InfoSphere Data Delivery Solutions: Change Data Capture (CDC)

Davendra Paltoo
November, 2010
Introductions…

Who Am I?

- 10 years experience working with CDC technologies and databases.
- Oracle DBA skills
- Joined Datamirror Corp. in 2000 (Data integration company).
  - Technical Support and Product Development roles
- Datamirror Corp. was acquired by IBM in 2007.
- Now work in IBM’s Software Group at the Toronto Lab
- Managing the CDC Quality Assurance Team
Presentation Overview

Challenges Facing Business And IT Today.
Understanding Data Delivery Methods.
What is “Change Data Capture”? 
Why do customers need incremental data delivery?
Supported Databases and Operating Systems
CDC Architecture Overview
InfoSphere CDC ➔ DataStage Integration
What’s New in InfoSphere Change Data Capture 6.5
   - Data Distribution and Synchronization.
   - Active Data Warehousing.
   - Application Consolidation and Migration.
   - Master Data Management.
   - High Availability and Disaster Recovery.

Summary
Where to find more information
Business & IT Challenges

1. Lower TCO through low impact, data capture solutions.
2. Reduce the time and costs associated with batch windows and data movement processes.
3. Improve business insight by providing continuous access to real-time information.
4. Ability to synchronize data in silos in real-time.

Smarter businesses are able to leverage information to drive smarter business outcomes.
Ask Me How Trusted Information Can Help Your Business?

“Get A Single View of Your Business”
Mastering Information

“Get Your Arms Around Your Data”
Data Discovery & Mapping

“Understand Who’s Who & Who Knows Who”
Identity Resolution

“Integrate and Optimize Enterprise Data”
Enterprise Data Integration

“Deliver Information in Right Time”
CDC Replication

“Deliver Better Business Intelligence Faster”
Data Warehousing

“Never Lose Sight of Your Assets”
Traceability

“Understand Who’s Who & Who Knows Who”
Identity Resolution

“Integrate and Optimize Enterprise Data”
Enterprise Data Integration

“Deliver Information in Right Time”
CDC Replication
Data Delivery & Transformation require multiple models

**Virtual Data Delivery**
- Analytical & Reporting Tools
- Web Applications
- federation
- Product Performance
- Real-time Inventory Level

**Bulk Data Delivery**
- extract, transform, load
- Region 1 Product Performance
- Region 2 Product Performance

**Incremental Data Delivery**
- change data capture
- Primary Data Center
- Backup Data Center
- Database
- Extract transform load (ETL)
- Message Queue
- Business Application
Key Elements Of Incremental Data Delivery Value Proposition

**IMPACT (Lowers Risk and Cost)**
1. Lower cost of changed data access.
2. Use of native DB logs, lower overhead.
3. Non intrusive to applications and databases.
4. No use of database triggers.
5. Management easily integrated into existing IT operations.
6. Reduces risk to operational systems.

**LATENCY (Increases Business Visibility)**
1. Near zero latency for pervasive integration projects.
2. ETL can also deliver low latency but at significantly higher impact to production systems and mission-critical applications.

**CONSISTENT DATA DELIVERY (Trusted Delivery)**
1. Data pushed, delivered in continuous stream, continuous with business operations.
2. Transaction consistency maintained to preserve units of work, referential integrity.
3. Full transaction granularity, before and after image of all transactional changes.
4. Data event aware, can be used to trigger specific business processes.
5. Fault tolerance, recover to last committed transaction.
Incremental Data Delivery

- Is provided by change data capture technologies for
  - Publishing to consuming applications,
  - Delivery to one or more consumers
  - Real-time integration
- Enabled by log-based capture of database changes
- With minimal impact to source systems
- Supporting a wide variety of sources and targets.
Why do customers need incremental data delivery?

*Reduce cost, reduce risk, improve service levels*

**Live Reporting – cost optimization and agility**
Moves reporting (e.g. for Cognos) from production/OLTP server to avoid impact on the production systems.
Increases accessibility and visibility of data to lines of business for faster and smarter decisions.

**Data Distribution/Consolidation – optimize assets and reduce risk**
Multiple Point of Sales (POS) systems.
Synchronize data to improve customer satisfaction and supply chain management.
Synchronize old & new databases during a migration in lieu of high-risk rip & replace.

**Continuous/High Availability – business continuity**
Includes workload balancing and disaster recovery scenarios. Need for a back-up solution in case of a failure. High risk of doing nothing (compliance, legal, discontinued operations).

*Most enterprises use replication/CDC for two or more of these scenarios*
### Overall Expansive Source, Target, and Platform Support

<table>
<thead>
<tr>
<th>DATABASES Source &amp; Target</th>
<th>DATABASES Source Only</th>
<th>TARGETS</th>
<th>MESSAGE QUEUE</th>
<th>OPERATING SYSTEMS</th>
<th>HARDWARE PLATFORMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informix</td>
<td>IMS</td>
<td>Teradata</td>
<td>JMS</td>
<td>IBM i OS</td>
<td>IBM i OS</td>
</tr>
<tr>
<td>Oracle</td>
<td>ADABAS</td>
<td>Information Server</td>
<td>MQ Series</td>
<td>z/OS</td>
<td>IBM System z</td>
</tr>
<tr>
<td>Sybase</td>
<td>CA-IDMS</td>
<td>Cognos Now!</td>
<td>TIBCO</td>
<td>AIX</td>
<td>IBM System p</td>
</tr>
<tr>
<td>MS SQL Server</td>
<td>CA-Datacom</td>
<td>Netezza*</td>
<td>WebMethods</td>
<td>HP-UX</td>
<td>HP PA-RISC</td>
</tr>
<tr>
<td>DB2 LUW</td>
<td>IAM</td>
<td></td>
<td>BEA</td>
<td>Solaris</td>
<td>Intel</td>
</tr>
<tr>
<td>DB2 z/OS</td>
<td></td>
<td></td>
<td></td>
<td>MS Windows</td>
<td>Sun</td>
</tr>
<tr>
<td>VSAM</td>
<td></td>
<td></td>
<td></td>
<td>RedHat, SUSE Linux</td>
<td>Linux on System z</td>
</tr>
</tbody>
</table>

* Customized solution, limited requirements
Log-based CDC captures data without interacting with database
No changes or upgrades to applications and schemas required
Peer-to-peer architecture does not require additional hardware
Sending only changed data requires minimal network bandwidth
Option 1: Database Staging

1. DataStage extracts data from source database using standard ETL functions
2. InfoSphere CDC captures change made to source database
3. InfoSphere CDC writes changes to a staging table
4. DataStage reads the changes from the staging table, transforms and cleans the data as needed
5. Update target database and internal tracking with last bookmark processed

Option 2: MQ based integration

1. DataStage extracts data from source database using standard ETL functions
2. InfoSphere CDC captures/collects changes made to remote database
3. Captured changes written to MQ
4. DataStage (via MQ connector) processes message and passes data off to downstream stages
5. Updates written to target warehouse

Option 3: File Based

1. DataStage extracts data from source database using standard ETL functions
2. InfoSphere CDC captures change made to source database
3. InfoSphere CDC writes each transaction to a file
4. DataStage reads the changes from the file
5. Update target database with changes

Option 4: Direct Connect

1. DataStage extracts data from source database using standard ETL functions
2. Custom operator running on regular intervals requests data changes from InfoSphere CDC
3. InfoSphere CDC captures and collects changes made to remote database
4. Captured changes passed to user exit and writes to comm port
5. Custom operator passes data off to downstream stages
6. Update target database with changed data
What’s New in InfoSphere Change Data Capture 6.5

• Performance, Performance, Performance
  – Single scrape
  – Reduced memory footprint
  – Differential & subset refresh
  – Target pipeline

• Consumability
  – Enhanced monitoring view
  – New performance view
  – Additional shutdown options
  –Disconnected scrape

• Extended platform and database support
  – DB2 pureScale
  – Oracle 11G & compressed tables
Enabling Solutions for...

- Data distribution and synchronization.
- Data Warehousing.
- Application Consolidation and Database Migration.
- Master Data Management Systems.
- High/Continuous Availability and Disaster Recovery.
Implementation Topologies

- Uni-Directional
- Cascade
- Bi-Directional
- Multi-Thread
- Local
- 2-Way
- Remote Capture
- Consolidation
- Distribution
Data Distribution and Synchronization

- Offload query and reporting from transactional systems
  - Enables real-time operational reporting

- Synchronization for
  - Geographic distribution of data
  - Business continuity

- Publish changed data events for use in business integration
  - Message brokers
  - Web apps and portals
Challenges with Synchronizing Data to Applications/End-Users

Reduce the time and costs associated with batch windows and data movement processes. Improve business insight & agility by providing continuous access to real-time information.

- **Business Need**
  - Get information on demand for access to the most current / fresh information needed for daily decision making and to improve customer service.
    - i.e. depending on the industry, up to the minute reporting may be needed for fraud detection or risk avoidance
  - Ensure the right information is available to the right person at the right time to enable timely decision-making

- **Business Pain**
  - Decision makers and business users are working with aged data, risking errors or missing opportunities.

- **IT Need**
  - Break out of reliance on nightly batch windows.
  - Optimize the workload balance for the transactional systems avoiding performance degradation for operational transactions.
  - Reduce cost of information infrastructure – processing & hardware

- **IT Pain**
  - Minimize cost of data processing.
  - Reduce strain on operational systems to capture and synchronize data in silos.
Challenge # 1: Application Synchronization

- Solution deployed to help organizations synchronize multiple instances of applications.
- Data is in silos. Need to propagate changes in each instance to keep the data synchronized.

“Billing Plan Rate Reference Data”

“Real Time Visibility To Changes In Plan Rates Critical To Correct Billing and Customer Care”
Challenge # 1: Application Synchronization

- Solution deployed to help organizations synchronize multiple instances of applications.
- Data is in silos. Need to propagate changes in each instance to keep the data synchronized.

"Change Data Capture using native database logs accesses reference plan rate data as soon as it changes"

"Remote billing applications have real-time visibility to changing reference data"
Large US Telco reduces costs while increasing customer satisfaction

**Challenge**

- Merger and acquisition has resulted in multiple billing applications to support customers across multiple geographies and regions.
- Challenge in synchronizing changes to price plan and reference data across all billing systems in real-time results in discrepancies between nightly generated billing rates and call center quoted rates.
- Result is increased customer dissatisfaction and increased costs associated with billing reconciliation and call center staffing.
- Development and long term maintenance costs of building in-house vs buying.

**Solution**

- Real-time application synchronization.
- Chose IBM InfoSphere Change Data Capture for z/OS to synchronize changed price plan and reference data across all billing applications in real-time.

**Benefits**

- Saved estimated in-house development costs of $500K to implement basic system and longer term cost of maintenance.
- Reduced billing errors.
- Lowered re-rating and re-billing costs.
- Reduced call volumes and staffing requirements in call centers to deal with customer billing reconciliation queries.
- Improved customer satisfaction.
- Reliability of IBM InfoSphere Change Data Capture. Risk concerns about stability and reliability of an in-house developed solution.
- Longer term ROI of leveraging investment in InfoSphere CDC in other projects at Verizon.
Challenge #2: Data Synchronization For Live Reporting

Sample MIPS Utilization

Pain Points

1. Direct reporting consuming production resources (resource control)
   - Cost of processing on demand query requests during normal business hours.
   - Physical data design of production not optimum for some ad-hoc requests.

2. Cost of data movement (growing data volumes, batch processes, development costs).
   - Low availability/poor latency on target data
Challenge #2: Data Synchronization For Live Reporting

Sample MIPS Utilization

Agency Auto - DB2 Connect CPU Usage in MIPS
Analysis Period: 01/01/20X9 - 01/31/20X9

Solution

1. Reduce cost of report generation.
   - Reduce cost of processing on demand query requests during normal business hours.
   - Near real-time data synchronization

2. Reduce cost of data extraction (growing data volumes, batch processes, development costs).
   - Near real-time data synchronization
Data Warehousing

• Improves currency of warehouse data, enabling dynamic warehousing

• Compliments ETL to enable batch window optimization
  • Less time and processing from source to target
  • Reduced network traffic
  • Makes data available when full extract is prohibitive

• Provides real-time feeds into transformation, cleansing, and ETL processes

• Lowers the cost of data mart provisioning
Processing claims data once a day during nightly batch cycles.

Business Challenges (Which Drove This Project)

1. **Operational Data Store is only as current as the previous batch run (Visibility to Business).**

2. Claims applications are not available during nightly batch runs (Batch Window).

3. No high availability plan for the Operational Data Store (Business Continuity).
Health Care Claims Provider – Proposed DW Architecture

**Production Server (z/OS)**
- BlueStar
- DB2
- Claims Application
- ODS Backup

**ETL Server (p/AIX)**
- DataStage
- DB2
- “CDC” Continuous

**Benefits**
- Increased availability of key claims business applications.
- Near real-time availability of health, claims data in ODS to support multiple business consumers.
- Expected lower MIPS utilization in z environment based on native log scraping architecture.
Application Consolidation and Database Migration

- Enables zero down time during migration.
- Provides transaction-consistent data across old and new systems
- Allows flexible configuration
  • Migrate all or only a subset of data
  • Migrate between different applications, databases, operating systems, and platforms
  • Upgrade to newer versions of databases
  • No distance limitation
- Reduces risks associated with migrations

![Diagram showing Old and New systems connected through Change Data Capture](image)
InfoSphere CDC for Application Consolidation & Migration

Keep data synchronized between current production server and a server deployed to test a new application upgrade/version, or a hardware/OS upgrade. In some cases once the migration has been completed, customers may want to rollback the migration to the old server. Mirroring and Refresh can be used to synchronize the old server with any data changes that have been made to the new server.

Lower Risk/Cost Data Access

Production Server

OLTP

Manufacturing

Consolidation Server

OLTP

Manufacturing

Log Based Synchronization

Table Refresh synchronization

Roll Back Capability

ERP

Native DB Log

ERP

Roll Back Capability
Financial services company saves costs with a zero downtime migration

**Challenge**

- Migrate critical production application and data from Sun servers environment to new HP hardware.
- How to accomplish migration with little to zero production downtime. Critical business application that deals with customer credit transactions. Application must always be available.
- Very large transactional data volumes.

**Solution**

- Real-time synchronization using CDC.
- Chose IBM InfoSphere Change Data Capture to synchronize changes taking place on the old Sun Server replicating them to the HP hardware platform.

**Benefits**

- CDC migration strategy enabled little to zero downtime saving **thousands of dollars** in lost production credit transaction processing for each hour of downtime.
- CDC ensured no data loss.
- CDC reduced risk and provided security of a roll back strategy and along with a co-existence strategy.
Master Data Management Systems

- Complements ETL to enable more effective load of master data
  - Less processing time from source to target
  - Reduced network traffic
  - Makes data available when full extract is prohibitive
- Enables low impact distribution of changes in master data
  - Provides real-time feeds of master data
- Provides high availability of master data
High, Continuous Availability and Disaster Recovery

- Protects your business for planned or unplanned outages
  - Scheduled maintenance
  - Power outages
  - Allows fast switchover with transaction-consistent data.
- Flexible configuration
  - Maintain all or only a subset of data
  - Use different operating systems
  - No distance limitation
  - Allows for lower cost hardware or platform on backup system
- Maximizes ROI
  - Ability to repurpose backup system for business activities such as query and reporting
  - Can supplement other HA solutions with Active-Active capabilities
How Much Interruption can your Business Tolerate?

Ensuring Business Continuity:

1. Disaster Recovery
   - Restore business after an unplanned outage

2. High-Availability
   - Meet Service Availability objectives e.g., 99.9% availability or 8.8 hours of down-time a year

3. Continuous Availability
   - No downtime (planned or not)

Global Enterprises that operate across time-zones no longer have any ‘off-hours’ window. **Continuous Availability is required.**

What is the cost of 1 hour of downtime during core business hours?

<table>
<thead>
<tr>
<th>Industry Sector</th>
<th>Loss per Hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial</td>
<td>$8,213,470</td>
</tr>
<tr>
<td>Telecommunications</td>
<td>$4,611,604</td>
</tr>
<tr>
<td>Information Technology</td>
<td>$3,316,058</td>
</tr>
<tr>
<td>Insurance</td>
<td>$2,582,382</td>
</tr>
<tr>
<td>Pharmaceuticals</td>
<td>$2,058,710</td>
</tr>
<tr>
<td>Energy</td>
<td>$1,468,798</td>
</tr>
<tr>
<td>Transportation</td>
<td>$1,463,128</td>
</tr>
<tr>
<td>Banking</td>
<td>$1,145,129</td>
</tr>
<tr>
<td>Chemicals</td>
<td>$1,071,404</td>
</tr>
<tr>
<td>Consumer Products</td>
<td>$989,795</td>
</tr>
</tbody>
</table>

Source: Robert Frances Group 2006, “Picking up the value of PKI: Leveraging z/OS for Improving Manageability, Reliability, and Total Cost of Ownership of PKI and Digital Certificates.”
Summary

• Incremental data delivery is key to leveraging information for smarter business outcomes

• Change Data Capture solutions provide low impact, low latency, and consistent delivery of information

• IBM has the breadth and expertise to help you with your information management challenges
To find out more about CDC …

CDC specific:

CDC Forum:
https://www.ibm.com/developerworks/mydeveloperworks/groups/service/html/communityview?communityUuid=a9b542e4-7c66-4cf3-8f7b-8a37a4fdef0c

Data replication topics:
QUESTIONS
Backup Slides
Full capabilities to deliver *Trusted Information* across heterogeneous systems

- Accelerate Data Integration Projects
  - Metadata-driven design
  - Business IT alignment
- Leverage Existing Resources
  - Broad native connectivity
- Architected for Growth
  - Massive scalability
  - Parallel processing
Insurance company optimizes costs and improves business visibility

**Challenge**

- Hosted z/OS environment supporting financial billing reconciliation reporting requirements (MIPS based pricing model).
- Projected data volume growth (approximated at 10% monthly) increasing MF MIPS consumption and total report generation costs.
- Reports generated nightly. Line of business looking for on-request/ad hoc reporting capability.
- Personal Lines Division looking for a more flexible/cost effective solution for report generation.

**Solution**

- Improved data movement resulting in a more cost effective query and reporting solutions
- Chose IBM InfoSphere Change Data Capture to:
  - Optimize resource consumption and mainframe costs related to data access and report processing.
  - Synchronize changed data in real-time to reporting ODS.

**Business Benefits**

- Able to leverage technology they already owned (DB/Hardware) to build reporting Operational Data Store (ODS) infrastructure.
- Improved visibility to Line of Business by supporting on request query and reporting capability.

**Cost Benefits**

- Cost savings of approximately $100K per/month.
- Return on software investment pay back in 3 months.
- Significant optimization in MF MIPS utilization.
Challenge #3: Data Synchronization For Batch Window Reduction

Pain Points

- File locking during extraction (post batch job completion) prevented application availability to business users.
- High processor utilization during nightly data extraction.
- 30% data growth/year which was lengthening extraction run time and impacting ability to complete EDW processing for start of next business day.

Business Cost

- Development costs of hand coding and maintaining customer transformation logic/routines..

Batch Window Requirement = 15 hours

System i Production

JD Edwards One World

DB2

Nightly Data Extraction

File Save/Restore
Flat File / FTP
SQL Extracts

System i EDW

Partition A

Custom Transformation

Load

Stage

Partition B

DB2

EDW
Challenge #3: Data Synchronization For Batch Window Reduction

Solution

1. (+) Real-Time
   (-) Resources

- File locking for data extraction eliminated.
- Shell can provide continuous application availability to users even while changed data extraction is occurring.
- In-flight transformation capabilities simplified or eliminated many batch data extraction jobs.
- Compressed batch window savings allows Shell to continue to meet EDW SLA’s with yearly data volume growth (scalability).

- Adding transformation capabilities of InfoSphere CDC has reduced development costs associated with custom coding.
Global Petroleum Refiner reduces batch window and development costs

Challenge
- Provide data extraction/delivery from System I based ERP application to support multiple lines of business (EDW, E-Business) while maintaining core application availability, system performance and reducing length of nightly batch window.
- Data volume growth of 30% yearly increasing nightly data extraction run time and processor utilization which threatened Shell’s ability to complete EDW processing for the start of the next business day.
- File locking during data extraction (post batch job completion) prevented application availability to business users for extended periods of time.

Solution
- IBM Infosphere CDC for System i implemented to provide continuous, low impact access and delivery of critical ERP data to EDW and E-Business line of business users.

Business/Cost Benefits
- Elimination of file locking during data extraction enabling continuous availability of critical business applications.
- Data extraction optimization which reduced the nightly batch window from 15 hours to 6 hours.
- Development costs reduced by simplifying or eliminating several custom developed batch processes.
- A scalable data extraction architecture allowing Shell IT to maintain EDW service levels with yearly data volume growth.
- A solution that worked with IT infrastructure and that was applied to solve multiple business problems.

“The InfoSphere technology represented such clearly superior solutions that we elected to proceed with aggressive implementations and then move on to other projects”. Senior Staff Systems Architect.