CLAUDIA IMHOFF

Reprinted from DM REVIEW

ETL in a Box (architect not included)

Claudia would like to thank Tom Kerr for his contribution to this month's column.

The data acquisition process of the Corporate Information Factory (CIF), in which the data warehouse and operational data store (ODS) are populated from operational sources, represents the most technically challenging part of any business intelligence (BI) environment. Some industry experts estimate that 60 to 80 percent of a BI project's effort is spent on this process alone. In today's high-volume, client/server environment, data acquisition techniques have to coordinate staging operations, filtering, data hygiene routines, data transformation and data load techniques in addition to cooperating with network technology to populate the data warehouse and ODS. Unless your CIF is very small in size or scope, these are usually separate processes that must function as one smoothly operating unit.

An efficient, scalable CIF won't naturally become a well-oiled and finely tuned machine by itself. To borrow a phrase from a former management instructor, you must succeed on purpose. Just as your BI environment's value depends on a proven architecture (such as the CIF), the complete extract, transform and load (ETL) environments also depend on a solid architecture.

A common misconception is that it is not necessary to waste time on upfront design or ETL architecture. The reasoning approximates the following: "Now that we have a tool to generate code so quickly and easily, there is no reason to delay getting started. We can afford to learn as we go. After all, data warehouse construction is iterative, right?" It is not unusual for BI project managers to have the attitude that ETL architecture all but goes away because the new software will create one for them. The only thing needed is some knowledge of the source and target databases and how to enter that information into the tool. These attitudes are often driven by the need to justify the expense of the tool by showing management quick results as promised during the tool selection process.

While it's true that software tools are invaluable for more efficient and timely movement and transformation of data, an integrated and sustainable business intelligence environment requires thought and planning on the interaction of the various components and processes.

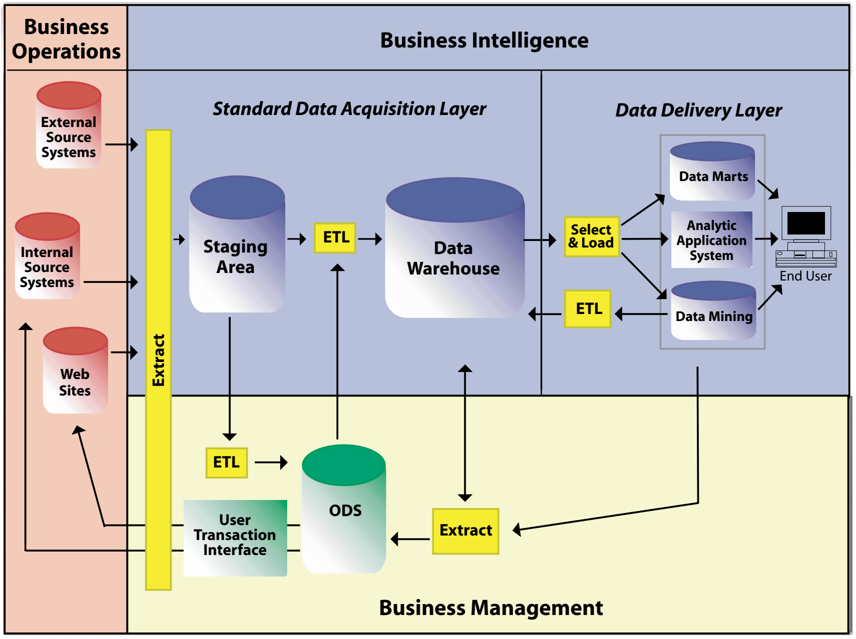

Figure 1: ETL in the Corporate Information Factory

The CIF architecture illustrates where the various ETL processes take place (see Figure 1). It's more than just in the data acquisition process! While data acquisition is the predominant process using the ETL tools, the data delivery process and movement of data from the analytical functions to the ODS or operational systems use ETL processing as well. The full-blown set of ETL operations must combine into a cohesive, integrated system – one that ensures each process will fit into the overall effort efficiently, determines how the tool will be used for each component and synchronizes all ETL events.

You should have an ETL master craftsman who ensures that the ETL processes have strength and endurance. This requires an overarching view and control over the entire environment and is the job of an ETL architect. The ETL architect ensures program efficiency by creating a cohesive ETL architecture to ensure that the various ETL functions form one cohesive system.

To quote a colleague, "You must go slow to go fast." Taking the time to properly architect a highly integrated set of processes and procedures up front is the fastest way to achieve a smoothly running system that is maintainable and sustainable over the long haul. To accomplish an efficient, scalable and maintainable process, the ETL architect must have the following roles and responsibilities:

- The ETL architect must understand the overall technical environment and strategic performance requirements of the proposed system. The architect must interact with the source system technical staff, the project database administrator (DBA) and the technical infrastructure architects (if different from the ETL architect) to develop the most efficient method to extract source data, identify the proper set of indexes for the sources, architect the staging platform, design intermediate databases needed for efficient data transformation and produce the programming infrastructure.
- An ETL programmer may only see his or her single-threaded set of programs. The architect must see the entire system of programs, how they will influence and affect each other and, ultimately, how the software coding tools must interact with the technical infrastructure to create a seamless environment. S/he must ensure the technical team understands the target database design and its usage so that the transformations which convert the source data into the target data structures are clearly documented and understood. The ETL architect oversees each and every one of the ETL components shown in Figure 1 and their subcomponents.
- The ETL process is much more than code written to move data. The ETL architect also serves as the central point for understanding the various technical standards that need to be developed if they don't already exist. These might include limits on file size when transmitting data over the company intranet, requirements for passing data through firewalls that exist between internal and external environments, data design standards, standards for usage of logical and physical design tools and configuration management of source code, executables and documentation. The ETL architect must also ensure that the ETL design process is repeatable, documented and put under proper change control.
- A key consideration for the ETL architect is to recognize the significant differences that the design and implementation methods for a business intelligence system have from an online transaction processing (OLTP) system approach. An OLTP system only changes in design when the operational process it manages changes, while BI systems must constantly adapt as business users discover new and different ways of analyzing their businesses. BI systems must be scalable from a volume perspective but must also adapt to changing business processes and technologies without requiring a complete redesign and conversion. For example, the current Web-based business environment demands that BI not only address strategic needs but integrate on a tactical level with day-to-day processing. This requires a forward-thinking architect who recognizes that the ETL processes must integrate with the needs of a real-time warehousing effort. (Note: This is indicated in Figure 1 by the lines linking the warehouse and the data delivery layer back to the ODS and then back through the user transaction interface to the Web site.)
- The role of the ETL architect also extends to that of consultant to the programming effort. The architect works closely with the programmers to answer questions and plays a key role in problem resolution. Depending on the size of the programming effort and the project organization, the ETL architect may also oversee the development of the programming specifications. In any case, the ETL architect plays a key role as a reviewer and approver during the peer review process.
- One last role for the ETL architect must be to ensure that the various software tools needed to perform the different types of data processing are properly selected. The yellow boxes in Figure 1 show each point in the CIF architecture requiring some kind of extract, data transformation or data load operation. Each of these ETL functions has a different purpose and, as such, may not necessarily require the same functions within the software tool. Under the guidance of the ETL architect, a well planned and documented ETL architecture, at least at a high level, will define these purposes and functions as input into the tool selection process.

ETL is one of the most important sets of processes for the sustenance and maintenance of your BI architecture and strategy. Time and thought are required to ensure the best architecture for its various components as well as for the selection of appropriate software tools and procedures within each component. Ongoing BI development demands a flexible, scalable and easily maintainable environment that can only come from an architected approach.

This type of architecture must be driven from a central focal point led by the ETL architect. The need for an ETL architect should be obvious for large systems growing exponentially. However, the need for an ETL architect is no less important in smaller environments. Small systems have a way of growing, and the smart development team will be ready, willing and able from the start to take on the growth with a well-architected environment.

Claudia Imhoff, Ph.D., is the president and founder of Intelligent Solutions (www.intelsols.com), a leading consultancy on CRM and business intelligence technologies and strategies. She is a popular speaker and internationally recognized expert and serves as an advisor to many corporations, universities and leading technology companies. She has coauthored five books and more than 50 articles on these topics. She may be reached at cimhoff@intelsols.com.

Tom Kerr is a senior project manager at IBM. Tom has 29 years of application development and management experience. He has spent the last nine years building and managing business intelligence systems for some of the nation's largest corporations. Kerr can be reached at tkerr@nc.rr.com.