a discount for me to place the order. This is an example of an organization capturing the data in their operational system, using my historical purchase history to then take action and to make an appealing offer for me to place the order. Apple are not calling me immediately; that would give the impression of "big brother" watching my every move on the Web site. They have decided to use a right-time approach, giving enough gap not to be intrusive, but not enough to lose the value of the information they have captured or to risk allowing me to decide to purchase a competitor's technology.

In the case of Apple, I believe they introduce an artificial delay between the message arriving at the call center and placing the call so that they don't appear intrusive. For Apple, it is a fine balance of not appearing intrusive over potentially losing the sale.

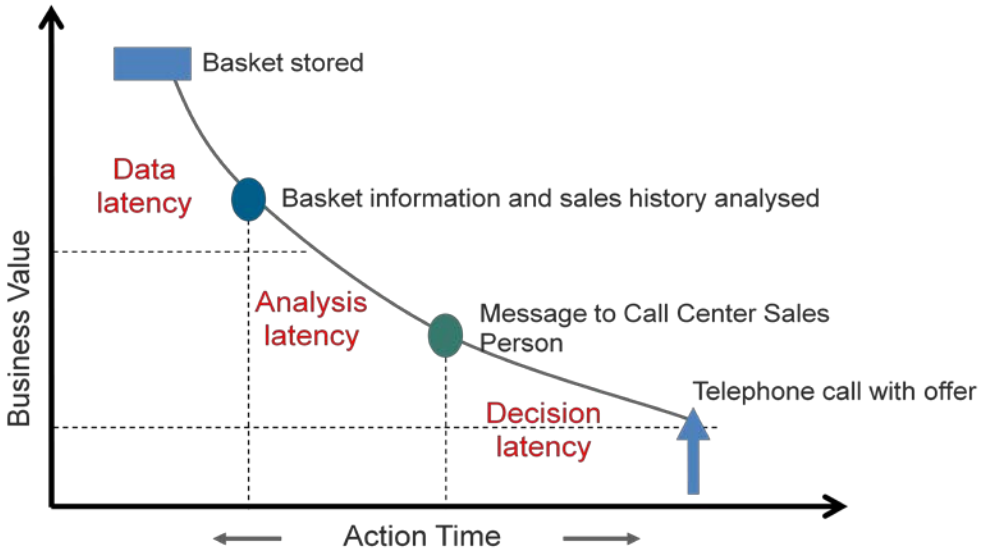

Figure 2. Business Value versus Action Time

The fundamental mechanism of the process is that the more quickly the business can capture the data, analyze it, make a decision, and then act on it, the more value it has for that business. If, for example, in the scenario, Apple had waited a day or a week to phone me, it would be highly likely that I would not be writing this document on an Apple PowerBook.

MORE DATA TO PROCESS

Data volumes are increasing. The IDC estimates that there is a compound annual growth of data of 57% per annum:

- 2006 : 161 Exabytes of data created (45 GB per person)
- Compound annual growth of 57%
- 2011 : 1800 Exabytes of new data

[1] http://www.tdwi.org/News/display.aspx?id=8368

Not only are volumes increase, but there are also increasing types of data that need to be processed, both structured and unstructured data, the majority of growth of data in an organization being in the unstructured data area, with e-mails and media. There are also increasing demands for compliance and security, where there is a requirement to store and have access to much more information and data.

With compliance and security, there is a requirement to store and have access to much more information and data. On top of all this extra data and regulation, organizations have shifted from a 5 day * 8 hours operation to more of a 24*7*365 operation, which means that they have lost the luxury of a large batch window to move data. Fundamentally, organizations have more data to process and less time to do it.

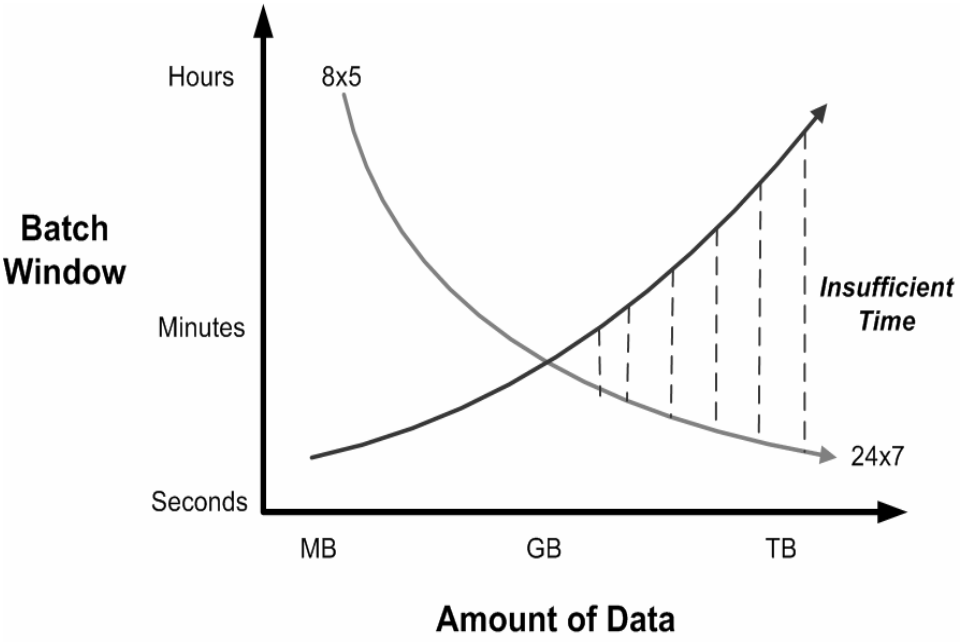

Figure 3. The Shrinking Batch Window

THE HIGH IMPACT OF REPORTING ON OPERATIONAL SYSTEMS

Before there was a general recognition that it was more efficient for there to be a separation of operational and analytical systems, reports were generated on the operational systems with an obvious impact on performance. Operational systems are designed to process transactions and are not designed for long-running analytical processes.

What we are seeing now, with the explosion of data volumes and the requirement for zero-down-time operational systems, is another architectural shift—especially in very high data volume organizations such as telecommunications or the finance sector. ETL technologies and EII technologies, which have provided the data extraction capabilities in the past, execute directly against the operational systems database, or APIs. This has the potential to significantly impact the performance of an operational system, especially if there isn't a simple way to identify changed records in the source system and a complete extract is required.

WHAT IS CHANGE DATA CAPTURE?

Wikipedia defines change data capture as "an approach to data integration that is based on the identification, capture, and delivery of the changes made to enterprise data sources." It also mentions that "CDC solutions occur most often in data-warehouse environments since capturing and preserving the state of data across time is one of the core functions of a data warehouse, but CDC can be utilized in any database or data repository system."

METHODOLOGIES FOR CHANGE DATA CAPTURE

TABLE DIFFERENCE OR TABLE COMPARISON

The most basic method for identifying changes in a database is to use table difference or table comparison. All you are doing with this technique is taking a complete copy of the source and comparing it with the target. It is really a "spot the difference" technique.

- Table difference or table comparison
  - *select * from new_version minus select * from old_version*
  - *select * from old_version minus select * from new_version* (to get deletes)
- Problems with this technique:
  - You are moving the whole table to the target.
  - Computational cost on target is high.
  - It doesn't capture intermediate values if the record reverts back to original value.

In general, I would avoid using the table comparison technique on any large data sets. If you are looking at several thousand records, this technique might be applicable. Anything above this will be extremely inefficient and should be avoided at all costs.

TIMESTAMPS, STATUS INDICATORS, AND VERSIONING ON ROWS

Tables whose changes must be captured might have a column that represents the time of last change. If the source system is more complex, you might also have a status "flag" or version number on the row. In these cases, a selection can be made to extract those rows that have changed after (or between) a certain date time parameter.

- Timestamps, status codes, and versioning
  - *select * from source_ table where date_time_stamp > '24/09/2008 10:00:00' and status_flag = 'U' and version_code > '1.1.0';*
- Problems with this technique:
  - You still have significant queries impacting the operational source system.
  - It requires timestamps and maybe other status codes to be available on the source system. Not all source systems have timestamps available for all tables that you might want to query.
  - It doesn't capture intermediate values if the record reverts back to original values, unless version numbers take this requirement into account.

DATABASE TRIGGERS

A slightly more sophisticated method for capturing changes in a database is to use database triggers. As an insert, update, or delete transaction modifies a table, a trigger is executed that maybe copies the modified row to another table where the change is logged and can be processed by an ETL tool in most scenarios.

- Database triggers
- Problems with this technique:
  - It is usually database specific, relying on capabilities being available within the database.
  - The ongoing maintenance of this method can be high, especially if you are capturing changes from a large number of tables.
  - Because this method is usually specific to the database ongoing upgrades, migrations of this solution can prove complex.

LOG READERS ON DATABASES

Most enterprise database management systems manage a transaction log that records changes that are made to the database contents. In most systems, this can be configured to be switched on and off, and the level of detail that is logged can be defined as well to tune performance. By scanning and interpreting the contents of the database transaction log, one can capture the changes made to the database in a non-intrusive manner.

- Log readers
- Potential problems with this technique:

- Cost: Unlike the other options, you are purchasing an additional component to process changed data.
- Coordinating the reading of the transaction logs and the archiving of log files. In some cases, the fact that archive logs are being used might be a benefit, because you have the option of processing the archive logs on a separate system, thus reducing the impact on the source still further.
- Dealing with changes to the format of the transaction logs between versions of the database management system.
- Eliminating uncommitted changes that the database wrote to the transaction log and later rolled back.
- Dealing with changes to the metadata of tables in the database.

I have labeled the above as *potential* problems because the majority of CDC log readers will address these issues "out of the box." CDC tools such as GoldenGate and Attunity have the option to process committed changes only. They have processes to automate the update of metadata structures, and they should automatically identify the database version so that changes in transaction log structure are addressed between releases.

From a beneficial perspective these technologies have distinct advantages over the previously discussed options, including:

- Minimal impact on the database —even more so if one uses log shipping to process the logs on a dedicated host.
- No need for programmatic changes to the applications that use the database.
- Low latency in acquiring changes. These technologies can cope with thousands of changes per second.
- Transactional integrity: Log scanning can produce a change stream that replays the original transactions in the order in which they were committed. Such change streams include changes made to all tables participating in the captured transaction.

CDC-TL

The traditional technology associated with data integration and data warehousing has been Extract, Transform, and Load technologies. These are mainly batched based, with scheduled jobs that extract the data from the source, transform it, and load it into the target, hopefully using a highly efficient bulk load method.

More recently we have seen the emergence of Enterprise Information Integration (EII) technologies, designed to create a federated query against the source system, used to complement the data warehouse with a view across operational systems or incorporating a new source system into the BI world without adding it immediately to the data warehouse. EII is subtly different from ETL in that, in most scenarios, whereas ETL is usually scheduled, EII is initiated by a query or by a Web service call against a business view or virtual database.

Some organizations have also used Enterprise Application Integration tools as the main mechanism for data integration, even to the warehouse.

In the table below, we can see the strengths and weaknesses of each technology. It is evident that ETL and CDC technologies are very complementary.

	Scripts	*ETL*	*EAI*	*Data Federation*	*CDC*
Data Volumes	Medium	Very High	Low	Low	High
Frequency	Intermittent	Intermittent	Continuous	Intermittent	Continuous
Trigger	Scheduled	Scheduled	Source or Target event	Query from Consumer	Source Event
Data Integrity	None	None	Guaranteed	None	Guaranteed
Transformations	Intermediate	Advanced	Basic	Intermediate	Basic
Processing Overhead	Intermittent High	Intermittent High	Continuous Medium	Intermittent High	Continuous Low

The benefits of linking the two technologies are:

- Both ETL and CDC can process high volumes of data. Linking the two technologies won't cause a bottleneck.
- With ETL you have the benefit of batch scheduling. With CDC you have the added benefit of introducing a real-time element. Data can be captured in real time as the operational system is updated, and it can be "trickle fed" to an intermediate storage area. It can then be post-processed in batch, when required, for the data warehouse.
- ETL doesn't provide a guaranteed level of data delivery. If you choose the right CDC technology, you have the option of ensuring zero data loss as part of the data movement process.
- CDC has the limitation, like EAI or Data Federation tools, of having limited data transformation capabilities. If you link CDC and ETL, CDC provides the Extract capability and the ETL tool provides the Transform and Load.

With the two technologies working together, you have what I would call CDC-TL. The change data capture mechanism captures the source data and passes it to the ETL tool to be transformed, cleansed, and loaded into the target.

SAS SUPPORT FOR CDC TECHNOLOGIES

SAS Data Integration Studio 4.2 provides four CDC transformations: Oracle CDC, DB2 CDC, Attunity CDC, and General CDC. The Oracle, DB2, and Attunity transformations work directly with changed data tables that are in native database format. The General CDC transformation loads change data from other vendors or from your own custom applications.

The CDC transformations are available in the Transformations tree under the heading Change Data Capture.

The CDC transformations require you to install and configure change data capture software on your relational database, and then to use that software to generate changed data tables.

SAS CDC SUPPORT FOR ORACLE ASYNCHRONOUS AUTOLOG PUBLISHING

Change data is captured from a set of redo log files managed by log transport services. Log transport services control the automated transfer of redo log files from the source database to the staging database. Using database initialization parameters, the publisher configures log transport services to copy the redo log files from the source database system to the staging database system and to automatically register the redo log files. Change sets are populated automatically as new redo log files arrive. The degree of latency depends on the frequency of redo log switches on the source database.

There is no predefined AutoLog change source. The publisher provides information about the source database to create an AutoLog change source.

Change sets for this mode of change data capture can be remote from or local to the source database. Typically, they are remote.

The diagram below shows a typical change data capture asynchronous AutoLog configuration in which, when the log switches on the source database, archiver processes archive the redo log file on the source database to the destination specified by the LOG_ARCHIVE_DEST_1 parameter and copy the redo log file to the staging database as specified by the LOG_ARCHIVE_DEST_2 parameter. (Although the image presents these parameters as LOG_ARCHIVE_DEST_1 and LOG_ARCHIVE_DEST_2, the integer value in these parameter strings can be any value between 1 and 10.)

It is important to note that the archiver processes use Oracle Net to send redo data over the network to the remote file server (RFS) process. Transmitting redo log files to a remote destination requires uninterrupted connectivity through Oracle Net.

On the staging database, the RFS process writes the redo data to the copied log files in the location specified by the value of the TEMPLATE attribute in the LOG_ARCHIVE_DEST_2 parameter (specified in the source database initialization parameter file). Then, change data capture uses Oracle Streams downstream capture to populate the change tables in the change sets within the AutoLog change source.

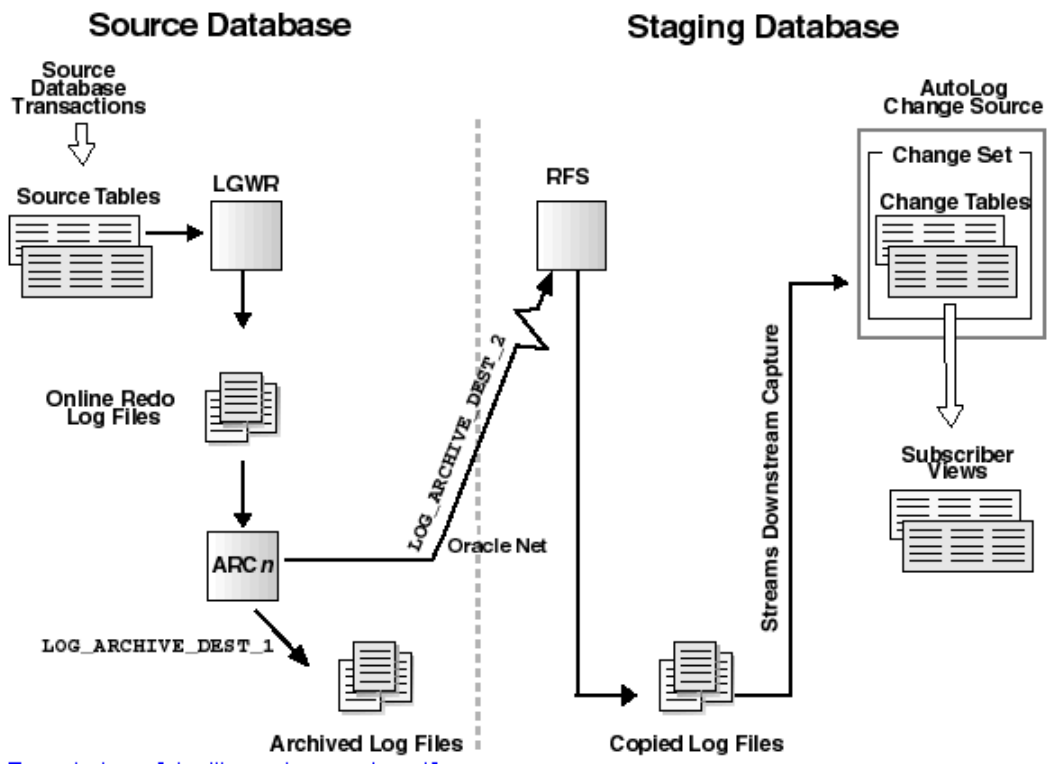

Figure 4. Asynchronous AutoLog Change Data Capture Configuration

Within SAS Data Integration Studio 4.2, once the Subscriber View has been configured, we can use the Oracle CDC transformation in a job to pick up the changes.

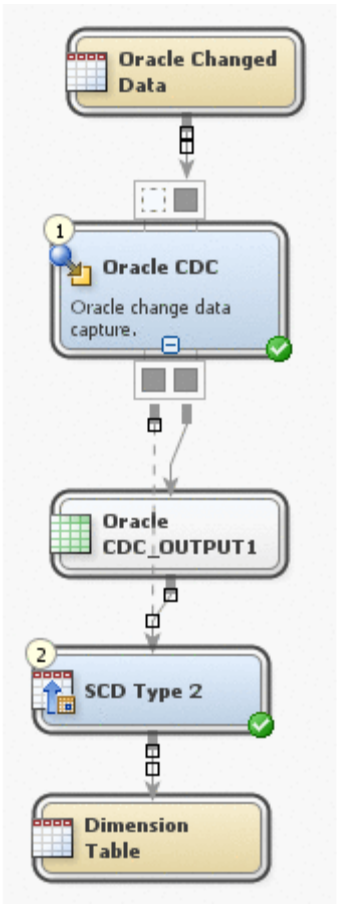

Figure 5. A Job in SAS Data Integration Studio 4.2 with the Oracle CDC Transformation

SAS CDC SUPPORT FOR ATTUNITY

Attunity is an independent vendor of CDC technology that is supported by SAS Data Integration Studio through the use of a specific Attunity CDC transformation.

Attunity provide a CDC Log Reader technology that supports multiple database types such as DB2, DB/400, SQL Server, Tandem NonStop and Enscribe, ADABAS, and many others.

The architecture of the product includes an agent that captures changes from the source database log. These are routed to a staging area, where they are stored. There is a change access service that exposes the data to the consumer. When used with SAS Data Integration Studio 4.2, the data is exposed as an ODBC source.

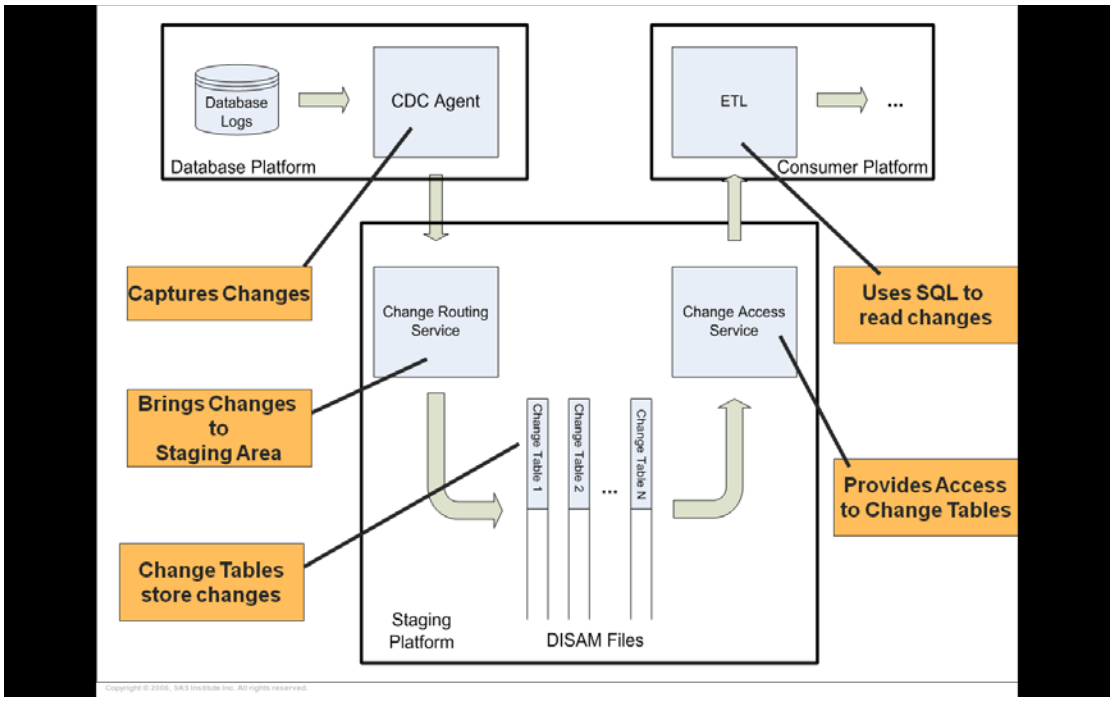

Figure 6. Attunity CDC Architecture for EDL

The Attunity CDC transformation makes it extremely easy to manage the processing of changed data records within SAS Data Integration Studio.

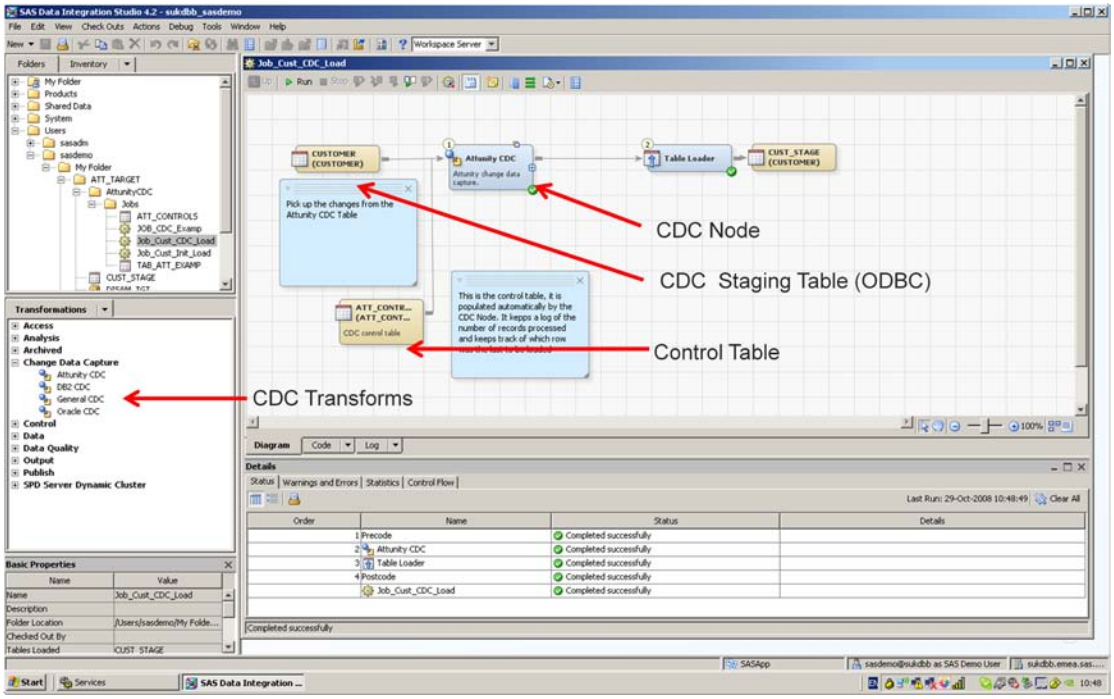

Figure 7. Attunity CDC Job with SAS Data Integration Studio 4.2

THIRD-PARTY OR GENERAL CDC TECHNOLOGIES

You can use the General CDC transformation in SAS Data Integration Studio jobs that implement change data capture. The transformation loads data from a source change table into a target. The source change table describes all of the inserts, updates, and deletes that have taken place on a relational database since a specified date and time.

Optional capabilities in the General CDC transformation enable you to use a control table to direct the loading process. You can also configure the generation of end date/time values in targets that are used to implement type 2 slowly changing dimensions. The end date/time values are used to close out target records in response to delete transactions in the source change table.

Third-party CDC technologies that have been used in conjunction with SAS Data Integration Studio include technologies such as GoldenGate. GoldenGate is an enterprise CDC solution that is used by organizations such as PayPal, Visa, and the United Kingdom National Health Service.

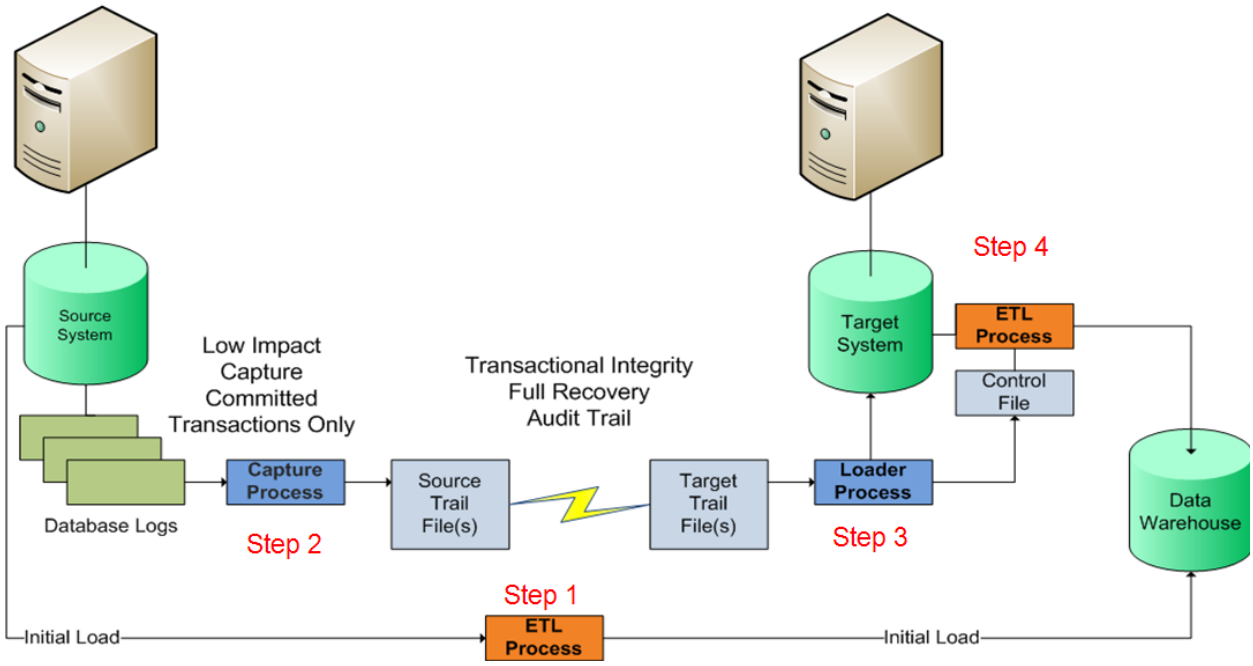

Figure 8. Typical GoldenGate CDC Architecture for ETL

BENEFITS OF CHANGE DATA CAPTURE FOR THE MODERN DATA WAREHOUSE

Below is a table that highlights the key benefits of change data capture for the data warehouse environment.

Technical Feature	Business Benefit
Log-based change data capture captures only committed changes and transfers it from the source to the target environment rather than performing queries directly against the source database.	Change data capture can improve operational efficiency and saves time and resources by eliminating redundant data transfer and saving network bandwidth. It also reduces the load on the operational systems by accessing the logs rather than the “live” operational database.
Low-latency integration provides the ability to propagate the data from source to target with sub-second data latency.	The earlier a business event can be captured, analyzed, and acted upon, the more value it has to the business.
Combining CDC and ETL provides a framework to filter, translate, transform, and cleanse data and to join data from multiple sources in real time.	Provides a platform that enables users to make better decisions faster with a solution that transforms, cleanses, and filters corporate data for enhanced data quality. Securely delivers sensitive data and makes it accessible to authorized

Technical Feature	Business Benefit
	consumers only. Leverages existing investment in ETL infrastructure.
Continuous trickle feed of data captured from the source that can either be propagated to the target data warehouse in real-time or batch mode.	Uninterrupted operational and business intelligence / analytic environments as there is no requirement for batch window outages. Removes risk of bulk load failure.
High-volume data processing. Log-based CDC technologies can process 1000s of transactions per second. Couple this with the high-volume capabilities of the ETL technology to transform and cleanse the data.	The ability to process enterprise volumes of data with low latency and low impact with the ability to cleanse and transform the data to provide trusted, accurate, and timely information to the business user.

CONCLUSION

Data warehouses have evolved from simple reporting environments with the ability to analyze last week's or last month's sales, to operational and active environments that are critical to the day-to-day activities of the business. An important requirement to enable these data warehouses is the ability to enable the lowest possible latency in which "fresh" data is delivered to the data warehouse, without impacting the existing operational systems.

Log-based change data capture technologies provide these capabilities. SAS recognizes that its customers' data warehouse environments are evolving into operational and active environments. In support of this evolution, SAS Data Integration Studio 4.2 provides direct support for DB2, Oracle, Attunity, and generic CDC capabilities.

REFERENCES

- GoldenGate Software, Inc. "Real-Time Data Integration for Data Warehousing and Operational BI." GoldenGate Software White Paper. San Francisco, CA. Available at http://www.goldengate.com/resources/GGateWhitePaper_RealTimeDW_abstract.html.
- Oracle Corporation. 2003. "Change Sources and Modes of Data Capture." In Chapter 16, "Change Data Capture," of Oracle Database Data Warehousing Guide, 10*g* Release 10.1. Redwood City, CA. Available at http://download-east.oracle.com/docs/cd/B14117_01/server.101/b10736/cdc.htm.
- Attunity Ltd. 2007. "Attunity Integration Suite White Paper." Burlington, MA. Available at http://www.attunity.com/Data/Uploads/whitepapers/Attunity_Integration_Suite_White_Paper.pdf.
- SAS Institute Inc. 2009. "Working with Change Data Capture." In *SAS Data Integration Studio 4.2: User's Guide* and in the online Help.
- "IBM InfoSphere Change Data Capture" data sheet. IBM Corporation. Armonk, NY. Available at http://www-01.ibm.com/software/data/infosphere/change-data-capture/features.html

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

David Barkaway
Enterprise Global Technology Practice, SAS Institute
Wittington House, Henley Rd, Medmenham
Marlow, Buckinghamshire, SL7 2EB
United Kingdom
Work Phone: +44 (0)1252845501
E-mail: david.barkaway@sas.com