



Design Tip #65 Document Your ETL System

By Joy Mundy

Whether you use an ETL tool or hand-code your ETL system, it's a piece of software like any other and needs to be documented. As your data warehouse evolves, the ETL system evolves in step; you and your colleagues need to be able to quickly understand both the entire system architecture and the gritty details.

There's a widespread myth that ETL tools are self-documenting. This is true only in comparison with hand-coded systems. Don't buy into this myth: you need to develop an overall, consistent architecture for your ETL system. And, you need to document that system. Yes, writing a document.

The first step in building a maintainable ETL system is to STOP and think about what you're doing. How can you modularize the system? How will those modules fit together into an overall flow? Develop your system so that you use a separate package, flow, module (or whatever your tool calls it) for each table in the data warehouse. Write a document that describes the overall approach – this can be a few pages, plus a screenshot or two.

Design a template module and group like activities together. The template should clearly identify which widgets are associated with extracts, transformations, lookups, conformation, dimension change management, and final delivery of the target table. Then, document this template flow in painstaking detail, including screenshots. The documentation should focus on what's going on, not on the detailed properties of each step or task.

Next, use the templates to build out the modules for each dimension and fact table. If you can control layout within your ETL tool, make the modules look similar, so people can look in the top-left for the extract logic, and can more easily understand the squiggly mess in the middle. The modules for each dimension table should look really similar to each other; likewise for fact tables. Remember: it's only a *foolish* consistency that's the hobgoblin of small minds. The table-specific documentation should focus on what's different from the standard template. Don't repeat the details; highlight what's important. Pepper your ETL system with annotations, if your ETL tool supports them.

Finally, your ETL tool may support some form of self-documentation. Use this feature, but consider it an appendix to the real document as it's either relatively lame (screenshots) or overwhelmingly detailed (all the properties of all the objects); it's not, in our experience, particularly useful.