



**Replacing DataStage**

**with**



**CloverETL**

*by Jiri Vojtek, Michal Tomcanyi  
April 2009*

## Objective

Proof of concept – replacing data transformation applications developed in WebSphere® DataStage® ETL<sup>1</sup> with solution based on CloverETL<sup>2</sup> suite of tools.

## Overview

OpenSys company was requested by a customer to carry out a Proof of Concept (PoC) project testing possibility to replace IBM WebSphere® DataStage® ETL platform with a more cost efficient product. Reduction of run and maintenance expenses was the driving motive behind this PoC. Not only license cost, but also HW requirements, support fees and cost of development resources were given as the key elements affecting the total cost of ownership.

## Project details

The PoC implements transformation of source system data (SRC I) carrying customer information (address and other contact details) into NEWACC standardized data file which is then processed by IBM WebSphere® QualityStage<sup>3</sup> tool where data cleansing (customer grouping) algorithm is implemented. The purpose of this transformation is to unify customer base of global company where customers are registered/entered into several independent systems in more than dozen of countries of the world.

### ***SRC I – structure of input data***

Raw data input of SOURCE I system consists of 6 pipe–delimited text files:

- *account* – account level information
- *account\_addr* – addresses for account level
- *contract* – contract level information
- *contract\_addr* – addresses for contract level
- *salesperson* – information about salesmen
- *tariff* – tariffs

## Business logic

The following specifications were used for designing & developing data transformations:

- SOURCE I File Specification – metadata for SOURCE I files (data types)
- Business Rules – rules for input filtering
- Data Mapping Template – mapping of input columns to output columns
- Channel lookups – sales channel lookups

---

<sup>1</sup>IBM WebSphere® DataStage® is an ETL tool and part of the IBM WebSphere Information Integration suite and the IBM Information Server. It uses a graphical notation to construct data integration solutions.

<sup>2</sup> CloverETL is a Commercial Open Source Java based ETL tool with low cost, excellent scalability and extendibility. CloverETL is accompanied by CloverETLDesigner, graphical designer of data transformations.

<sup>3</sup> IBM WebSphere® QualityStage® is a data profiling and cleansing tool.

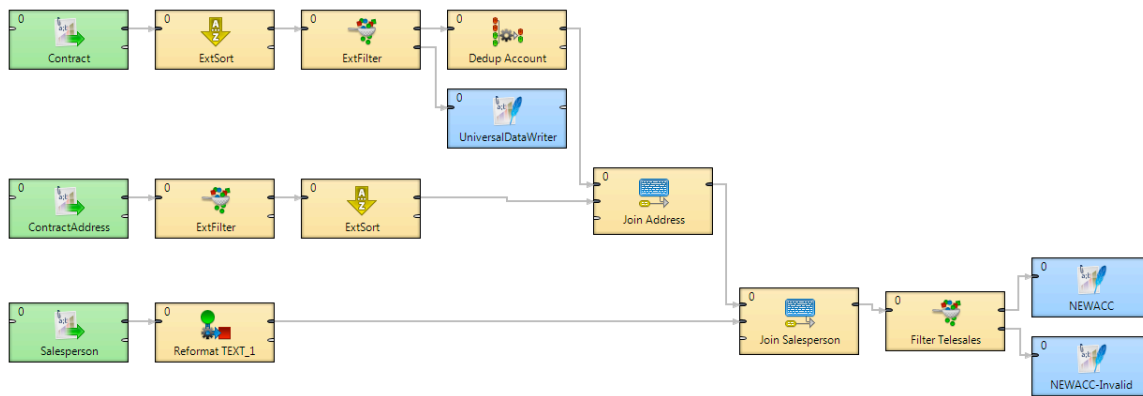
## Transformation Overview

CloverETL transformations below display account-level and contract-level processing. The result of running the graphs is the NEWACC file ready for grouping (further processing by QualityStage). Both graphs perform in general the following sequence of steps:

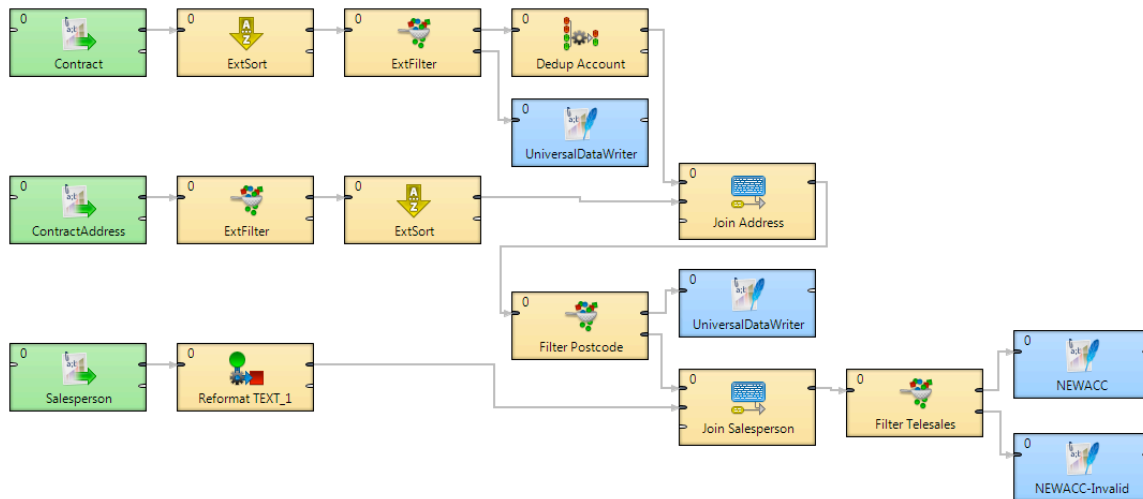
1. Data records are read from SRC I input files – delimited data
2. Applicable business filtering rules are applied to filter out invalid records
3. Data records from separate files are joined together into single NEWACC-like record
4. Sales channel lookups are performed
5. The rest of business rules is applied
6. Final fixed-length records are written into NEWACC file

CloverETL transformation graphs:

Transformation 1: SOURCE I contract-level transformation graph



Transformation 2: SOURCE I account-level transformation graph



## Run Results

During transformation runs, we have processed the following source data files (one day extract):

- account – 106 491 records (30.91 MB)
- account\_addr – 85 969 records (16.59 MB)
- contract – 25 633 records (9.96 MB)
- contract\_addr – 455 307 records (98.13 MB)
- salesperson – 615 records (0.99 MB)
- tariff – 51177 records (7.68 MB)

The resulting NEWACC file contained 39 963 records (33.33 MB). The output file from DataStage run was exactly equal to the file produced by CloverETL application.

### *Run Environment and Times for CloverETL:*

The total run time was 35 seconds (account-level) and 54 seconds (contract-level) respectively. We have used stock HP nc6120 notebook with 512 MB of physical memory with 2.00 GHz CPU.

### *Run Environment and Times for DataStage:*

The original transformation was run on 12CPU PA9000 HP-UX box (where several other unrelated transformations were executed concurrently). The respective run times were about 25 seconds (account-level) and 50 seconds (contract-level).

## Summary

- This PoC has demonstrated CloverETL can be used to fully replace DataStage ETL tool –for most types of transformations.
- During the PoC we were able to shorten the development time by approximatively 40% compared to the original transformation implementation project in DataStage.
- CloverETL provided comparable performance figures on significantly cheaper hardware. However, there were other processes loading the machine used for DataStage run which may have been distorting these performance figures.
- There is no quality difference between CloverETL and DataStage transformation outputs.
- With the **price/performance** ratio being the key comparison factor, replacement of DataStage with the CloverETL product brings great savings in both development and run project stages while retaining most (if not all) of the DataStage features.

## **Contact**

If you wish to get more information about this particular project or with any other inquiry, please visit OpenSys's web site at [www.opensys.com](http://www.opensys.com) or CloverETL portal [www.cloveretl.com](http://www.cloveretl.com).

You may also contact us via e-mail: [info@opensys.com](mailto:info@opensys.com)

### ***OpenSys / North America - East Coast***

113 N Henry St.  
Alexandria, VA 22314  
USA

### ***OpenSys / North America - West Coast***

2880 Zanker Rd.  
Suite 203  
San Jose, CA 95134  
USA

### ***OpenSys / Europe***

Kremencova 18  
110 00 Prague 1  
Czech Republic