

# QUICK TAKE



February 22, 2005

## Microsoft Addresses Enterprise ETL

SQL Server 2005 Integration Services Will Replace Data Transformation Services

by **Philip Russom**

with Connie Moore and Colin Teubner

### EXECUTIVE SUMMARY

Microsoft's new tool for extract, transform, and load (ETL) addresses enterprise ETL requirements like collaborative development, dedicated administration, and server scalability. It also goes beyond ETL to include functions related to data integration, such as data quality, data profiling, and text mining. These advancements bring Microsoft out of the data utility ghetto where Data Transformation Services (DTS) played and into the enterprise ETL arena, where SQL Server Integration Services (SSIS) will compete favorably with midtier vendors.

### RESEARCH CATALYST

Microsoft executives recently unveiled Integration Services, a new ETL tool to be released concurrently with SQL Server 2005 (code-named "Yukon") in mid-2005.

### WHAT IS INTEGRATION SERVICES AND WHY SHOULD ANYONE CARE?

SSIS is a new product for ETL and other data integration functions that replaces the older product SQL Server 2000-based DTS. In a nutshell, the new SSIS has far better tools for development and administration, which increase its usability and collaborative capabilities. Plus, it adds advancements to SQL Server 2005 for greater scalability, reliability, and real-time data cleansing.

All totaled, these advances put Microsoft in the enterprise ETL arena. SSIS' enterprise capabilities will have a strong impact on the ETL vendor landscape, because it's quite competitive with other ETL offerings. SSIS will also have a positive impact on data warehousing professionals, who have wanted to use DTS because of its low price, but found it lacking key enterprise ETL requirements like usability, collaboration, and administration — requirements now satisfied by SSIS.<sup>1</sup>

As with DTS and SQL Server 2000, enterprise licenses of SQL Server 2005 include SSIS at no additional charge. DTS has always had an appealing price-to-performance ratio, but the new performance level of SSIS makes the ratio downright seductive. This will bring many new data warehousing professionals and other integration specialists to the Microsoft technology stack.

### SQL SERVER INTEGRATION SERVICES TAKES A GIANT STEP TOWARD ENTERPRISE ETL

On the one hand, DTS has advanced very little since its initial release in 2000, so it's overdue for an overhaul. On the other hand, in one fell swoop SSIS moves ahead in areas important to enterprise ETL:

- **Collaborative development tool.** In the user survey of a recent ETL Wave™ project, Forrester found that 25% of user organizations interviewed were already practicing collaborative ETL, which is a best practice for teams of multiple ETL developers, administrators, and managers. Forrester predicts that 45% of ETL users will practice collaborative ETL by 2007.<sup>2</sup> Microsoft SSIS enables collaborative ETL with the new development tool Business Intelligence Design Studio. It's hosted in Visual Studio, where it automatically takes advantage of Visual Studio's collaborative capabilities for source code management, version control, and multi-user project management.
- **Separate management tool.** In a recent survey of enterprise ETL users, 17 out of 28 respondents said they prefer having an ETL administration tool that is separable from the ETL development tool. That's because more and more enterprise ETL is practiced by large teams that have a member or two devoted to administrative tasks exclusively. With SSIS, the separate admin tool is the SQL Server Management Studio, which is where users operate and maintain business intelligence database objects, as well as all SQL Server 2005 components.
- **Data quality and profiling functions.** Like all data integration technologies, ETL reveals problems with data quality that ETL must deal with. ETL specialists incorporate data quality functions either by using a third-party product outside of the ETL data flow or by designing transformations that handle data quality directly in the data flow. SSIS enables the latter with new capabilities called Fuzzy Lookup (which matches incoming "dirty" data with clean records in a reference table) and Fuzzy Grouping (which detects similarities between input rows and merges duplicates). SSIS complements these runtime data quality functions with design-time data profiling capabilities.
- **Data lineage and impact analysis via broad metadata management.** All the services of SQL Server 2005 share common metadata, resulting in broad metadata-based visibility from data sources, through ETL processes, and into reporting and analytic processes. Furthermore, developers can link an SSIS package to a report. This enables downstream impact analysis, so IT can see how a change in a source system will affect not only ETL but also reports. It also enables the reverse, namely upstream data lineage, where a report consumer can drill all the way back to the source of a report's data.
- **Scalability and reliability.** These are hallmarks of enterprise ETL, where data volumes are huge and data has to be delivered on time. SSIS addresses scalability with data flow pipes that are multithreaded by default. With dual-core CPUs coming from Intel, SSIS packages will automatically process in parallel to get the most out of Wintel platforms.<sup>3</sup> And SSIS now includes a fully parallelized sort engine. For reliability, SSIS supports check-pointing, although this depends on a package's design.
- **ETL that behaves like enterprise information integration (EII), on occasion.** Whereas most ETL tools load data into a database or file from which applications access it, SSIS can pass

data directly to an application, similar to the way EII works. The benefits are that information delivery approaches real time and that IT needn't design and maintain a persistent data store. In SSIS' new Design Studio, an ETL developer simply defines the target application as a "data reader" (a .Net concept). For example, SSIS could pass data directly to Reporting Services to refresh a report in real time. The need for real-time data delivery via enterprise ETL is heating up, so EII-ish capabilities like this are receiving greater prominence.

### DON'T WORRY: THE IMPORT/EXPORT WIZARD IS STILL THERE

Forrester estimates that less than half of DTS users use it for real ETL. Instead, a majority of users apply DTS to simply copy data from one database to another or to load data into SQL Server from flat files, spreadsheets, OLEDB sources, and so on. Users perform these tasks with the user-friendly Import/Export wizard, which is a much beloved tool among database administrators. Users should rest assured that this popular wizard lives on in SSIS, ready to help administer SQL Server 2005 and to help with other nondata warehouse uses of ETL.<sup>4</sup>

### DTS-TO-SSIS MIGRATION WILL TAKE SOME WORK — BUT IT'S WORTH IT

Users of DTS will recognize in SSIS' Business Intelligence Design Studio the familiar paradigm of the ETL "package." DTS packages are analogous to the "objects" and "jobs" found in other ETL tools. The catch is that SSIS packages differ in format from DTS packages, so DTS users must convert their packages before using them with SSIS. Recent chatter on SQL Server blogs shows that users are eager to use SSIS' advancements, but are concerned about the migration from DTS. This is natural, since all DTS implementations include lots of packages, and there's a brisk trade of packages through communal Web sites.

But Microsoft is on top of it. SSIS includes a conversion utility that can handle almost all packages. No doubt, as thousands of DTS users embark on the journey to SSIS, blogs and community Web sites worldwide will bulk up with conversion tips. After conversion to SSIS, users can again swap packages, as long as they're in the new format.

As with any radical software upgrade, a few package features will require manual revision:

- **You have to modify your self-modifying packages yourself.** A common practice with DTS is to create packages that modify themselves on the fly to achieve dynamic behaviors. Because each is unique, SSIS' conversion utility cannot convert self-modifying packages, so users must convert these manually. Furthermore, SSIS supports variables and configuration infrastructures that result in dynamic packages. So the self-modifying package is probably an endangered species.
- **Many references to DTS' object model require manual conversion.** SSIS' object model differs considerably from that of DTS. So, references to parts of the model that no longer exist will

require developer attention. For example, ActiveX script references via the DTSGlobalVariables Parent property are no longer valid. Since Dynamic Property Tasks refer to the absent parts of the old object model, these too will require attention.

- **Many data tasks need manual updating.** This is because a new Data Flow Task in SSIS replaces DTS' Data Pump Task, Data Transformation Task, and Data Driven Query Task.

## RECOMMENDATIONS

### GET READY TO TAKE SSIS WHERE YOU DIDN'T DARE TAKE DTS

- **Use SSIS even more broadly than DTS.** SSIS serves the same broad community of users that DTS did, ranging from DBAs to data warehouse professionals to analytic application vendors. Yet, to this community, SSIS adds corporate organizations that need a more productive, powerful, and modern tool than DTS to accommodate the demands of enterprise ETL.
- **Consider SQL Server Integration Services for larger projects.** Now that Microsoft has addressed many of enterprise ETL's requirements, Integration Services is better suited than DTS for large, enterprise-scope projects that require a collaborative development environment, separate administrative tool, and options for data quality and profiling.
- **Expect to spend time converting old DTS packages to the new format.** But this is time well spent, if you want to reap the benefits of SSIS — especially its new development and admin tools, greater server scalability, and real-time data quality functions.
- **Try out SSIS today, but wait for customer references before committing.** After all, SSIS isn't yet released, so there aren't many customer references to check. In the short term, however, companies in the SQL Server 2005 beta program can get a copy of Integration Services and conduct an evaluation.

## ENDNOTES

- <sup>1</sup> Three requirements distinguish enterprise ETL: scalability, connectivity, and collaborative development. See the December 8, 2004, Tech Choices "How To Evaluate Enterprise ETL."
- <sup>2</sup> With ETL teams evolving from one or two members to as many as seven, software automation and best practices for collaboration will make or break ETL team productivity. See the December 17, 2004, Trends "Collaborative ETL Is Coming, So Get Ready."
- <sup>3</sup> More than half of ETL production platforms run Microsoft Windows 2003 on Intel-based hardware, and the number is increasing. See the December 21, 2004, Trends "Hubs And Windows Dominate ETL Production Environments."
- <sup>4</sup> Nondata warehouse usage of ETL has increased this decade to almost 20% of all ETL usage. See the November 30, 2004, Best Practices "ETL's Brave New World Beyond Data Warehousing."