Top data integration trends and best practices

Data integration software is now much more than just extract, transform and load tools – it encompasses a wide range of platforms and technologies supporting real-time integration as well as traditional batch-oriented approaches. In this E-Book, learn about the trends and emerging best practices that are shaping the data integration process at leading-edge organizations.

Sponsored By:

talend®
open integration solutions
E-Book

Top data integration trends and best practices

Table of Contents

Understanding key data integration trends and business drivers

Companies choosing real-time data integration over batch-oriented techniques

Successful data integration projects require a diverse approach

Business users can no longer ignore the integration demands of SaaS

Resources from Talend
Understanding key data integration trends and business drivers

By Rick Sherman, SearchDataManagement.com Contributor

The past decade, especially with the recession and the continuing soft economy, has seen a tsunami in demand for the data needed to make sound business decisions. Yet businesses continue to fall behind when they don't approach data integration as a business-wide effort that not only drives sales and profitability but also allows data to provide transparency, privacy and security.

As information needs have evolved and grown, so has the path of data integration. Some of the important data integration trends and business drivers are described here.

Businesses need more and more data

Businesses today are hungrier than ever for information. They depend on accurate, timely information to fuel efficient operations, growth and customer responsiveness. As the volume of data grows, so does the complexity of integrating it.

Some trends fueling the exponential growth of data:

- Companies are generating more data internally. For example, the marketing group is collecting more detailed customer data from Web analytics and other customer touch points. Global companies have data from various countries to integrate, analyze and manage.

- There's more external communication with partners and suppliers. As communication between organizations increases, so does the amount of data being passed back and forth. Inventory levels, ship dates, product descriptions – each company needs the latest information so it can share the data internally and with its customers. Are we going to receive our shipment of widgets in time to meet our production goals?
the partner's product that we're selling on our website available in enough quantity to meet our holiday rush?

- There's a movement from structured data to other, unstructured sources, such as spreadsheets, documents and Web pages. Unstructured data can come from all over the enterprise. It's easier to generate but harder to integrate. This data was often ignored in the past, but businesses now realize that it's an invaluable source of company knowledge that needs to be integrated.

- Where batch data was once the norm, real-time data often is now expected. With BlackBerrys and iPhones in hand, people expect immediate gratification. Getting more data faster contributes to the growing volume.

**Businesses are understanding data integration benefits**

In order to be useful, data has to be integrated. This may sound obvious, but many businesses are really just starting to understand this. They've learned it the hard way: by allowing spreadmarts – spreadsheets created by individual users and then used for data analysis purposes – to proliferate across departments. Not only did this not deliver the information they needed, it created data silos that spawned more problems.

These spreadmarts provide inconsistent views of the enterprise and put businesses in the risky position of making decisions using faulty data. They're expensive, because each one is usually created and babysat by business professionals who should be spending time analyzing data, not gathering, massaging and attempting to integrate it.

Just knowing that they have a problem with spreadmarts doesn't resolve the problem for businesses. It takes a methodical plan to renovate or replace spreadmarts in a way that preserves the value of their business information while yielding the highest information value. Many businesses across industries have embarked on projects to leverage the business knowledge in these spreadmarts while designing data integration processes that truly incorporate that data into business decision making.
Data integration is evolving

Data integration is moving beyond data warehousing and extract, transform and load (ETL). While the basic tasks of data integration – gathering data, transforming it and putting it into a target location – sound like ETL, new data integration trends and versions of data integration tools offer processes and technologies that extend beyond basic ETL tasks. These technologies help turn data into comprehensive, consistent, clean and current information. The tools support data migration, application consolidation, data profiling, data quality, master data management and operational processing.

These tools allow businesses to determine the state of the source systems, perform cleansing, ensure consistency and manage all of the processing, including error handling and performance monitoring. In the past, IT groups had to manually build these processes into their data integration routines. Often, there wasn’t enough time or the required experience to build them properly. The latest tools on the market come pre-built with these capabilities.

In the past, ETL was limited to batch-driven, overnight operations. Data integration suites now incorporate enterprise application integration, enterprise information integration and service-oriented architecture features coupled with ETL tools to offer data integration in batch, interoperating with applications or in real-time from BI applications. As the business demands more current information, IT can perform data integration to deliver it.

Hand-coding is a hard habit to break

Despite the fact that data integration tools have evolved substantially in recent years, there’s a battle in IT: hand-coding versus ETL tools. Enterprise data warehousing has standardized on ETL tools, but downstream applications like data marts and cubes are often hand-coded. The result is that IT cannot be as responsive as the business would like, so the business then creates spreadmarts in a do-it-yourself attempt to get what it needs.

Hand-coded applications are often undocumented, hard to update and costly to modify. There's no need to reinvent the wheel and hand-code ETL when there's a large range of
excellent tools at different price points. Some are even free when bundled with other products. It is a better use of IT time and resources to use the pre-built processes to transform data, rather than building them from scratch.

Staying in touch with the evolving nature of data integration will help enterprises create deliberate processes for data integration, saving money and getting more people the information they need.
Open Source Data Quality

Regain Control of Data
* Data Analysis (profiling)
* Data Cleansing
* Monitoring of the data quality

Quality for All
* Open & extensible solutions
* No barrier to adoption: download it!
* No proven enterprise-grade solution

A Complete Solution
* Data profiling and cleansing in an integrated suite
* Native integration with Talend Integration Suite
* Professional services & support

Find out more >>

www.talend.com
**Companies choosing real-time data integration over batch-oriented techniques**

By Jeff Kelly, SearchBusinessAnalytics.com News Editor

Is there anything more frustrating – or useless – than out-of-date data? Ask any corporate-level decision maker and odds are the answer will be no.

Companies are increasingly turning from traditional batch-oriented techniques to real-time data integration to eliminate the scourge of out-of-date data. Real-time data integration can be achieved through a variety of methods, but the goal is the same: to communicate accurate, timely data from point A to point B in real time so users can make better-informed business-critical decisions.

Experts agree that real-time data integration is gaining popularity but also warn that it is not a methodology to adopt lightly.

"Recognize that the world is not a black-and-white place," said Ted Friedman, an analyst at Gartner Inc. "Any given company is going to have data integration requirements that span the latency spectrum. There are going to be pieces that are best suited to be delivered in a high-latency, batch-oriented mode, and there [are] going to be other things where real-time data integration really does have value."

**Real-time data integration options**

The most common real-time data integration method is change data capture (CDC), which also is called data replication. CDC tools and technologies recognize when an important change has occurred in one data source and, in real time, transmit the change to a given target.

Bloor Research’s Philip Howard explains: "As a change is made to a database record in your transactional system, for instance, it's also actively captured and fed through to your data
warehouse or business intelligence system, or whatever you've got running, so it's ready to answer real-time queries."

CDC is used most often to synchronize operational applications and for real-time business intelligence (BI) purposes, according to Friedman. Indeed, business intelligence is a major driver of real-time data integration adoption, he said, especially among businesses that require BI reports at a moment's notice.

For example, "if you've got some type of short-cycle business and you need up-to-the-second analysis of how your supply chain is performing, then you need to be delivering data from some data sources to your BI application in more of a real-time fashion," Friedman said.

CDC is less ideal, however, if the goal is a comprehensive real-time view of a single entity via data housed in multiple sources. For that, users more often turn to data federation, sometimes called enterprise information integration or data virtualization.

"Data federation is better suited to people that are looking ... at a more narrow slice of the data landscape," Friedman said. "They want to get a complete view of a single instance of an entity – a customer, a product, an employee – as opposed to somebody who's doing historical trending in the data warehouse."

For example, an insurance agent on a customer call might use an application supported by data federation technology to search multiple data sources to obtain a comprehensive view of that customer while still on the call. "That needs to be [done] in real time," Friedman said.

Both the CDC and data federation markets are well established, Howard said, having already gone through the consolidation phase "that you tend to get once products start to mature." Large vendors like IBM – which acquired data integration specialist DataMirror in 2007 – and Oracle – which scooped up Sunopsis in 2006 and GoldenGate Software in 2009 – as well as smaller players like Teradata offer a variety of solid CDC and data federation real-time data integration tools, he said.
Friedman also identified a third approach, what he calls the messaging-middleware method, in which real-time data integration is achieved through middleware technologies that connect applications.

"Think of IBM WebSphere MQ and Microsoft BizTalk Server, and products like that, that are really meant to do granular, message-oriented propagation of data," Friedman said. "An application on one end spits out a message of something meaningful that happened, and these technologies propagate that message to another system or application in a low-latency fashion. So it's sort of like the data replication idea, but working at the application layer as opposed to the database layer."

The middleware approach is ideal for inter-enterprise scenarios, when there's a need for real-time data integration among organizations that may not have access to one another's data sources, Friedman said. A vendor might communicate an important data change to a supplier in real time using this method, for instance.

**Data quality raises real-time data integration concerns**

Both Howard and Friedman noted, however, that while there are many benefits to real-time data integration, there are numerous drawbacks as well – first among them, poor data quality. In more traditional, batch-oriented data integration processes, there is ample time to scrub and cleanse data before it reaches its destination. Not so with real-time data integration, regardless of the method.

"In the middle of that process [batch-oriented data integration], you've got a chance to actually analyze and cleanse that data," Friedman said. "In the world of real-time data integration, there's less opportunity to apply very sophisticated tools for analyzing the quality and cleansing the data." There is a higher risk, then, that data integrated in real time will be of poorer quality, incorrect or misleading.

Friedman said current real-time data integration tools are better at data transformation and cleansing than they've been in the past, but there is still plenty of room for improvement. It
is possible that someday near-perfect real-time data integration quality could be achieved, he said, as the problem is more technological than conceptual.

Both analysts said it’s also important to recognize that real-time data integration isn’t ideal for all companies and in some cases may even prove detrimental. Friedman advises users to match their data integration methods to their latency requirements. An organization that routinely analyzes certain data sets on a weekly basis, for example, would have no need for real-time data integration, which could actually cause more harm than good, partly because of the data quality concerns.

Organizational structure and corporate politics also play a role in determining the appropriateness of real-time data integration, Friedman said. If users aren’t ready to accept and use real-time data, there’s little point in integrating data in real time in the first place.

"Frankly, I know some companies that if they had real-time BI, it wouldn't matter at all because the way they make decisions, the culture and the politics of the organization are not set up for them to act on real-time information," Friedman said. "I think that's a limiting factor for many organizations today."

Howard agreed, pointing to what he called decision-making latency. "How soon can you as a human being make a decision based on new information that you're given? If you have to have a meeting with five other people and it takes two days to arrange that, or even two hours to arrange that, then you don't need real-time [data integration]," Howard said.

He added: "If you can make a decision instantly – 'Ah, this has happened, therefore I know to do such-and-such' – then that's where real-time decision making becomes important."
Successful data integration projects require a diverse approach

By Mark Brunelli, SearchDataManagement.com News Editor

NEW YORK – Organizations should avoid the tendency to take a “one size fits all” approach to data integration projects and start thinking about the best ways to unify multiple integration tools and methodologies, according to attendees and speakers at Composite Software’s Data Virtualization Day conference here.

But don’t run out and purchase every data integration-related technology on the market just yet. Instead, conference attendees said, the message of integration diversity is more about choosing “the right tools for the job” and then thinking about innovative yet sensible ways to combine various approaches.

Choosing the proper tools for data integration projects

Methods for data integration and data movement include bulk processes such as extract, transform, load (ETL); granular, low-latency data capture and propagation; message-oriented data movement; and abstracted, federated or virtualized views of data from different source systems in addition to others. And choosing the right approach – or combination of approaches – can be daunting.

Research from Stamford, Conn.-based Gartner Inc. indicates that bulk data movement is by far the most widely used and most valued choice for data integration projects. However, conference attendees pointed out that oftentimes bulk processes are a lot like throwing a bomb when all that’s needed is a bullet.

Bulk processes are useful and necessary in situations when, for example, a user is trying to store historical information but dealing with large data sets that do not have create and update times, Mike Linhares, a conference speaker and research fellow at pharmaceutical maker Pfizer Inc., said in an interview. But it’s not always the right choice.
“I think that choosing virtualization – pure virtualization where there’s no caching going on – makes a lot of sense, especially when you have very transactional systems and you need low latency and systems have a very high availability,” Linhares said. “But when you get into a situation where a system’s availability starts to become a little not-so-routine, caching becomes a very selective way of making sure that the data is available. It also becomes very useful if you’re looking at a medium-sized set of data and you actually want to improve query performance but not impact the transactional systems very much.”

Conference attendee Sravan K. Kasarla, chief information architect at Springfield, Mass.-based MassMutual Financial Group, said that while he’s seen data virtualization used for reporting and dashboards, he hasn’t seen it widely used to deliver information to a wider array of business applications.

Data virtualization, or data federation, is the process of virtually separating data from the underlying hardware on which it resides, and housing it in a semantic, or middleware, layer that can be easily accessed by applications and processes.

“I know for the [business intelligence] layer it can work very well,” Kasarla said. “But I’m trying to solve the challenge across the board, [including] information access for structured and non-structured data. That is my challenge.”

Kasarla, who was at the conference investigating innovative ways to leverage data virtualization, said he ultimately plans to deploy the technology at MassMutual as part of an information architecture revision.

While Kasarla sees data virtualization as a “must-have” technology, he warned that it’s easy for users to fall into the trap of investing in integration tools before implementing the organizational structure and acquiring the skill sets needed to manage them properly. Kasarla said he prefers to keep the number of data integration tools he uses to a minimum.

“There is not a single platform which can offer you soup to nuts, from granular data access ... all the way to ETL,” Kasarla said. But don’t, he added, “interpret that to mean that I can go and get as many choices as possible.”
Gartner offers its keys to successful data integration projects

An increasing number of organizations are spending time and energy to derive greater value from their information assets, and a greater focus on different approaches to data integration is a fundamental part of that process, said conference speaker Ted Friedman, a vice president and member of the information architecture team at Gartner.

“There is a renaissance around data,” Friedman told the audience.

Citing Gartner surveys and frequent conversations with clients, Friedman said that there are five keys to data integration success. They include standardization, diversification, unification, the ability to leverage data integration technology to its fullest and governance.

In the context of data integration, standardization means that organizations should focus on repeatable processes and approaches for dealing with data integration problems, the analyst said.

Diversification, meanwhile, is about employing a wider variety of tools, provided that they meet the needs of the business. “This discipline of data integration has many facets and many faces, and there are many ways to skin the cat, to use another cliché,” he said.

Unification, Friedman explained, is all about determining how best to link together combinations of available tools and architectures “in a synergistic way.”

Companies that manage to standardize, diversify and unify will now have some good leverage, which means that data integration will have had a positive impact on the business. But those organizations will still need to focus on ways to increase the breadth of the business impact, he added.

Organizations are increasingly looking at ways to govern data quality, data privacy, security, lifecycle management – and the list goes on, Friedman said. But they also seem to be “missing the point” when it comes to the governance of integration tools and architecture. “[Governance] is certainly an insurance policy, in a way, to get the optimal value out of all these investments,” he said.
Friedman said users that steadfastly adhere to his five points will have an easier time with integration in the future. “If you do these things, I can assure you that you have a very good chance of being a successful data integration practitioner or leader,” he said.
Business users can no longer ignore the integration demands of SaaS

By Barney Beal, SearchDataManagement.com News Director

For years, Software as a Service (SaaS) applications were the domain of the business user. Freed from the constraints of IT, a vice president of sales could subscribe to Salesforce.com, or an HR director could sign up for Workday.

And as isolated, niche applications, SaaS tools served their purpose – they were up quickly, they were easy to use and there was no huge upfront capital investment.

Then the bill came due.

Once the SaaS applications were in place, users came to like them and to want more from them. In particular, they wanted access to more data from other systems. That meant IT suddenly needed to find a way to integrate SaaS applications with one another and with legacy in-house systems.

Business users key to SaaS integration success

While SaaS-based vendors have bolstered their APIs and connectors to large, legacy systems like SAP, and while a new breed of SaaS integration vendors has emerged to fill in the holes, integrating SaaS data remains a difficult endeavor. And, according to experts, it is not something business can now simply hand off to IT. The business side needs to remain involved.

"Typically the situation is, people buy SaaS the way they bought best-of-breed applications," said Ray Wang, a partner at San Mateo, Calif.-based consulting firm Altimeter Group. "They have a specific problem and want [SaaS applications] integrated into whatever their back-end system is. As you add a bunch of SaaS applications, the question is, 'How does this fit together with my business processes?'"
Ideally, according to Jeff Kaplan, managing director at Wellesley, Mass.-based THINKstrategies, SaaS application buyers would consider integration before they make a purchase.

"Up until now, the primary concerns were the security and reliability of SaaS," Kaplan said. "More and more, people are recognizing those hurdles are far easier to overcome than integration questions, which have more unique ramifications within each organization. Before people adopt a specific SaaS solution – based on, say, a 30-day trial – what they have to consider is how that application is going to be integrated into specific workflow and legacy applications and the data source environment."

**SaaS integration challenges: a familiar problem**

Of course, technology rollouts very seldom go ideally. Most organizations have neither the time nor the resources to make long-term strategic decisions about integration when they launch SaaS applications. Despite the advances in application implementation, organizations still need to bring together IT, business and any systems integrators they may use.

"The more things change, the more they stay the same," Kaplan said. "The same three parties need to be working together just like in the old days."

As with on-premise applications, it is incumbent on the business side to ensure that the customer record is the same across all applications, according to Wang.

"The standard data integration problems come up again," he said. "You still need really good business architects that can identify the issues upfront. This is why the business side needs to get involved. You still need an architect and need to map out what are the important data values and analytics you're trying to measure."

Along with some of the familiar integration concerns that come with SaaS, companies also need to worry about data quality, Kaplan warned. SaaS integration tends to uncover the dirty data in an organization and compounds integration issues because SaaS integration requires data migration as well.
The silver lining on SaaS data integration

And while businesses can look to the past and their experience with on-premise application integration for guidance, the good news is that SaaS integration is easier.

"The good news is – the way I like to describe it – there's a shorter distance between the dots," Kaplan said. "Even though there are new data sources and applications that need to be integrated because of APIs, Web services and other de facto standards or best practices, it is possible to get the job done more quickly and cheaply than in the past."

SaaS integration vendors have done a good job of making it easy to tie into back-end financial systems, but SaaS-to-SaaS integrations with different data or process models present a challenge, according to Wang.

In addition to advances by SaaS vendors, a new set of integrators and consultants have emerged around SaaS integration. Organizations need not turn solely to the old systems integrators.

"There is certainly a growing segment in the market who recognize these newer players are more in tune [with the] challenges of integration but also the expectations of the business side," Kaplan said. "They're going to get the job done as quickly and cost effectively as possible, as opposed to the old guard, who would send an army in and camp out as long as possible."
Resources from Talend

Download Talend Open Profiler, the First Open Source Profiling Solution

Video Tutorial: Talend Open Profiler Quick Demonstration

Webinar: Leverage Open Source for Data Quality

About Talend

Talend is the recognized market leader in open source data management & application integration. Talend revolutionized the world of data integration when it released the first version of Talend Open Studio in 2006.

Talend’s data management solution portfolio now includes operational data integration, ETL, data quality, and master data management. Through the acquisition of Sopera in 2010, Talend became also a key player in application integration.

Unlike proprietary, closed solutions, which can only be afforded by the largest and wealthiest organizations, Talend makes middleware solutions available to organizations of all sizes, for all integration needs.

Visit us at www.talend.com