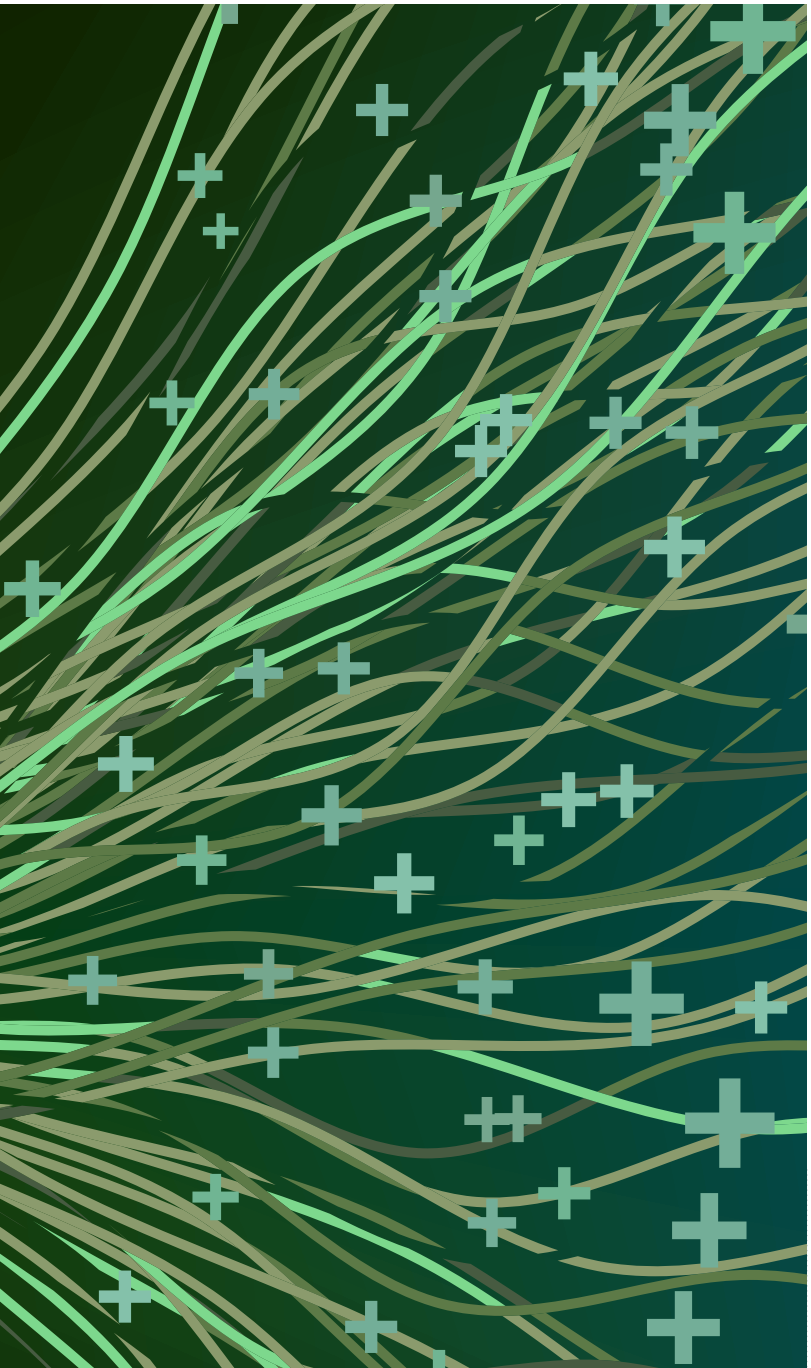
SearchDataManagement



*E-Guide*

# THE CHALLENGES BEHIND DATA INTEGRATION IN A BIG DATA WORLD

TechTarget

SearchDataManagement



**ON ONE HAND,** while big data applications have eliminated the rigidity of the data integration process, they don't take responsibility for generating new business rules to help organizations better manage large quantities of data. This e-guide reveals the increased complexity behind big data projects, and how new approaches to data integration and governance are required. Additionally, consultant Rick van der Lans offers insight on how centralizing data integration may not be the best answer to tackling organizations' ever-increasing amounts of distributed data.

SearchDataManagement



# BIG DATA APPLICATIONS REQUIRE NEW THINKING ON DATA INTEGRATION

*Craig Stedman, Executive Editor*

Edmunds.com Inc., which publishes automobile pricing data, vehicle reviews and other car-shopping information online, is driving deep into big data applications territory to power its data warehousing and business intelligence (BI) operations. In February, 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 (HDFS) 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 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 June 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 conference in San Jose, Calif. "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.

## A FEDERATED FORMAT FOR BIG DATA APPLICATIONS

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. in Cambridge, Mass. "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.

Consulting and market research company Enterprise Management Associates Inc. (EMA) has mapped out what it calls a "hybrid data ecosystem," an architectural framework for big data applications 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.

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.

## SLOW START, FAST FINISH

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. But, he added, the team eventually found its footing and adopted strategies such as abstracting complex data sets to simplify interactions with other information, and "continuous refactoring" of the code base to incrementally improve scalability and reliability in a controlled way.

As of June, according to Potloff, 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 keyword bidding processes.

"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."

# DISTRIBUTED DATA NEEDS INTEGRATING, BUT IS CENTRALIZATION THE ANSWER?

*Rick Van Der Lans, Managing director and founder*

As described in the first part of this series, there are many good reasons why data entry and data storage are dispersed. Still, data has to be integrated. And there are many different reasons why data has to be integrated. For example:

- Customer care can be improved when sales data is integrated with complaints data and data from social media -- all three being different data sources.
- Transport planning can be more efficient when internal packing and delivery data is integrated with weather and traffic data -- both being external data sources.
- Internal sales data becomes more valuable to an organization when it's integrated with, as an example, demographic data. The combination

may explain why certain customers buy certain products. The sales data may be stored in an ERP system running on local servers, whereas the demographic data is available from an external website whose physical location is completely unknown.

- To develop the right key performance indicators (KPIs), sales data must be integrated with manufacturing data.

Another reason for integrating data is purely to be able to make sense of the data. For example, sensor data coming out of high-tech machines can be highly cryptic and coded. The explanations of these codes may be stored in a database residing on another system. So to make sense of the sensor data, the information must be integrated.

It's obvious that the need to integrate data from different sources is important for every organization. And now that data has been distributed, data integration becomes an even bigger technological challenge.

For the last 20 years, the most popular place to integrate data has been the data warehouse. In most data warehouse systems, data from multiple sources is physically moved to and consolidated in one big database (at one site). Here,

the data is integrated, standardized and cleansed, and made available for reporting and analytics.

This centralization and consolidation of data makes a lot of sense from the perspective of the need to integrate data. And if there isn't too much data, it's technically feasible. But can we keep doing this? Can we keep moving and copying data, especially in this era of big data? It looks as if the answer is going to be no, and for some organizations it's already a no. Here are four problems with the data warehousing approach:

**The ever-growing amount of data.** There's a reason why big data is the biggest trend in the IT industry. The word "big" says it all. Big data is about managing, storing and analyzing massive amounts of data. And sometimes big data can be too big to move. For some organizations, the amount of data generated each day is more than can be moved across the network (depending on the network characteristics). In such a situation, when data is moved to a central site for integration purposes, the network cables will start to look like snakes swallowing pigs.

**The growing importance of operational intelligence.** End users want to work with zero-latency data that is 100% or close to 100% up to date. If data is first transmitted in large batches over the network and stored redundantly,

there will always be a delay. When users demand operational intelligence, it's better to request data straight from the source.

**Privacy.** More and more international legislation addresses storing data on individuals, such as customers, patients and website visitors. These rules are becoming tighter and tighter -- and rightfully so. This implies that when an organization needs access to demographic data on individual customers, it can't just copy and store that data in its own systems for integration purposes. The data must be used where it's stored.

**The sheer cost of storing data.** Consolidating big data is starting to become too expensive when stored in traditional SQL database servers.

Until now, centralization may have been the right approach for data integration, but as more data is entered and stored in a distributed fashion, it may not be the right solution in the near future. In the 1980s, distributed database technology moved data to the user, and for integration purposes data was moved to the point of query processing. It's now time to move the query processing to the location where the data is collected. This minimizes network traffic duplication on stored data, and it lowers the risk that data will become inconsistent -- or just plain incorrect -- and/or out of date. If the mountain will not come to Mahomet, Mahomet must go to the mountain.

SearchDataManagement




## FREE RESOURCES FOR TECHNOLOGY PROFESSIONALS

TechTarget publishes targeted technology media that address your need for information and resources for researching products, developing strategy and making cost-effective purchase decisions. Our network of technology-specific Web sites gives you access to industry experts, independent content and analysis and the Web's largest library of vendor-provided white papers, webcasts, podcasts, videos, virtual trade shows, research reports and more —drawing on the rich R&D resources of technology providers to address market trends, challenges and solutions. Our live events and virtual seminars give you access to vendor neutral, expert commentary and advice on the issues and challenges you face daily. Our social community IT Knowledge Exchange allows you to share real world information in real time with peers and experts.

## WHAT MAKES TECHTARGET UNIQUE?

TechTarget is squarely focused on the enterprise IT space. Our team of editors and network of industry experts provide the richest, most relevant content to IT professionals and management. We leverage the immediacy of the Web, the networking and face-to-face opportunities of events and virtual events, and the ability to interact with peers—all to create compelling and actionable information for enterprise IT professionals across all industries and markets.