

**Delivering information you can trust**

April 2008

**IBM** **Information Management** software



## **Evaluating real-time data integration solutions**

---

**Contents**

---

- 2 Introduction**
- 2 The demand for real-time integration**
- 5 Need for a solution that extends ETL tools functionality**
- 6 Change Data Capture technology and real-time data integration to enhance ETL tools**
- 7 About IBM® InfoSphere™ Change Data Capture**
- 9 Benefits of IBM InfoSphere Change Data Capture**

**Introduction**

Access to timely and accurate information has become a critical driver of business success. Organizations need to have access to timely accurate information to make better business decisions, serve their customers better and respond to changes in the market. But increasingly, they are hitting roadblocks in their quest for up-to-date accurate business information.

Today, most mid to large-sized corporations have a data warehouse capable of analyzing information from production systems on a daily or weekly basis, however many are finding that extracting information from overtaxed mission-critical production databases is becoming more time consuming and complex. At the same time, the demand for more timely information is increasing to the point where daily or weekly updates are no longer sufficient.

**The demand for real-time integration**

What keeps today's Chief Information Officers (CIOs) up at night? It is likely to be a long list of challenges, and chances are that data integration ranks high amongst those. According to a Forrester survey of more than 600 technology decision makers and enterprises across North America; improving integration between applications is one of the top priorities for organizations today. Why?

Let us look at some business challenges that organizations face:

- *Increasing demand for real-time information for reporting and analytics. Traditionally, reporting was done from warehouses which were updated on a daily or weekly basis. For many types of reports, that data is current enough. For others, though, nothing short of up-to-the-minute would suffice, such as inventory data where product inventory is very high or very low, or billing information where billing is done by the minute or every fraction of a day.*
- *Large volumes of information are difficult to handle in a batch window. As more information is gathered – such as online transaction data, inventory data, and customer information – the effort involved in moving it to the warehouse increases drastically. Many organizations are finding that an eight hour batch window is no longer sufficient for traditional ETL (extract, transform, load) tools to integrate all of the needed data.*
- *Necessity to conduct business 24/7 is reducing batch windows. As more business is done across time zones and over the web, many organizations are faced with the problem of shrinking batch windows, making it more difficult for traditional ETL tools to extract data in the short time available.*
- *Growing need to detect and react to business events as they happen. Many organizations are looking for ways to detect business events in production systems and have those events trigger a response in another system. For example, a cell phone company would like to send a text message to a customer running low on minutes asking if he would like to purchase more.*

- *The need to track all changes for auditing purposes. Organizations need to comply with regulations, which often requires them to continuously track all changes to data and not just the net result of those changes.*
- *Increasing need to keep data in sync across the enterprise. Customers want up-to-the-minute access to order, payment and inventory data so they can buy products, pay bills and check delivery status online. Employees need much of the same so they can better service customers and make wise business decisions. To accomplish this, eCommerce data needs to be in sync with business applications and data needs to flow in real-time across the enterprise.*
- *Organizations want to deploy new applications using data on legacy systems without paying for an increase in workload. Often, legacy systems are already maxed out and new capacity is very expensive. Organizations want new applications on UNIX or Windows to avoid this cost, but integrating the data from those legacy systems without increasing the load on them is a key challenge.*

### **Need for a solution that extends ETL tools functionality**

The majority of organizations use ETL tools to extract data in bulk from their production systems and load it into other systems including data warehouses. The main strengths of traditional ETL tools are that they extract data from many different applications, perform complex transformations on that data, and then bulk load large volumes of data into data warehouses. But when it comes to extracting data from production systems, there are certain specific requirements of an ETL tool, such as:

- *ETL is a batch process that requires a batch window – ETL tools can move batches of data out of production systems into data warehouses. For many corporations, running a nightly batch process when systems are not running at full capacity, works well. However, as businesses are increasingly becoming global and as more and more business is conducted 24/7 over the Internet, finding a time when production systems can be taken offline to run a batch extraction of data becomes increasingly difficult.*
- *ETL batch windows increase with the amount of data extracted – Even where corporations can tolerate an overnight batch window, many are finding that with increasing amounts of data to be extracted, the batch window required to do the extraction is expanding beyond the hours that their production systems are offline.*
- *ETL tools typically require that changes be made to the production systems from which data is extracted – ETL tools often require changes to production systems – for example, adding a date/time stamp to allow incremental updates. For many corporations, the additional risk this would pose to their mission-critical systems is not something that they can tolerate.*

- *ETL tools do not track all changes that have been made to the system. Rather, they just track the net result of those changes. For example, traditional ETL tools would not track if the inventory level of a certain product decreases due to purchases but then increases due to restocking, or if the status of an order changes.*

### **Change Data Capture technology and real-time data integration to enhance ETL tools**

Change Data Capture (CDC) technology provides a way to capture changes on production systems so that they can be applied elsewhere without directly querying the database. CDC has many benefits and is complementary to traditional ETL tools.

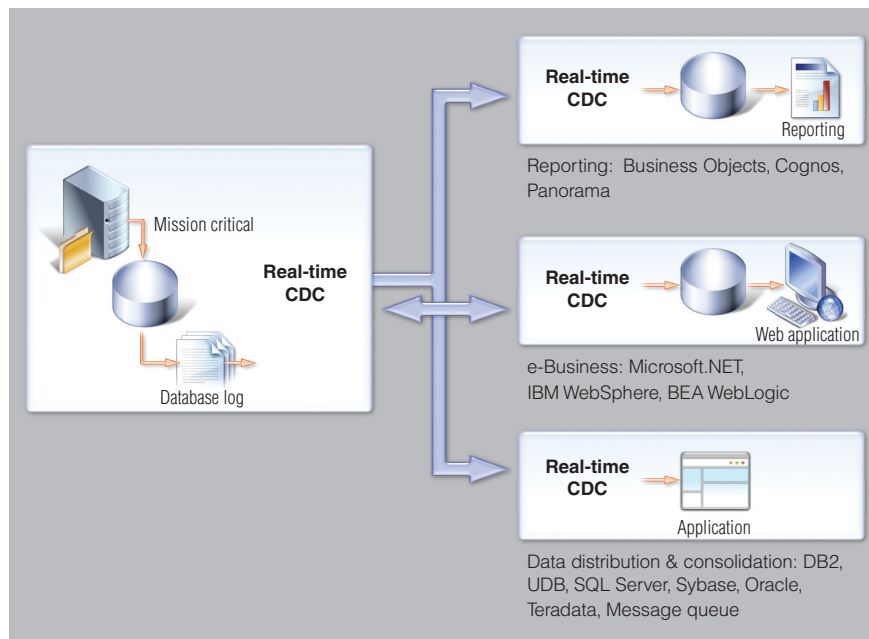
- *Changes are captured in real-time so information is always up-to-date. CDC captures changes continuously as they occur. The result: information is always up-to-date rather than being only as current as the last batch window.*
- *No impact on the performance of production systems. CDC reads database log files rather than querying the database directly. There's no impact on mission-critical production systems.*
- *No requirement for batch windows. With changes captured, transformed and applied continuously, there is no need to take the systems down to extract data.*
- *Easily scales to very large databases and large numbers of transactions. Only changes are replicated rather than all of the data in the changed tables. The result is much greater scalability through less data being moved across.*

- *Does not require changes to the source system. Because CDC is only reading the log, it does not require changes directly to the source database, yet it can detect all transactions including descriptive information about the change (user, application, time, etc).*
- *Logs all changes to the system, not just the net results of those changes. For audit and compliance, all insert, update and delete actions are recorded rather than just the net results of those actions.*
- *Complements ETL tools. Many corporations combine the strengths of CDC and ETL tools. By using CDC to flow changes in real-time to ETL tools for transformation customers get the best of each product.*

#### **About IBM InfoSphere Change Data Capture**

IBM InfoSphere Change Data Capture provides real-time, bi-directional data integration and transformation between diverse relational databases and other data stores on different platforms. Unlike batch load, query or message-based systems, IBM InfoSphere Change Data Capture was designed to deliver scalable, high performance data integration in real time with minimal latency. Its unique technology has a minimal impact on the performance of operational applications.

Figure 1: Architectural overview of IBM InfoSphere Change Data Capture





### **Benefits of IBM InfoSphere Change Data Capture**

- *Easy Integration - IBM InfoSphere Change Data Capture is heterogeneous and integrates seamlessly with existing systems, thereby helping reduce costs of replacing or upgrading existing infrastructure.*
- *Low Impact – Rather than using triggers or performing queries against the database, IBM InfoSphere Change Data Capture uses best-of-breed log-based Change Data Capture (CDC) technology to capture changed data from database logs. This ensures that the performance of even the most demanding mission-critical applications running on the source system is not adversely impacted.*
- *Easy to deploy – IBM InfoSphere Change Data Capture’s easy-to-use GUI makes it simple to select source and target databases and then configure transformations. The net result: it can help lead to faster implementation, reduce cost, and increase return on investment.*
- *Real-time – Data changes are captured in the source system as they happen and the changes flow immediately to the target systems. All information is up to date and in sync.*

- *Flexible – Customers use IBM InfoSphere Change Data Capture to solve many different business problems across a wide range of platforms – from mainframe to mobile. Customers use it to load a data warehouse, synchronize data between existing systems and web applications, distribute and consolidate data between different applications, or manage other replication-based requirements. It can be deployed in a variety of architectures, on many different platforms, and between all major databases including DB2<sup>®</sup>, UDB, Microsoft SQL Server, Oracle, Sybase and Teradata.*
- *Scalable – IBM InfoSphere Change Data Capture is a high performance solution that easily scales to large data volumes without impacting the performance of mission-critical applications running on the source system.*

**For more information**

For more information about IBM Information Server, contact your IBM marketing representative or visit [ibm.com/software/data/integration](http://ibm.com/software/data/integration)



© Copyright IBM Corporation 2008

IBM Software Group  
Route 100  
Somers, NY 10589

Printed in the United States  
April 2008  
All Rights Reserved.

IBM and the IBM logo, InfoSphere, DB2 are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries or both.

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc., in the United States, other countries or both.

Other company, product or service names may be trademarks or service marks of others.

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

***TAKE BACK CONTROL WITH*** **Information Management**