



Home > News

<http://www.tdwi.org/News/display.aspx?ID=9109>

"New" ETL Technologies Challenge "Legacy" Stigma

9/10/2008

By Stephen Swoyer

[Looking for BI/DW news?](#)

Sign up to receive TDWI's free newsletters, offering the latest business intelligence and data warehousing news and analysis.

More information & subscribe

It's a puzzle. After a round of heavy consolidation in the data integration (DI) market -- during which DI best-of-breed products from Ascential, Sunopsis, Business Objects (which acquired its Data Integrator tool from Acta), and DataMirror were snapped up by larger, non-best-of-breed vendors -- many concluded start-up innovation in the DI segment was all but dead.

After all, how could start-ups with a fraction of the resources of an IBM Corp., Oracle Corp., or SAP -- not to mention an Informatica Corp., Microsoft Corp., or SAS Institute Inc. -- hope to compete?

Something close to the opposite happened. Innovation might not exactly be *thriving* in the DI segment, but it's far from moribund. In the last year alone, upstart DI players such as Expressor Software Corp. and Talend arrived, touting two decidedly dissimilar -- but characteristically "next-gen" -- takes on DI. Their marketing pitches might differ dramatically, but both Expressor and Talend tout a flavor of "fresh" DI -- offerings they claim are faster, more flexible, and less "legacy" than established products.

According to Expressor officials, for example, the DI suites marketed by IBM, Informatica, and other heavyweights are powered by aging or "legacy" ETL technology that doesn't scale as well or which isn't as "smart" as Expressor's stuff. Expressor officials might not use the term "monolithic" to describe their established competitors, but there's little doubt that's what they mean (see <http://www.tdwi.org/News/display.aspx?ID=9029>).

"Our strategy is to say, we do it better than Informatica and [IBM] and Ab Initio; we simply scale better." says Expressor chief scientist John Russell. "Just look at our architecture. We designed it from the ground up, over [the course of] a few years. It's fast -- we believe our engine is going to be the fastest engine in the industry. It's flexible: whether we do batch processing or real-time processing, we use the same engine. We're a lot more flexible [than competitive offerings]."

Talend officials don't take quite so stark a position, preferring to talk up Talend's open source pedigree and its intriguing (if not unique) approach to data integration. However, even Talend officials don't shy away from playing the legacy card, contrasting the "overhead" of an IBM or an Informatica with what they say is the "simplicity" of Talend Enterprise Suite.

"One of the big advantages of our [software] is its simplicity. It's just a job file and a batch execution. You don't have all this overhead like you get from our competitors," says Yves de Montcheuil, vice-president of marketing with Talend. "We can design data integration processes and expose them as services. [You] execute them on the fly whenever you want to get the data."

These attacks have some DI players who feel their products are unfairly cast as legacy and monolithic (in effect, *obsolete*) crying foul.

"While 'freshness' is certainly preferable for fruits and vegetables, its value in technology is more questionable," asserts Ken Hausman, product marketing manager for data integration with SAS.

Hausman believes that such marketing is misleading. It's true that most of today's market-leading DI platforms were first built 10, 15, or even more years ago to address then-pressing integration needs, he says. It isn't as if data integration hasn't changed, however. While the practice of data integration changed, DI players failed or were too slow to respond. Integration platforms evolved over time -- almost always in response to the changing state of DI.

The result, Hausman claims, are DI stacks that both technologically and conceptually bear little semblance to their seminal predecessors. "[I]n its formative stages, the data integration market needed to create enterprise-level order within its data management practices and the concept of 'one version of the truth' served its purpose," he comments. That just isn't the case anymore. "Over the last thirty or so years, this market has evolved, and many companies recognized that they needed to move beyond this cliché to what I would call 'one truth with many versions.'"

There's a further wrinkle here, too, he maintains: in several cases, the new features being touted by today's DI upstarts aren't necessarily all that new.

Take Expressor, which likes to trumpet its "semantic" metadata repository. According to Expressor, it lets individual business units preserve the illusion of autonomy -- i.e., manufacturing gets to call its widget an "element" while logistics gets to retain its "type" designation -- even as it reconciles metadata definitions across an organization. "In those two

areas [i.e., manufacturing and logistics], you're able to set your own context, so they can say, 'We still have our own way of storing that,'" explains Michael Walclawiczek, Expressor's marketing chief.

There really isn't anything earth-shattering here, Hausman argues. Both SAS Enterprise ETL and competitive DI platforms from IBM, Informatica, and others are able to achieve the same thing, he claims.

"[DI players] have addressed this issue by allowing for different repository architectures based on the needs of their customers," he points out, citing SAS' own approach, which draws on the capabilities of its BI and analytics platforms. "[B]ecause SAS repositories also manage information from its ... business intelligence and analytics applications, as well as ... [vertically-oriented] industry solutions, users are able to use the business or technical vocabulary to which they are accustomed," Hausman says.

Consider Talend, which markets an open source ETL engine that produces extracted, cleansed, and transformed data in the form of an executable binary. It's a neat idea, to be sure -- but it isn't exactly original. Evolutionary Technologies International, or ETI, has marketed a similar DI technology for going on two decades now (see <http://www.tdwi.org/News/display.aspx?ID=8958>). In addition to the Java executables produced by Talend, ETI can produce C or COBOL binaries. ETI, on the other hand, isn't open source.

There's also Expressor's claim to a kind of data integration-economy-of-scale -- i.e., a model in which subsequent DI projects (which benefit from semantic abstractions, captured business rules, and other products of earlier projects) tend to come together much more quickly.

It's a nice idea, says Hausman -- but it's one that's true of almost any metadata-driven DI toolset. "[W]ith any metadata-driven platform -- such as SAS and others -- development efficiency improves dramatically as metadata created in first round projects are subsequently reused," he argues.

Against Ageism

Like Hausman, Michael Curry, who heads product strategy and management for IBM's Information Platform and Solutions, rejects the idea that just because a technology is old (or "legacy") -- or is based on a technology franchise that's a decade or more old -- it's somehow inefficient or obsolete.

He cites the example of another so-called "legacy" platform -- the IBM mainframe -- which today bears little resemblance to its Big Iron predecessors. The modern mainframe plays host to Linux and J2EE workloads, in addition to applications written in vanilla COBOL or Assembly. Mainframes are at the heart of enterprise-wide service-enablement efforts, Curry continues, and -- with the help of specialty processors such as Big Blue's zSeries Integrated Information Processor (zIIP), which aim to make it possible for customers to move data processing workloads back to the mainframe at nominal cost -- IBM is trying to position Big Iron as an attractive host platform for enterprise data integration efforts.

The apposite point, Curry says, is that Big Blue's InfoSphere DI stack, like its mainframe franchise, hasn't gestated in a vacuum.

Thanks to technology infusions from the former CrossAccess, Ascential Software, and DataMirror -- in addition to IBM's own homegrown DI development efforts -- it simply doesn't make sense to depict the InfoSphere platform as a legacy holdover, Curry maintains.

"Sure, some of this stuff was created 15 years ago or 20 years ago -- but so was EAI. So was so much of the technology we use. Everything that's in the market now started a long time ago and evolved forward," he maintains. "It isn't a question of age. Just because something is built on older technology -- and we've done a lot of [modernization] work on our own with that product [InfoSphere Information Server] -- doesn't mean it's inefficient."

What about monolithic? Isn't there a sense in which InfoSphere -- which has become an everything-but-the-kitchen-sink repository for IBM's disparate DI and data quality technologies -- really is huge, as some of Big Blue's upstart competitors allege?

Curry prefers to position it as a "modular" play: customers buy the base InfoSphere Information Server (formerly Ascential DataStage), and then choose the options (data source connectors; data quality or data profiling features, etc.) they need. He rejects allegations of bloat as competitor-fomented fear, uncertainty, and doubt. "You look at it as [finding] the best tool for the job. When you look at the scope of problems that you solve with ETL or data integration or whatever we're going to call it, there's really not another solution that can do what [InfoSphere Information Server] does with the enormous volumes of data that [enterprises] are working with."

Stephen Swoyer is a technology writer based in Athens, Ga. You can contact Stephen via e-mail at stephen.swoyer@spinkle.net.