



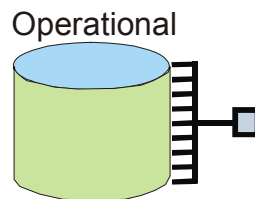
CHANGED DATA CAPTURE REDUX
By W H Inmon

In the early days of data warehousing as some of the early pioneers were working out the challenges and logistics of making data warehousing a reality, a particular problem struck these pioneers. The problem was this – data warehouses needed to be periodically refreshed. As transactions were updated into the operational environment, there was a need to periodically go and retrieve those transactions and update them into the data warehouse. The need for periodic update was a pervasive need, occurring in many corporate environments.

FINDING TRANSACTIONS

So how was data that needed to be updated into the data warehouse to be discovered? One approach was to periodically – usually daily – scan the operational data bases. This approach worked as long as the transactions that needed to be updated in fact could be located in the operational data base. But periodically scanning the operational data bases was an onerous task. The first problem was that in order to find the 1% of the data needed to go into the data warehouse refreshment file that 100% of the data in the operational environment needed to be scanned. Repeating this wasteful act of scanning 99% of a file needlessly every day was a large waste of resources. The second problem was that in many cases the operational environment was part of an online procedure and there simply weren't any spare operational machine cycles to be devoted to a full table scan.

Fig 1 shows the practice of scanning an entire operational data base looking for data that has been updated in the last day or so.



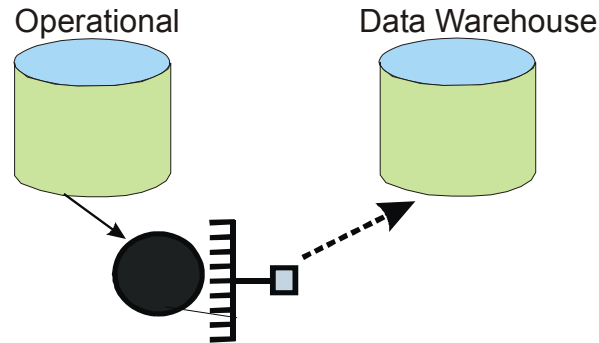
From an operational standpoint doing full table scans where 99% of the data did not need to be scanned was a practice that was unwelcome. Yet data warehouses still needed to be refreshed.

ENTER THE LOG TAPE

An alternate strategy was to use the log or journal tape that was created by transaction processing. Every OLTP environment has either a log or journal tape for the purposes of backup and recovery, should the online system go down. Nearly all of the data needed for a data warehouse refreshment was included in that tape. Even though log tapes were difficult to read and were never created for the purpose of refreshing a data warehouse, the log tapes nevertheless contained the data that was needed for refreshment. Furthermore, the log tapes were able to be processed off line. This means that by processing the log tapes there was no disturbance to the online system.

Thus born was the practice of “changed data capture” (or “CDC”).

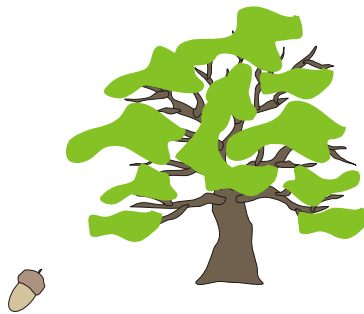
Fig 2 shows the early form of CDC.



The problem originally solved by CDC was that of the need for efficient and unobtrusive refreshment of a data warehouse. From that humble origin has grown a sophisticated and wide ranging technology that is used far beyond the boundaries of data warehousing (although CDC is still one of the mainstays of data warehousing.)

EXPANDING THE HORIZONS OF CDC

From the tiny acorn planted years ago has grown a mighty oak.



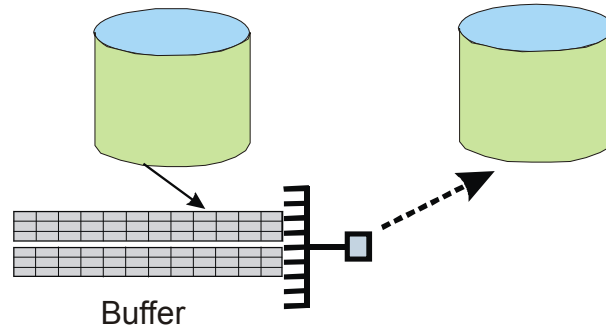
When the oak was planted years ago, who could have foreseen the many functions the oak would serve. The oak gives shade on a hot summers day. The oak provides a habitat for birds and insects. The oak provides wood for building and fireplaces. The oak provides acorns for squirrels to hide and ultimately create more oaks. In short the original pioneers of CDC had no idea of the far reaching consequences of their concept.

READING THE BUFFER

One of the first advancements made in CDC technology was to operate on a different basis than reading log tapes. Log tapes serve their purpose quite well. But even in the best of circumstances, log tapes take some amount of time to process. If there is a need for real time or near real time data, log tapes are simply too slow.

It was soon discovered that there is another source of data other than log tapes. That source is the buffer area of output for transaction processing. Upon the execution of a transaction, the output of the execution of the transaction is placed in the buffer area. It is at this moment that the results of transaction processing can be captured. The advantage of capturing the results of processing here is that the capture can be done in near real time speeds, much faster than capture can be done from a log tape. So if near real time CDC is desired, capturing the results in the output buffer area accommodates.

Fig 4 shows the capture of results in the buffer area.



ADVANTAGES AND DISADVANTAGES

While the capture of transaction processing done in the buffer area is a possibility, there remains the capture of data from the log tape. There are advantages and disadvantages to each approach –

Log tape capture

- can be done off line
- requires no operational machine resources
- requires a utility to “understand” the contents of the log tape
- requires a fair amount of time to process in the best of cases

In buffer capture

- can be done in near real time speeds
- requires some amount of operational resources
- adds a degree of complexity to online processing

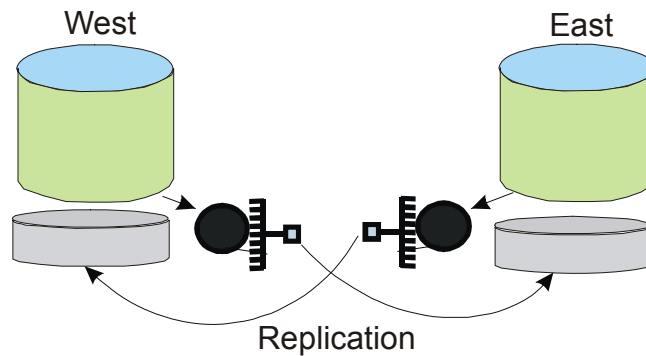
From the start as a part of a data warehouse solution, the uses of CDC have expanded over the years. Indeed CDC is still part of the data warehouse technical infrastructure. But there are legitimate uses for CDC well beyond data warehouse.

REPLICATING TRANSACTIONS

One such use is for the purpose of replicating transaction data. This is an operational need and has little or nothing to do with data warehousing.

Suppose an organization awakes one day to find that it has multiple processing centers. Suppose that the organization determines that it needs to have a replication of transaction processing spread across many different locations. In such a circumstance CDC can be used to replicate the transactions that have occurred across many different physical locations.

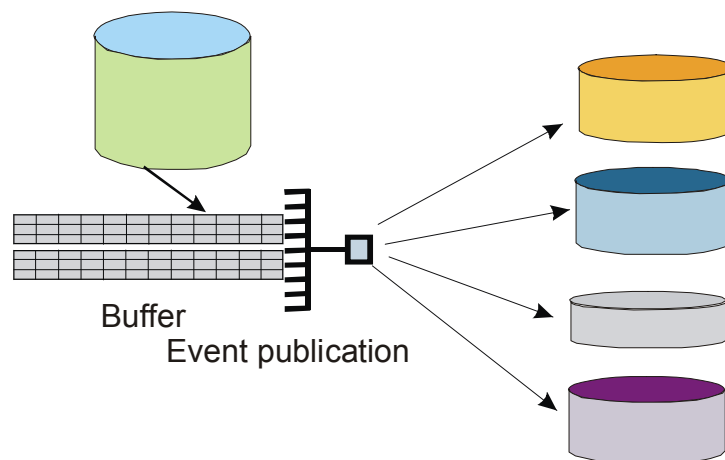
Fig 5 illustrates this use of CDC technology.



EVENT NOTIFICATION

Another operational use of CDC technology occurs when there is a need for event publication. In this case, one environment has processing done. Upon the completion of certain kinds of processing, the fact that the processing has occurred and the results of the processing are “broadcast” to other operating locations. Once those locations receive notification that an even has occurred, the processing that occurs in the outlying locations is affected.

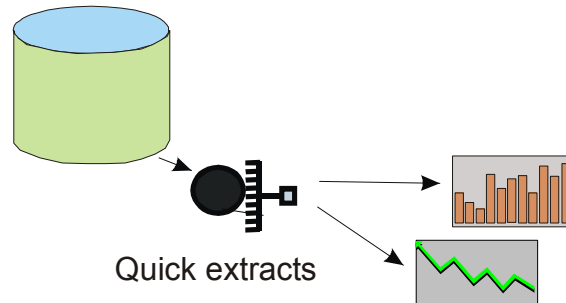
Fig 6 shows the usage of CDC as part of an event notification processing.



OCCASIONAL EXTRACTS

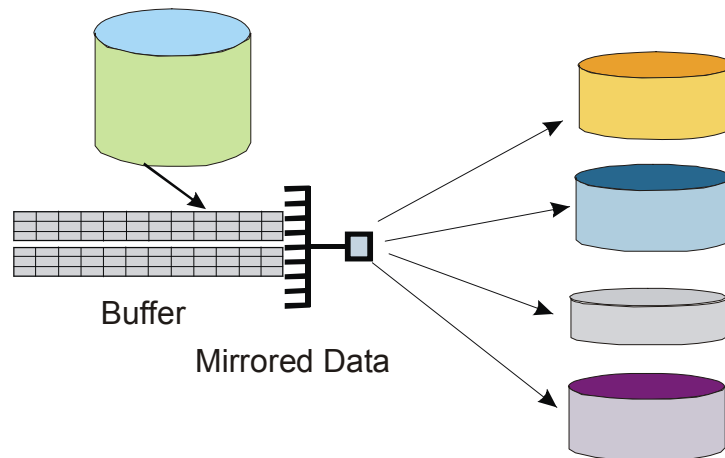
Yet another usage of CDC is for an occasional extract. Often times it is simpler and easier write a CDC script than it is to write a Cobol or vb.net program.

Fig 7 shows this use of CDC technology.



MIRRORING DATA

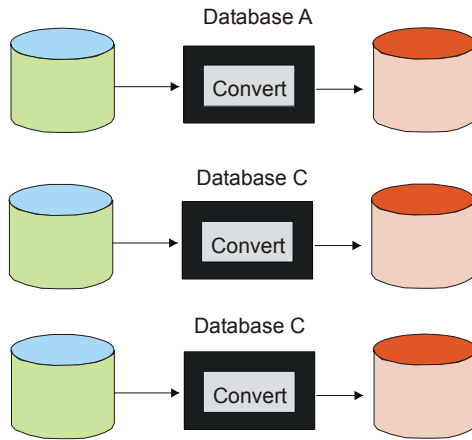
Yet another (somewhat related to event notification) type of processing is mirroring. Occasionally an organization finds itself needing to mirror processing that has occurred in one location in other locations. The mirroring needs to be done on a near real time basis and the mirroring needs to include many different activities (or even ALL activities). In this case CDC technology can be used.



LARGE CONVERSIONS

Yet another strong case for the usage of CDC outside of data warehousing is that the occurs in a massive conversion. In a massive conversion, there are often many files and data bases that need to be converted. In fact there are so many files that need to be converted that they cannot all be done at once.

Fig 9 shows this circumstance.

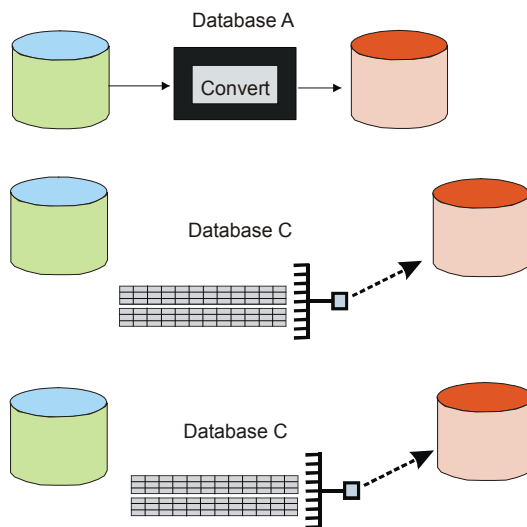


The problem with not being able to convert all data at the same time is that the system – for a period of time, is in a state of limbo. The system still needs to operate even though part of the data is converted and part of the data is not converted. This “state of limbo” poses a real challenge for organizations that struggle with this problem.

It is into this arena that CDC offers some real advantages.

CDC can be used as a “quick and dirty” to simulate the results of a conversion. While some of the data in fact has really been converted, other sources of data use CDC to simulate the data to appear as it would look if a real conversion had been done. In such a fashion, CDC allows the organization to operate in a state of “limbo”.

Fig 10 shows this state of affairs.



In the beginning were a few organizations that support CDC. As time has passed, the speed and sophistication of CDC has grown in addition to the cost going down.

Organizations such as SQData continue to pioneer the uses of CDC and the efficiency of operation.

SQData offers enterprise-class, near-real-time data integration solutions, which includes high-performance changed data capture (CDC), data replication, data synchronization, enhanced ETL and business event publishing. SQData specializes in the high-speed delivery of legacy mainframe data (IMS, DB2, VSAM, etc.) into data warehouses and downstream applications.

SQData allows you to address many different business issues with a single product, providing a comprehensive solution for changed data capture, replication, enhancing existing ETL processes, data migrations/conversions and straight ETL...all within a single package.

No longer do you have to purchase multiple products to accomplish what SQData can do for you out-of-the-box, enabling you to enjoy the consistency, performance and reliability of an industrial strength integration framework.

- **Near-Real-Time Changed Data Capture (CDC)**
SQData offers near-real-time changed data capture for both non-relational (IMS, VSAM) and relational (DB2, Oracle) datastores. SQData's high-performance data capture agents, coupled with cross-platform integration engines, provide the most versatile CDC solution available.
- **Real-Time Replication**
SQData provides a complete end-to-end replication solution for customers who have a need to synchronize their relational and/or non-relational data with near-real-time latency.
- **CDC Enhanced ETL**
SQData's CDC technology provides a solution for optimizing existing ETL processes by eliminating the need for costly bulk unloads of source data, saving significant CPU processing cycles and time. Of course, you can also use SQData for pure ETL should the need arise.

Customer Benefits

Simply stated, the SQData solution offers the best customer value available in the marketplace.

- **Significantly Lower Cost of Ownership**
SQData has proven to provide customers with considerable cost savings over other 3rd party integration products.
- **Address Multiple Business Needs**
SQData's rich capabilities allow customers to utilize the technology for many business initiatives.

- **Alternative to Existing Replication and/or ETL Products**
Replacing existing replication / ETL products with SQData results in enhanced capability at a lower cost.
- **Rapid Deployment Time**
A short learning curve and eliminating of custom programs results in faster, production-ready interfaces.

IN SUMMARY

From the humble origins as a solution for data warehouse refreshment, changed data capture has evolved into a multi functional technology serving many different needs in the organization. Now CDC can be done offline or in real time, CDC can serve operational needs, CDC can serve a diversity of organizations in the world of IT.

About SQData

SQData was founded in 2000 by a team of Database Software Developers and Consultants with a goal of providing customers with the most comprehensive data capture, replication and event publishing framework available.

SQData's customers tend to be large companies who have a need to integrate their legacy data with newer technologies. Companies choose SQData for its legacy data capability, reliable operation and simplicity.

For more information regarding SQData technology and professional services, please visit www.sqdata.com.