The Benefits of Centralizing Data Mappings for ETL and Data Warehouse Projects

Creating a Knowledge Gap

Despite all kinds of advances in database and data warehouse technology, it still takes just as long to complete an ETL or data warehouse project as it did 10 years ago. Why? Part of the reason is that manual processes continue to bog down development.

One of those manual processes is collecting and coding source-to-target mappings. Identifying data sources and data transformations is an essential preliminary step in any data integration or data warehouse project. Business and data analysts and data modelers specify the sources of data feeds and how data from those sources should be transformed. Where these specifications are typically stored? In spreadsheets.

Analysts give the spreadsheets to developers, who begin building the integration solution or data warehouse.

More often than not, developers end up writing code that varies from these specifications. They might discover that new data sources have superseded old ones. New industry regulations might change project requirements. Newly identified sources might require different transformations.

Developers rarely update the original spreadsheets once coding has begun. As the project progresses, the original set of spreadsheets becomes increasingly obsolete in unpredictable ways, creating an ever widening knowledge gap for everyone involved. Business analysts and architects still have their original spreadsheets, but they can’t be sure how closely the spreadsheets describe the feeds being built. If new developers or analysts are assigned to the project, or another team wants to embark on a similar project, they have to look at the code base itself to figure out which data sources are really being used and how data is being transformed.

This knowledge gap between spreadsheets and code leads to busy work, more time on projects, and increased chances of error. Of course, the problem is bigger than just ETL source mappings. Most projects also lack an authoritative centralized source for architectural and technological standards, naming conventions, coding techniques, and history management rules.

Streamlining Development through Centralization

Rather than relying on spreadsheets, a better approach is to first create a searchable centralized repository for all these specifications and standards and then to ensure that the code is generated from the current specifications and standards in this repository. The centralized repository should include:

- Source-to-target mappings and transformations.
- Naming conventions.
- Architectural standards and implementation patterns.
- History management rules.

- Selection of coding techniques.
- Operational data.

The repository should enable developers and others to trace dependencies, analyze impact of changes, and produce data lineage across projects. It should also support versioning, so that business managers, developers, and other stakeholders can see how specifications have changed over time.

Ideally, organizations would be able to automatically generate ETL and SQL processes directly from the specification repository. By automating code generation, development teams could easily keep up with ongoing requests for changes from business managers and architects. Business managers would be able to focus on business goals and requirements, developers to concentrate on best practices and code quality, and both groups could be confident that the latest code was systematically correlated to the latest specifications.

Spreadsheets are powerful tools. But just as they're not a substitute for data warehouses, they're not a substitute for a centralized repository for data warehouse specifications.

Gamma Systems Data Warehouse Studio

Gamma Systems Data Warehouse Studio is a development platform that enables software architects, data modelers, and business analysts to define business rules, data mappings, and other design elements and store them in a searchable central repository. Once these aspects of the integration or data warehouse project have been defined, the platform automatically generates over 95% of the SQL and ETL code required for the project.

If requirements or specifications change, the development team can quickly generate new code. Specifications and code are always in sync, and line-of-business managers can always be sure that the latest code reflects the latest requirements, business rules, and regulatory guidelines.

By centralizing project requirements and generating SQL and ETL code, Gamma Data Warehouse Studio enables enterprises to reduce the costs and development times of data warehouse and data integration projects by 50% to 70%. Now that's a result that business managers would like to see anywhere even in a spreadsheet.

Article by Simon Eligulashvili is founder and CEO of Gamma Systems, Inc.

Visit our Block on Informatica Marketplace

Gamma Data Warehouse Studio

INFORMATICA
The Data Integration Company™

Worldwide Headquarters, 100 Cardinal Way, Redwood City, California 94063, USA
phone: 650.385.5000   fax: 650.385.5500   toll-free in the US: 1.800.653.3871   www.informatica.com