
5214
infomat
Inspiring Business Software

Proof of Concept

OpenEdge Pro2
OpenEdge Change Data Capture

Dominique Demeyer
Development Manager @ Infomat

infomat

Agenda

- Who, what and why?
- ERP data replication for BI
- How does Pro2 fits in this story?

infomat

Who?

- Software house with in house developed ERP software for SME which is our core business
- a lot of 'connected' solutions:
  CRM, DMS, BI, Accounting,...
- 3 sites, Antwerp, Liege (Belgium)
  and Arnhem (The Netherlands)
- Founded in 1997
- + 150 Customers

infomat

A constellation of applications...

infomat

What we will bring today

…a case story of 1 particular integration trail…

BI Application

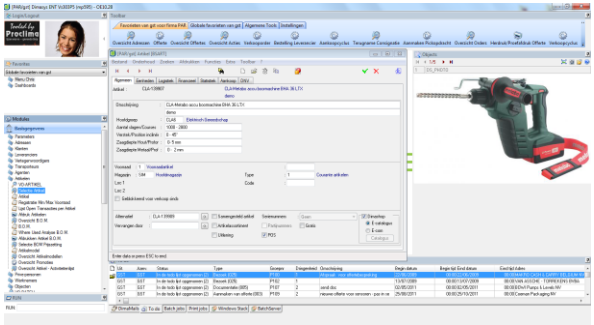
- How is the BI application 'connected' as 1 of the satellites of our DIMASYS universe
- in other words : How ERP data is replicated to the BI application

infomat

Customer's point of view

dimasys | ENT ENTERPRISE

ERP Application



i8

BI Application

infomat

This data replication is done by an ETL process

- How is the ETL structured:
  - Originally (back in 2010)
  - Now, where we started an ETL refactoring project where OpenEdge Pro2 is involved
- And what where the underlying decision triggers

What and why?

- It's not an in-dept technical session on Pro2
- An insight in:
  - The details of the integration project
  - The decision key-points
- Which explain why Pro2 became an option.
  - It seems interesting to bring this angle of a development project.

infomat

Connecting a BI application to an ERP system

2 different worlds…

- Usage level:
  - ERP = operational system = transaction processing (OLTP)
    Day-to-day operations like sales, purchases, stock, manifacturing,…
  - BI = managing tool = analytical processing (OLAP)
    Data analysing needs = base for decisions on enterprise level
- Technical level:
  - ERP typically runs on relational DBMS => OpenEdge
  - BI is build on a data warehouse DB => Microsoft SQL Server Analysis Services (SSAS)

infomat

Connecting a BI application to an ERP system

In general there are more reasons to have # datasources with # data structures :

Higher performance for each system

- DBMS : tuned for OLTP (access, indexing, concurrency control, recovery)
  => Data design = application oriented
- DWH : tuned for OLAP (complex queries, multi-dimensional views, consolidated)
  => Data design = subject oriented

infomat

Connecting a BI application to an ERP system

Conclusion: an ETL process is definitely required.

ETL = glue between ERP and BI

- **Extract** data out of a 2-dim ERP database
- **Transform** application oriented structure towards subject oriented structure
- **Load** transformed data into a multi-dim DWH database

infomat

Connecting a BI application to an ERP system

How to implement ETL for our case

- What data to transfer:
  - Which tables?
    - only a subset of tables or the complete ERP DB ?
    - an 'educated' selection of historical and master data. In our case +- 80 out of 300 tables.
  - What data per table?
    - Complete or partially (from start point in time) ?
    - Potentially a ERP system consist a lot of data
      - Historical data (detail of transactions, # years of data,...)
      - Sometimes also master data

Partially : a setting to choose a starting point is foreseen

infomat

Connecting a BI application to an ERP system

How to implement ETL for our case

- Transfer type: full or incremental?
  - Full = complete dump & load
  - Incremental = only changed data
    - Is obviously the smartest thing to do
    - Back in 2010 this was no real option
      - could only be done by triggers
      - unwanted negative impact on performance of ERP

Full was unfortunately the only option at that time

infomat

Connecting a BI application to an ERP system

How to implement ETL for our case

- When to transfer/on what frequency?
  - Real-time or delayed?
    - Real-time looks of course preferable
    - however : For analyzing purposes, a stable dataset during analyzing time may be preferable

So a daily (or rather nightly) refresh of the DWH data could be a good compromise

infomat

Connecting a BI application to an ERP system

How to implement ETL for our case

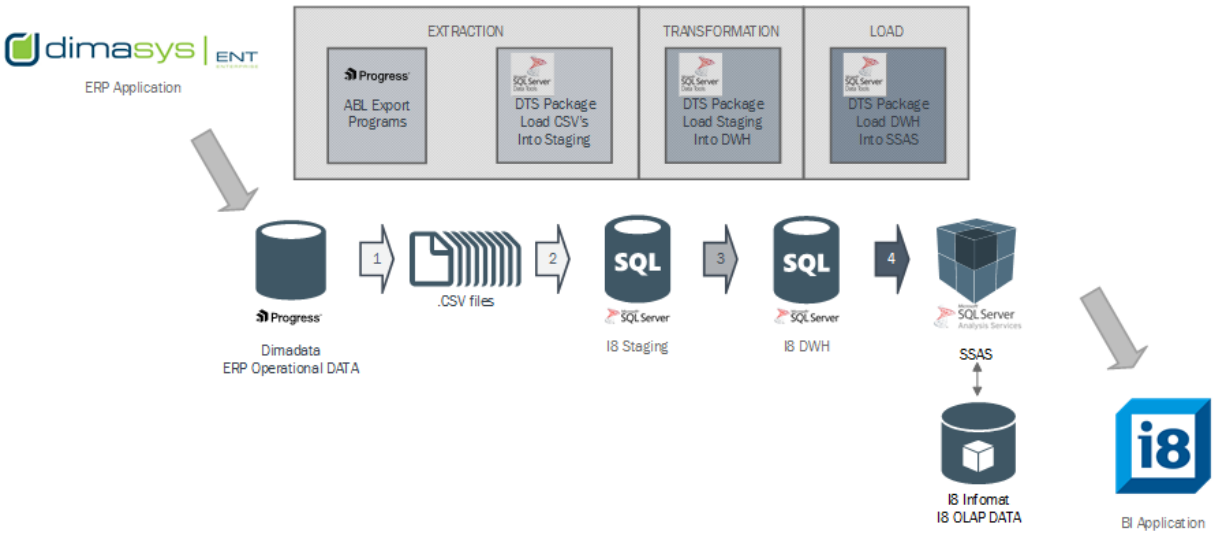
- How to connect the 2 datasources?
  - a direct connection Between OpenEdge DB and SQL Server
    - SQL script using ODBC on OpenEdge
    - Progress .R using ODBC on SQL

    Both where found not suitable for bulk processing a large dataset within the given timeframe limits

  - Export / import
    - Export to text file (csv) and bulk import to SQL
    - 'ugly' but still best option at the time

infomat

ETL structure in original setting



infomat

ETL structure in original setting

- Data extraction is done with filter on a starting point in time
- This is loaded and forwarded in the different intermediate 'stations'
  - Each time by performing a complete truncate first
- Sequence of steps
  - Step 1 is run on the Dimasys batch system (ABL based)
  - Steps 2 – 4 are run by a SQL Agent job

infomat

ETL structure in original setting

Conclusion

- … This works …
- successfully implemented at +- 30 customers
- The customers need to have an analyzing tool, running on ERP data is met !
- The aspect of having data with a delay of TODAY – 1 is generally accepted

infomat

ETL structure in original setting

Although…

- At some of the implementations, the elapse time is getting tight
  - More than 12 hours to run the ETL
    - need to limit the number of years of historical data to transfer
    - is not always what customers wants

- Some problems with 'disturbing' characters like CHR10+CHR13 in CSV
  - Need to implement extra sanitation processes

infomat

ETL structure in original setting

- The only way to solve this is :

Process less data

- If only we had a tool to ETL only changed data…

infomat

ETL structure
… A New beginning …

- No we have:

OpenEdge 11.7 with CDC

OpenEdge Pro2

infomat

ETL structure
… A New beginning …

OpenEdge CDC / OpenEdge Pro2

- What is what, and what is the difference between them…
- …and what should we use for our BI integration project ?

infomat

ETL structure
… A New beginning …

What is CDC in general?

- General industry term
- Capability to determine and track data source changes …in order to use this info in another process
- A typical application is to replicate data to a second data source incrementally

infomat

ETL structure
… A New beginning …

What is CDC in general?

- 4 classical ways to implement CDC
  - Date_created / date_modified
    - should be on DB level, not on application level, no deletes
  - Diff
    - compare : requires current + previous data state
    - a lot of resources required on storage and to compute differences
  - Triggers
    - slows down operational system
  - Log-based

Or … 'Database tied' CDC…

MS did it already for **free**, so Progress had to make a move to…

infomat

ETL structure
… A New beginning …

What is OpenEdge Change Data Capture…
dixit Progress :

CDC :

"provides a flexible and scalable capture process to facilitate the data extraction, transformation, and eventually the loading of the data to an external data source."

infomat

ETL structure
… A New beginning …

What is **OpenEdge Change Data Capture… dixit Progress :**

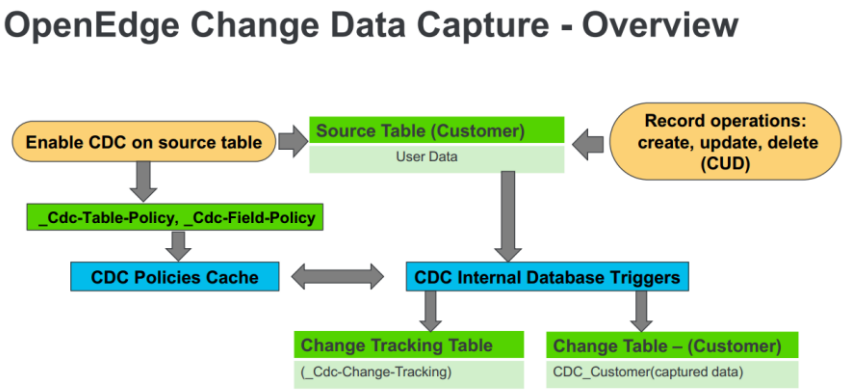
- OpenEdge CDC is flexible because:
  - ❑ Captured data is maintained in the same database
  - ❑ Captured data is maintained in relational form

- OpenEdge CDC is scalable because you can define CDC policies such that:
  - ❑ The amount of data captured is variable by table. You can capture no data, some data, or the whole record.
  - ❑ You can index the data for easier retrieval.
  - ❑ The amount of data captured is controlled through policies defined at the table and field level.

- Benefits of OpenEdge Change Data Capture include:
  - ❑ It identifies and tracks all changes made within the OpenEdge RDBMS
  - ❑ It guarantees accurate tracking of all data changes regardless of where they occur
  - ❑ It **increases efficiencies** and **availability** of changes for **ETL** to sync identified changes with other data sources, datarepositories or **data warehouses**
  - ❑ It can be activated with **zero changes to the application**, just configure and run
  - ❑ It can be managed completely online—no downtime required

infomat

ETL structure
… A New beginning …



infomat

ETL structure
… A New beginning …

What is OpenEdge Change Data Capture…

Bottom line :

CDC makes a registration of content changes on DB level

- But does only that…
- Exploitation of those changes is not done by the feature itself
- if you want to fill another data source based on CDC, f.i. the Staging SQL DB in our case, we have to establish connection to SQL instance and have to code data replication logic.

infomat

ETL structure
… A New beginning …

What is OpenEdge Pro2… dixit Progress :

"Progress® OpenEdge® Pro2™ data replication

- is a solution that provides easy, fast replication from OpenEdge into a separate OpenEdge, SQL Server or Oracle database.
- Gain quick and easy access to mission-critical data from your OpenEdge system, without disrupting normal business operations or risking transactional database stability."

Conclusion : this does exactly what we have to do ourselves, in case of only using CDC.

infomat

ETL structure
… A New beginning …

What is OpenEdge Pro2

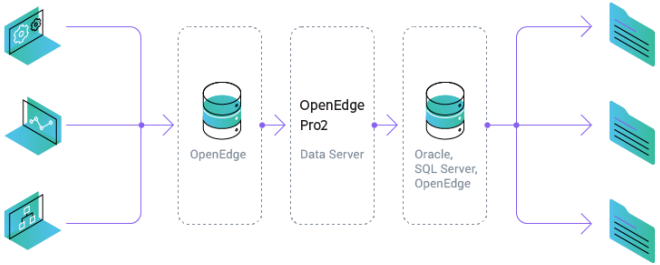
In fact the latest version op Pro2 uses internally the CDC feature of OpenEdge.

- previous versions of PRo2 uses replication triggers for that…
- Both methods are still possible, but using underlying CDC is strongly recommended
- As well for maintenance as performance reasons…

infomat

ETL structure
… A New beginning …

What is OpenEdge Pro2



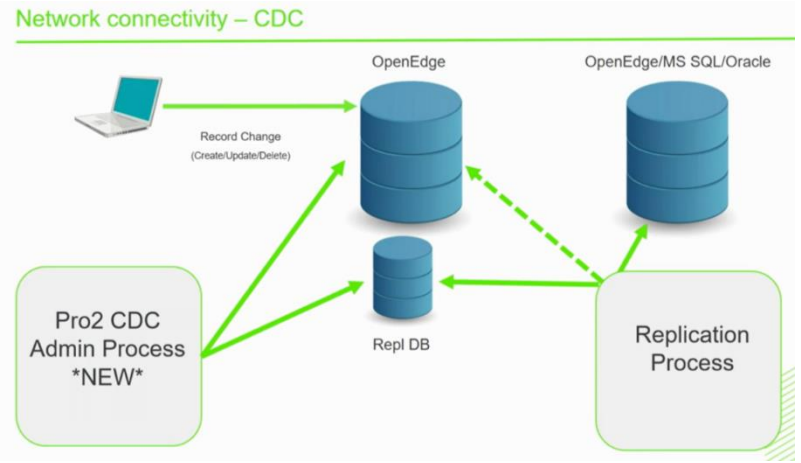
1. change data capture automatically captures all data changes
2. next, the multi-threaded replication process retrieves the updated record
3. the queued data in the replication is moved to the target database.

Done in (near) real-time

infomat

ETL structure
… A New beginning …

What is OpenEdge Pro2



infomat

ETL structure
… A New beginning …

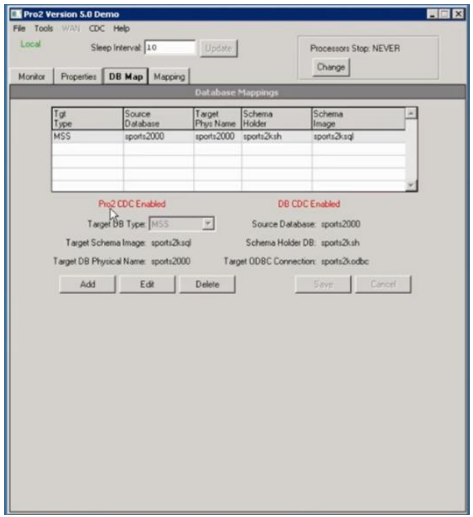
The Solution Is Pro2

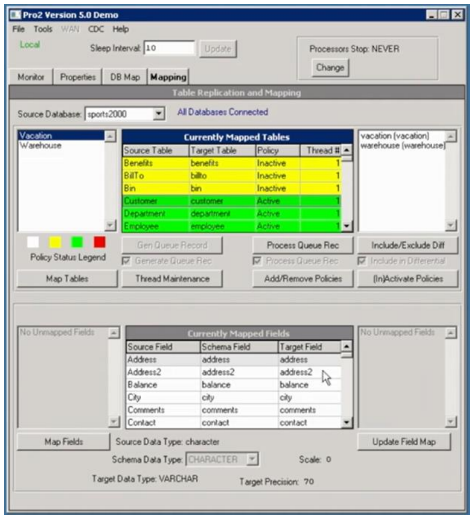
- Real-time data replication
- No connectivity limitations
- No disruption to normal business operations or risk to your system of record.
- Create a channel for transferring Progress OpenEdge data into a target databases
- Use your third party reporting solutions and tools

infomat

ETL structure
… A New beginning …

Implementation tool : The Pro2 Admin console

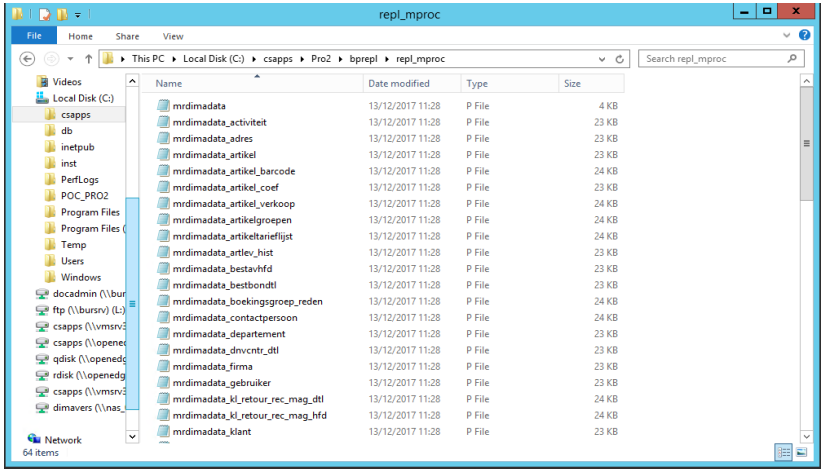




infomat

ETL structure
… A New beginning …

Possibility to 'tweak' the replication process :



infomat

ETL structure
… A New beginning …

Back to (our) business :

- It seems logic to take advantage of the Pro2 functionality right away:
  - We get a replication to a SQL Staging DB in an 'out of the box' kind of way…
- Alternatively : if we use only CDC we have to develop more of the extraction logic ourselves

infomat

ETL structure
… A New beginning …

Back to (our) business :

- So we decided to do a PoC on Pro2.
  - Set up 2 Virtual Servers
    - 1 with Progress 11.7 and Pro2 5.0
    - A copy of a medium size customer DB
    - Dimasys ERP

    - 1 with MS SQL
    - 18 BI application
- For setup/implementation of Pro2 itself we got help from a Progress Pro2 specialist Vladimir Zalda

infomat

ETL structure
… A New beginning …

Conclusion:

The POC involved evaluation on :

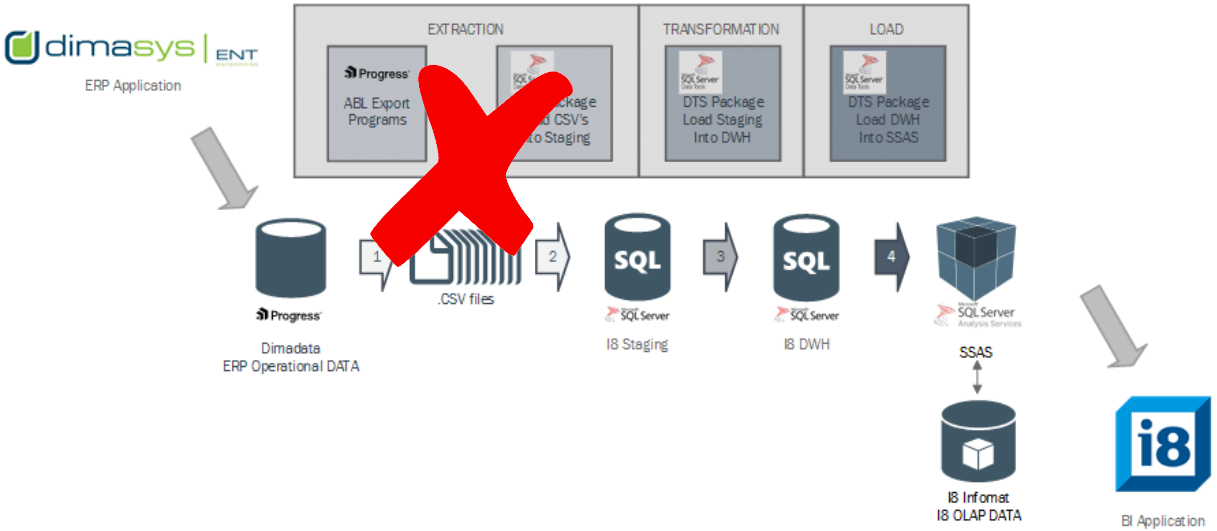
- Setup/implementation : time and complexity
- Exploitation :
  - Activate/de-activate CDC policies
  - Activate/de-activate Pro2 queue processing
  - Bulk load/reload
  - Performance on create/update flows
  - Create/update records via ERP => see result in Staging DB

- Impact on OLTP with CDC/Pro2 activated
  - User-experience testing
  - Still to do : Technical bench marking

General conclusion on Pro2 PoC

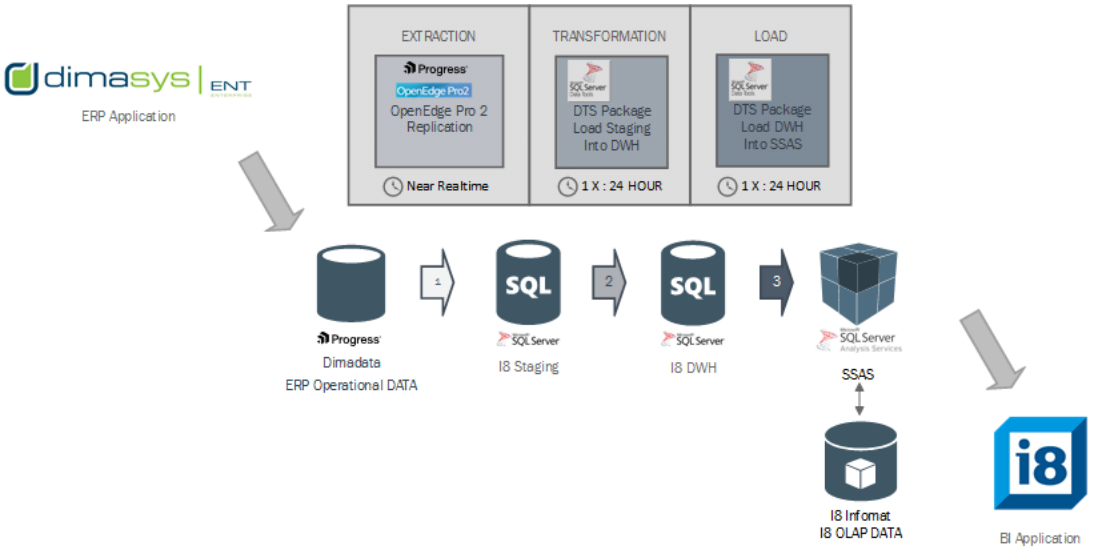
infomat

ETL structure
in original setting ...reminder



infomat

ETL structure
… A New way …



infomat

ETL structure
… A New way …

- Still more improvement to the ETL can be done :
- CDC 'intelligence' is used via Pro2 to synchronize to the SQL Staging
- We like to use change data info again in the next step.

infomat

Questions ?

infomat