



Building ETL Systems with SAS Data Integration Studio

November 14, 2007

Three Building Blocks of DIS

- Define target data sets to store your transformed data using Target Designer
- Write custom transformations using Transformation Generator
- Build job modules using Process Designer

DIS – Target Designer

- Target table creation
 - Copy existing data set structure
 - Design new data set structure
- Cube designer wizard

DIS – Transformation Generator

- Base SAS code
- Macros
- Define input & output parameters

DIS – Transformation Generator

The screenshot displays the SAS Data Integration Studio 3.4 interface. The main window is titled "QMS Elector Load - Set Elector Insert Dim Key Properties". The left pane shows a tree view of repositories, including "Foundations", "Integration Technologies", "QMS", "Jobs", "Reports", "SAS Data Sets", "STP", "Tables", "Transformations", and "ELECTOR". The right pane shows the source code for the transformation job, which includes two data steps and a merge operation.

```
1 data <ELECTOR_NEW_RAW_DIM (keep=ELECTOR_DIM_KEY ELECTOR_ID ELECTOR_STATUS_CODE BIRTH_DATE MATURITY_DATE
2 GENDER PD_UNKWN_IND ED_ID PROVINCE_ID ADDRESS_ID GEO_CODE_STATUS_CODE
3 ADDED_REGISTER_DATE ADDED_ELECTORAL_CYCLE_NUMBER ADDED_DS_TYPE_ID EFFECTIVE_FROM_DATE
4 EFFECTIVE_TO_DATE CURRENT_IND NROE_LAST_UPDATE_DATE ETL_UPDATE_DATE);
5 merge <ELECTOR_NEW_RAW (in=enr )
6 <QMS_SEQUENCES (in=qs);
7 by CURRENT_IND;
8 if enr=1 then do;
9 ELECTOR_DIM_KEY=ELECTOR_DIM_SEQ+_n_;
10 output <ELECTOR_NEW_RAW_DIM;
11 end;
12 call symput('elector_dim',ELECTOR_DIM_KEY);
13 run;
14
15 data <QMS_SEQUENCES;
16 set <QMS_SEQUENCES;
17 if <elector_dim > 0 then
18 ELECTOR_DIM_SEQ=<elector_dim;
19 run ;
```

DIS – Process Designer

- Visual representation of processes
- Resemble design diagrams
- Execution capability
- Deploy job for scheduling

DIS – Process Designer

The screenshot displays the SAS Data Integration Studio 3.4 interface. The main window is titled "Process Designer : QMS Transaction DS Type Exclude Load : PhobLev2". On the left, a "Repositories" tree shows a hierarchy: Foundation > Integration Technologies > QMS > Jobs > NROEBASE. The main workspace contains a flowchart with the following components and connections:

- Top Row (Data Sources):** Three orange circular nodes: "DATA_SOURCE_MOVE_EXCLUDE", "DATA_SOURCE_TYPE", and "DATA_SOURCE_NEW_EXCLUDE".
- Second Row (Transformation Processes):** Three blue square nodes: "QMS Transaction DS Type Exclude Load - Move Filter Sort", "QMS Transaction DS Type Exclude Load - Sort DATA_SOURCE_TYPE", and "QMS Transaction DS Type Exclude Load - New Filter Sort".
- Third Row (Data Sources):** Three orange circular nodes: "DATA_SOURCE_MOVE_EXCLUDE_SORT", "DATA_SOURCE_TYPE_SORT", and "DATA_SOURCE_NEW_EXCLUDE_SORT".
- Fourth Row (Transformation Processes):** Two blue square nodes: "QMS Transaction DS Type Exclude Load - Merge Move" and "QMS Transaction DS Type Exclude Load - Merge New".
- Fifth Row (Data Sources):** Two orange circular nodes: "DATA_SOURCE_MOVE_INCLUDE" and "DATA_SOURCE_NEW_INCLUDE".

Flow connections: The top row nodes connect to their respective second-row nodes. The second-row nodes connect to their respective third-row nodes. The "DATA_SOURCE_TYPE_SORT" node connects to both "QMS Transaction DS Type Exclude Load - Merge Move" and "QMS Transaction DS Type Exclude Load - Merge New". The "QMS Transaction DS Type Exclude Load - Merge Move" node connects to "DATA_SOURCE_MOVE_INCLUDE". The "QMS Transaction DS Type Exclude Load - Merge New" node connects to "DATA_SOURCE_NEW_INCLUDE".

At the bottom of the window, there are tabs for "Process Editor", "Source Editor", and "Log". The system tray at the bottom right shows the user "PhobLev2", session "sastest2 as sastest2", and the time "8:21 AM".

Important Useful Features

- Check in / Check out
- Use of other SAS tools to write Base SAS code
- Source Designer
- Update metadata

Challenges

- Diagram size in Process Designer
- Runtime space used by pre-defined transformations
- Partial Check in capability
- SPDS Clusters
 - When cluster members are generated from different processes

Environment Migration

- Export
 - From Foundation
 - By object type
- Import
 - From Foundation
 - Cleanup first

Execution & Scheduling

- Job execution from DIS
- Deployment for scheduling
- SAS – Management Console – Schedule Manager

Conclusion

- The SAS Data Integration Studio tool has the capabilities to construct full ETL systems.
- Three steps to build ETL systems using DIS:
 - Build target data sets
 - Write transformations
 - Develop processes
- Important extra features:
 - Check In / Check Out
 - Migration tools (Export/Import)
 - Job execution and deployment
- Some limitations:
 - Size of processes
 - Using predefined transformations
 - Partial Check In

Questions & Answers