ORCHESTRATING A BIG DATA INTEGRATION STRATEGY

Big data environments create the opportunity to ease some of the rigidity of ETL-driven data integration processes. But increased flexibility also means more complexity—and, perhaps, a need for new skills, tools and data management practices. 

BY CRAIG STEDMAN AND JACK VAUGHAN
In February 2013, the Santa Monica, Calif., company replaced its existing relational data warehouse with a Hadoop-based system in an effort to speed up data processing and enable its business users to run more complex and data-heavy analytics applications than the old platform could support.

But the Hadoop Distributed File System isn’t the only engine under the hood of the new environment. After initially being processed in HDFS, dealer inventory information, vehicle configuration data sets and other forms of structured data are passed along to HBase, its companion NoSQL database, for storage. From there, aggregated information correlated with Internet clickstream data is transmitted to IBM Netezza and Amazon Redshift systems for ad hoc querying and to BI tools from MicroStrategy and Platfora for reporting uses, according to a June 2013 blog post by Philip Potloff, chief information officer at Edmunds.

Doing the required data integration work to tie everything together wasn’t a simple matter. Edmunds had to replace the traditional extract, transform and load (ETL) processes that fed the relational data warehouse with new manually coded integration programs, using Java, MapReduce and Hadoop’s Oozie job scheduler. Paddy Hannon, the company’s vice president of architecture, said in an interview at the Hadoop Summit 2013 in San Jose, Calif., that the work took four developers about 18 months to complete.

Copying data sets from the file structure of HDFS into a database table format for storage in HBase wasn’t that big of a challenge, said Hannon, who took part in a panel discussion at the Hadoop conference. “The more difficult part,” he said, “was unpacking the 10 to 15...
years of ETL we’d done to find out what rules were important and which weren’t.” Then the developers had to incorporate the business rules deemed worth keeping into the new implementation.

Such challenges are common on big data projects—and in many cases, the data integration process is likely to become more complicated to manage as all-encompassing data warehouses and rigid ETL routines give way to more dynamic environments involving a variety of different systems and flexible, on-the-fly integration to support specific data analysis needs. That can require a big shift in data management principles and procedures, covering data integration as well as related data cleansing and governance initiatives.

In the past, data integration in the form of ETL typically was “a self-contained process” that focused simply on moving cleansed and consolidated data from source systems to a target data warehouse, said Michele Goetz, an analyst at Forrester Research Inc. “Now you’ve got this federated environment where data can be anywhere,” she said. “And a lot of times you want to leave it where it is and just call it when it’s needed [for use on another system].”

At least, that’s where things are heading, according to Goetz and other analysts. The most prevalent big data deployment approach that Forrester is seeing among its clients is a Hadoop system tied to an enterprise data warehouse (EDW), with the two technologies augmenting one another. For example, a Hadoop cluster could serve as a staging area for data on its way to the EDW or become the primary repository for specific types of information.

Enterprise Management Associates Inc. has mapped out what it calls a “hybrid data ecosystem,” an architectural framework for big data environments that incorporates eight different categories of systems, including EDWs, data marts, Hadoop clusters, NoSQL data stores and specialized analytical databases. In a survey on big data initiatives conducted jointly by EMA and 9sight Consulting in the summer of 2012, 72% of the 255 business and IT professionals who responded said their organizations were using more than one of the eight technology platforms. Forty-six percent said they had three or more in place (see Figure 1).
Big data systems make it more feasible to store data “in a very crude fashion and refine it as needed” for particular uses, EMA analyst Shawn Rogers said. Hadoop systems and NoSQL databases can serve as “a sort of loading dock” for raw data, with data models and schemas being applied to data sets later. In such scenarios, data integration processes morph from ETL into more malleable extract, load and transform (ELT) approaches. And once data is ready for BI and analytics uses, it can be moved to the system that’s the best fit. “It doesn’t have to be so rigid now,” Rogers said. “We can apply some freedom, and some common sense, to our architectures.”

Another factor that encourages ELT over ETL in big data environments is a desire on the part of data scientists doing advanced analytics to have access to unfiltered information. “Data scientists are used to working with dirty data and dealing with the noise,” Goetz said. In fraud detection applications, for example, “you don’t clean the data at all.” The goal is to find anomalies in the information that point to suspicious transactions and activities.

But as organizations move away from treating big data analytics as a siloed application and look to use the analytical results to drive their mainstream business processes, data quality and seamless upstream
integration become more important. And the increased flexibility of big data architectures also brings a higher level of development and management complexity, which might require an infusion of new processes and skills—and even a cultural overhaul—in IT departments.

At Edmunds, Potloff wrote in his blog post, the first few months of the data warehouse replacement effort “were pretty slow going” as members of the development team learned the basics of using Hadoop technologies. Greg Rokita, the company’s senior director of software architecture and leader of the Hadoop team, said in a Q&A section of the post that the developers had no prior experience with HDFS, HBase, MapReduce and other Hadoop tools. After finding its footing, Rokita added, the team adopted strategies such as abstracting complex data sets to simplify interactions with other information and doing “continuous refactoring” of the code base to incrementally improve scalability and reliability in a controlled way.

Potloff wrote that as of June 2013, the newly combined data sets and improved processing capabilities of the Hadoop-based environment had enabled business analysts using the HBase-fed query and reporting systems to save more than $1.7 million in paid-search marketing expenses through better optimization of Edmunds’ keyword bidding process. “We gave capabilities to the business that they had never had before,” Hannon said at the Hadoop Summit. “It was well worth it in the long run.”

To help grease the data integration skids in a federated systems environment, Goetz recommends that organizations create a “contextual services” layer consisting of components such as a metadata repository, data quality and governance policies, master data management models and an enterprise-wide glossary of business terms. “Unless you have that, you’re not going to be able to put all the pieces together,” she said.

Another danger in developing data integration applications involving Hadoop clusters and other big data systems is overloading them with too much data movement. “It’s easy to write a MapReduce program, but it’s also easy to write one that doesn’t perform very well,” said David Loshin, president of consul-
tancy Knowledge Integrity Inc. “You don’t want to flood your network with just sloshing data back and forth.”

The good news is that vendors of big data technologies and data integration tools are trying to one-up each other in developing automated integration capabilities for big data environments. For some users, the tools that are available now are good enough to get them over at least basic integration hurdles. For example, there’s Amadeus IT Group SA, a travel reservations system operator based in Madrid, Spain, that is using Hadoop, MapReduce and NoSQL technologies to reduce its IT costs and support new services for travel agencies and other users of its system—including an application called Extreme Search that provides proposed trip itineraries to consumers based on a variety of customizable parameters.

A June 2013 report on big data usage and issues in the travel industry, written by university professor and author Thomas Davenport, puts creating integrated data sources first on a list of challenges that need to be overcome. Integration can be a particularly thorny task for travel companies because of their continuing use of mainframe systems at the heart of their IT architectures, according to the report, which was sponsored but not controlled by Amadeus.

Hervé Couturier, head of research and development at Amadeus, said during a joint interview with Daven-

---

THREE KEY INTEGRATION CONSIDERATIONS

Forrester Research Inc. analyst Michele Goetz cited the following questions to ask internally when developing plans for an integrated data warehousing and big data environment:

1. How does the data you’re collecting need to be used? Is it for operational purposes or analytics applications—or both?

2. What’s the nature of the data, and where is it located? Is the information structured, unstructured or semi-structured? And what systems, internal or external, is it being stored in?

3. How long do you need to retain the data and the integration links you develop? Will the integrations be persistent or one-off connections to meet short-term requirements?
port that the company’s mainframe isn’t going away anytime soon. But the integration problem is solvable, he added. “The challenge is using 30-year-old technology and how do you merge that with new technology,” Couturier said. “But we can do that. The technology is here, and now the question to a large degree is how you can get to a usable business case.”

There’s no shortage of packaged tools to choose from for use in big data integration, and there isn’t necessarily one right answer. ETL technology isn’t completely out of the picture—it still has viable applications in big data environments. Data virtualization software that pulls together information from source systems without physically moving it is another option offered by various integration vendors. Data replication, change data capture and compression technologies can all play valuable roles in integrating big data.

Database vendors that offer a mix of relational, columnar and appliance technologies are integrating the products up front to enable data to flow between them, although Rogers said that creates the potential for “stack lock-in.” In addition, vendors of all stripes have introduced connector software that can shuttle data between Hadoop systems and SQL databases. Gartner Inc. analyst Merv Adrian also pointed to Apache HCatalog, a table and storage management technology that’s being developed by the Apache Software Foundation; it’s designed to provide a shared schema and table abstraction capabilities to free Hadoop users from having to worry about where and in what format their data is stored.

But taken as a whole, the current set of integration tools still has some maturing to do. “Some easy things have been done,” Rogers said. “Over the next 18 to 24 months, I think we’ll see more sophisticated tools.” Tony Baer, an analyst at Ovum Ltd., has a similar expectation. Baer said the state of big data tools is similar to the one for BI and data warehouse software circa 1996. “Back then, the industry had to introduce things like data cleansing because, for the most part, people had simply been dealing with transactional data up until then,” he said, adding that more functional tools are needed “to help civilize and manage big data integration.”
One of the data types that fits particularly well with big data systems is machine data collected from sensors on industrial equipment and from IT system and network log files. Such data can be voluminous; Rogers noted that an airplane flight from New York to London might generate 40 TB worth of machine data. In the past, much of that information was either quickly jettisoned because of storage limitations or largely ignored because of a lack of analytics firepower. But low-cost and highly scalable Hadoop and NoSQL systems running on clusters of commodity hardware make it more feasible to retain the data and put it to use.

Compared with social media data, which carries with it as much noise as it does analytics hype, machine data is straightforward information that can provide insight into how computers and other devices are performing.

For example, General Electric is moving to collect, store and analyze sensor data on the operations of products such as its electrical turbines and jet engines. Vestas Wind Systems, a Denmark-based maker of wind turbines, is using big data analytics tools from IBM to analyze geospatial and sensor data as well as a variety of other information in an effort to optimize the location and layout of turbine farms. And Goetz cited an unnamed oil and gas company that utilized sensor data to increase the production time on its oil wells. Other potential uses include monitoring power grids to detect and avert outages before they occur and tracking the performance of IT equipment to spot performance problems or security threats.

Because machine data has a defined structure, capturing and feeding it into a Hadoop or NoSQL system typically isn’t as complex as integrating other, more unstructured forms of information is. “With process and metering data, you know what you’re getting, and it’s not necessarily taxing the system in the same way as when you’re dealing with highly unstructured data,” Goetz said. The biggest challenge, she added, is extracting the data from nontraditional data sources, such as the on-board computers on tractors. But system performance needs to be watched closely during the loading process to make sure that things don’t grind to a halt—a recommendation that Loshin seconded. “The level of scalability that’s required is immense,” he said.
ABOUT THE AUTHORS

CRAIG STEDMAN is executive editor of TechTarget’s SearchBusinessAnalytics.com and SearchDataManagement.com websites. He has been an IT journalist for more than 25 years and was an editor and reporter at Computerworld before joining TechTarget in 2009. Email him at cstedman@techtarget.com.

JACK VAUGHAN is news editor of SearchDataManagement.com. He covers topics such as big data management, data warehousing, databases and data integration. Vaughan previously was an editor for TechTarget’s SearchSOA.com, SearchVB.com, TheServerSide.net and SearchDomino.com websites. Email him at jvaughan@techtarget.com.

Freelance writer ALAN R. EARLS contributed to this report.

Orchestrating a Big Data Integration Strategy is a SearchBusinessAnalytics.com e-publication.

Scot Petersen
Editorial Director

Jason Sparapani
Managing Editor, E-Publications

Joe Hebert
Associate Managing Editor, E-Publications

Craig Stedman
Executive Editor

Melanie Luna
Managing Editor

Mark Brunelli
News Director

Linda Koury
Director of Online Design

Neva Maniscalco
Graphic Designer

Doug Olender
Publisher
dolender@techtarget.com

Annie Matthews
Director of Sales
amatthews@techtarget.com

TechTarget Inc.
275 Grove Street, Newton, MA 02466
www.techtarget.com

© 2013 TechTarget Inc. No part of this publication may be transmitted or reproduced in any form or by any means without written permission from the publisher. TechTarget reprints are available through The YGS Group.

ABOUT TECHTARGET: TechTarget publishes media for information technology professionals. More than 100 focused websites enable quick access to a deep store of news, advice and analysis about the technologies, products and processes crucial to your job. Our live and virtual events give you direct access to independent expert commentary and advice. At IT Knowledge Exchange, our social community, you can get advice and share solutions with peers and experts.