Real-time replication vs ETL - How analytics requires new technologies designed for it, not '70s technology
Topics

In today’s webinar, we will discuss:

• What is ETL?
• What is CDC?
• What is replication?
• Differences
Extract, Transform, and Load (ETL)

SELECT * FROM XXX

Transform (Datatypes, structure, Embedded types)

INSERT INTO YYY
ETL Latency

Large Blocks

Every Hour, or Day, or Week
ETL Incremental

Track primary key, load PKey from Last

Track existing timestamp, extract from last TS

Modify structure to add timestamp or change tag
ETL Parameters

Advantages
• Loads large blocks of data efficiently
• Allows for complex and detailed transformation
• Allows for full data extraction
• Can be easy to implement
• Efficient data loading for some target environments

Disadvantages
• Can be time consuming
• Places extreme load on source database
• Transformation can take excessive time
• Bulk load can create data ingestion performance problems on the target database
• Incremental loading difficult or complex
  – Either timestamp based
  – Data/structure changes
Change Data Capture (CDC)

INSERT, UPDATE, DELETE

Trigger or Background Process

Modified Rows

Copy to Target
CDC Latency

CDC Generation → Row Extraction → Load
**CDC Parameters**

**Advantages**
- Easy method of identifying or extracting data
- Requires no schema modifications to the source database
- Medium latency
- Medium data loading cadence

**Disadvantages**
- Requires database level support
  - Either custom CDC or Triggers
- Requires additional database and table space
- Implies overhead on database (both for generation and extraction)
- Extraction latency can increase as data volumes increase
- Requires source access to DB
- No or few transformations
Replication

Master Replicator: Extractor

DBMS Logs

Slave Replicator: Applier
Replication Latency

Transaction

Batch
# Replication Parameters

## Advantages
- Low latency replication
- No source database access
- No source database load
- No source database changes

## Disadvantages
- Limited transformation and combination of data
- Per-transaction loads slower
  - Mitigated by batching or parallel apply
**Comparison**

<table>
<thead>
<tr>
<th>Feature</th>
<th>ETL</th>
<th>CDC</th>
<th>Replication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Rate</td>
<td>Low, periodic</td>
<td>Periodic</td>
<td>High, real-time</td>
</tr>
<tr>
<td>Transformations</td>
<td>All data and modifications possible before load</td>
<td>Very low complexity only</td>
<td>Low complexity only</td>
</tr>
<tr>
<td>Combination</td>
<td>Complex combinations possible</td>
<td>Very limited</td>
<td>Very limited</td>
</tr>
<tr>
<td>Source DB Load Impact</td>
<td>High to Very High</td>
<td>Medium</td>
<td>Low</td>
</tr>
<tr>
<td>Incremental Support</td>
<td>Difficult, or requires DDL changes</td>
<td>Fully supported</td>
<td>Fully supported</td>
</tr>
</tbody>
</table>
Example of modern Heterogeneous Deployments

- Replicator
- Warehouse
- Per User
- Per Database
- Audit

continuent
Conclusions

• ETL
  – Efficient and simple system
  – Slow, and not usable for modern deployments
  – Incremental Complex

• CDC
  – Capable with database load
  – Lower latency
  – DB dependent

• Replication
  – Very low latency/High performance
  – Limited transformations and combinations
Next Steps

• If you are interested in knowing more about Tungsten Replicator and would like to try it out for yourself, please contact our sales team who will be able to take you through the details and setup a POC – sales@continuent.com

• Read the documentation at http://docs.continuent.com/tungsten-replicator-5.2/index.html

• Subscribe to our Tungsten University YouTube channel! http://tinyurl.com/TungstenUni
For more information, contact us:

Eero Teerikorpi  
Founder, CEO  
eero.teerikorpi@continuent.com  
+1 (408) 431-3305

Eric Stone  
COO  
eric.stone@continuent.com

MC Brown  
VP Products  
mc.brown@continuent.com

Chris Parker  
Director, Professional Services EMEA & APAC  
chris.parker@continuent.com