TDWI E-BOOK

FEBRUARY 2013

ENHANCING YOUR DATA WAREHOUSE

How Big Data Technologies Can Augment and Complement Your DW

Sponsored by

IBM

tdwi.org

tdwi
THE DATA WAREHOUSING INSTITUTE

BIG DATA TRENDS: DATA WAREHOUSE AUGMENTATION

By Stephen Swoyer

Can big data technologies benefit your relational database and data warehouse environments?

Imagine using big data technologies to enrich and optimize your beleaguered relational database (RDBMS) and data warehouse (DW) environments.

It's a popular idea, and it goes by many different names. For simplicity's sake, let's call it "data warehouse augmentation." This describes a scheme for enriching traditional decision support and data warehouse–driven applications with new and nontraditional kinds of information—things the DW wasn't designed to manage.

Big data technologies, such as the open source software (OSS) Hadoop framework, are also playing host to an emerging class of complex analytics; unlike traditional analytics, these new algorithms aren't expressed in SQL—but that's not all. As a platform for both distributed computing and *distributed storage*, Hadoop gives organizations a new option for managing aged data. Instead of moving "cold" information out of the warehouse and into other contexts, such as a data vault or archive, it can be shifted into Hadoop, where it's available for interactive (and highly responsive) querying.

Crucially, data warehouse augmentation doesn't propose to displace or do away with the warehouse. Its proponents are crystal clear about this. Properly understood, augmentation is an effort to *optimize* the DW, stresses Christy Maver, big data product marketing manager with IBM.

"If we talk about some of the things that Hadoop might be able to do that the warehouse can't, that's only half the story. In doing those things the warehouse wasn't designed for, what ends up happening is that the warehouse becomes more efficient," she says. "You're moving the [workloads] to the warehouse that it's most efficient at running. You're taking *away* [the workloads] that it wasn't designed for. This isn't displacement. This is optimization."

Real-Time ROI

Data warehouse augmentation isn't a technology idea dreamed up by vendors, either. Information management specialist Marc Demarest, a principal with consultancy Noumenal, Inc., says he has several clients who are actively augmenting conventional RDBMSs with big data technologies. In most cases, Demarest explains, this means new information is flowing into—and aged information is flowing out of—the DW.

DW augmentation addresses two prevailing trends: namely, the growing glut or backlog of information and the increasing complexity of analytic algorithms. More important, data warehouse augmentation can be linked to tangible ROI: the cutting-edge or greenfield big data project *might* ultimately generate ROI, significant or otherwise; there's no guarantee

that it *will*, however. Using big data to augment your existing data warehouse, on the other hand, can significantly boost DW performance; it permits you to leverage more sophisticated (i.e., non-SQL-based) analytic algorithms and enables new (or previously unfeasible) application or workload use cases.

Take information glut, which is an acute pain point for many enterprises. In the average organization, Demarest says, the rate at which information is accruing outstrips the capabilities of its existing data management infrastructure.

Using big data to augment your existing data warehouse can significantly boost DW performance; it permits you to leverage more sophisticated analytic algorithms and enables new application or workload use cases.

"[The Hadoop File System] is a great platform [for this]," says Demarest. "You can persist a lot of file-based data quickly, with compression if you need it, and you can use MapReduce to do simple—or not so simple, if you want—ETL-like things to produce load sets for either conventional DBMSs or 'NoSQL' DBMSs."

This approach makes even more sense "if you need only a small bit of what you're persisting but want to preserve the ability to go back and get more things later, or want to have a historical/longitudinal 'data preserve,'" he explains.

Augmenting in Practice

Everyone has a different vision for DW augmentation. IBM, for example, views it as a function of three separate technologies: InfoSphere BigInsights, its enterprise-grade Hadoop platform; InfoSphere Streams, its platform for streaming or machine-generated data; and InfoSphere Data Explorer, its heterogeneous data federation layer. This approach isn't strictly self-serving. Big data implies a kind of information continuum: from structured (primarily SQL) information, which is the bread and butter of the traditional data warehouse, to semi-structured information (e.g., social media entries as well as streaming and event data from machines or sensors), to mostly unstructured data types such as voice and video. You *could* use Hadoop to process, manage, and analyze all of this information, but how is doing this any different from using a conventional RDBMS or DW to address all of your information management and analytic needs?

Monolithic solutions—or a single monolithic architecture—just won't work.

The logic of data warehouse augmentation is to use fit-for-purpose technologies to store and process information of any conceivable type. In other words, you want to leverage the technology that's best suited to each particular use case.

This means using the RDBMS, or an analytic database such as Netezza, for SQL-centric workloads; using Hadoop to parse and process semi-structured and unstructured information; and using a streaming database (such as Streams) to parse and process machine-generated and sensor data.

There's another wrinkle here, too. The architectures of the past used to be inescapably physical (e.g., codified structure, such as the data warehouse and its data model); those of the future will be highly virtual. That's where a data federation tool (such as IBM's Data Explorer) can be helpful: it uses a virtual abstraction layer to deliver consistent information access. Data virtualization (DV) knits the whole information management ecosystem together. Even though a three-pronged approach such as IBM's might seem self-serving, it nonetheless addresses the unique challenges of DW augmentation.

Killer Use Cases

DW augmentation typically targets three common use scenarios.

The first, says Vijay Ramaiah, product line manager for IBM, is a kind of big data "landing zone." This involves using Hadoop and HDFS as a preprocessing hub for information. BigInsights, IBM's enterprise-grade implementation of the Hadoop framework, fills this role, Ramaiah continues.

"The Hadoop environment [i.e., BigInsights] becomes a nice pre-processing hub, or landing zone, for your data. You can park the data there and determine [which portion of it] you want to bring into your data warehouse," he explains. "You're

using Hadoop as a staging area. You can do the same thing with streaming data: cleanse the data [and] transform it before you load it into the warehouse."

Hadoop can do a lot of things, but it isn't optimized for data in motion. That's why IBM touts Streams as a complement to BigInsights—in this case, as a tool for parsing and analyzing streaming data before it's loaded (in bulk) into HDFS. "You're going to want to parse and analyze and do some analysis on [streaming data] before you even store it. Once you decide you want to store it, you have to decide if you want to put it into a database or data warehouse," he comments. "If there's minimal value to you, put it in your Hadoop infrastructure; if it is valuable [to persist this information in a cleansed and normalized structure], then load it into the warehouse."

The decision about which data to persist and where to persist it anticipates another scenario, which IBM calls "query-able archive." This consists of shifting aged (or "cold") information out of the DW and into Hadoop and HDFS. It likewise provides a blueprint for persisting not-so-valuable data from big data sources, such as Streams.

"This is using the Hadoop environment as another way to archive your data. You might have a year or so of data in the warehouse that you're using, as well as maybe [several] years' [worth] that you're not accessing that frequently. Why not just bring this low-touch data over to Hadoop?" Ramaiah suggests. "What we're saying is that you can archive data from the warehouse into BigInsights and Hadoop and store it there, where it's always available [for interactive querying via] Data Explorer."

The third scenario utilizes MapReduce (or other, more advanced, analytic algorithms) to perform ad hoc analysis on extremely large data sets.

This might *sound* like a traditional big data or Hadoop use-case, but it isn't. That's because your data set includes structured information from the data warehouse, streaming information from Streams, and semi-structured or unstructured information from a host of different sources. "This can be done in seconds or hours instead of days or weeks. It's [a scenario] that wasn't practical or feasible but now is, thanks to these big data technologies," says Ramaiah.

Demarest and other experts say that DW augmentation is one of the most popular applications for big data. IBM's Maver believes this popularity has a practical basis: it addresses existing customer pain points. "Customers want to be able to combine the data that's in their warehouses with other data that is structured and unstructured—and get insights," she points out. "This is all about making the warehouse efficient and optimal, about being able to do things that you previously could not do. Before [DW augmentation], we just didn't have the capability to do this. Now we do."

Stephen Swoyer is a contributing editor for TDWI.

Q&A: DATA WAREHOUSE AUGMENTATION BENEFITS AND BEST PRACTICES

What is data warehouse augmentation, and is it right for your organization? If your enterprise can benefit, what's the best way to get started? For answers, TDWI spoke with Christy Maver, who handles big data product marketing at IBM.

TDWI: What is data warehouse augmentation and why is it so important?

Christy Maver: Data warehouse augmentation is building on your existing data warehouse infrastructure, leveraging big data technologies to "augment" its value. Most organizations today have one or more data warehouses, and most organizations are experiencing the pain of growing data volumes, velocity, and variety. They want to be able to leverage different types of data (structured, unstructured, and streaming) generated from numerous internal and external sources.

The trouble is that they can't easily handle these new data types with their existing data warehouse. Simultaneously, the growth of big data is having both a performance and an economic impact on the data warehouse environment. Large amounts of cold data (by which I mean rarely accessed data) in a data warehouse is straining performance and driving up costs. This can make it difficult, if not impossible, to use all the data available or even determine which data is needed.

Organizations that use big data technologies to augment—not replace—their data warehouse environment can increase operational efficiency, reduce total cost of ownership (TCO), and improve business insights using underleveraged data.

What is the biggest misconception data warehouse professionals have about augmentation?

The biggest misconception is that big data technologies replace the data warehouse. That's simply not the case. It's not a data warehouse replacement; it's an enhancement. The addition of big data technologies plugging into your data warehouse environment can make your warehouse run better by not overstraining it. The idea is to complement the data warehouse with big data capabilities that are purpose-built to process specific data types and volumes and not overtax the data warehouse.

How can an organization tell if it needs to augment its data warehouse?

A "yes" answer to any of these questions means that your organization is a good candidate for data warehouse augmentation:

- Are you drowning in very large data sets (terabytes, petabytes, or more)?
- Do you use your warehouse environment as a repository for *all* of your data?
- Do you have a significant amount of cold (low-touch) data that is not often being accessed?
- Are you facing rising maintenance/licensing costs?
- Do you have to throw data away because you're unable to store or process it?
- Do you want to perform analysis of data in motion to determine, in real time, what data should be stored in the warehouse?
- Do you want to perform data exploration and navigation on complex and large amounts of data?
- Are you interested in using your data for traditional and new types of analytics?

How do I know I can trust the data coming from these new sources? Must trust be built in a special way—that is, in a way that's different from how trust about data was built for an enterprise's *current* data warehouse?

To trust data coming from new sources, especially external sources, organizations need to govern this data just as they would other types of information, such as master data and reference data. This has been coined as "big data governance." At a high level, organizations should incorporate big data into their existing information governance frameworks. Similar to cataloging physical assets, organizations need to optimize their big data as follows:

- Metadata: Build information about inventories of big data
- Data quality management: Cleanse big data just as companies conduct preventive maintenance on physical assets
- Information life cycle management: Archive and retire big data when it no longer makes sense to retain these massive volumes

What are some specific examples of a data warehouse augmentation?

There are three main scenarios for a data warehouse augmentation: pre-processing hub, query-able archive, and ad hoc analysis.

A **pre-processing hub** is when an enterprise-grade Hadoop capability is used as a staging area or landing zone for data before determining what data should be moved to the data warehouse. Stream computing can also be used as a real-time component by processing and analyzing streaming data, without having to store it first, and determining what data should be saved—either in HDFS or the data warehouse. In some cases, data won't need to be saved; being able to process and act on information as it is happening can also reduce storage in the warehouse. With this landing zone approach, data can be cleansed and transformed before loading it into the data warehouse.

In the **query-able archive** approach, infrequently accessed or aged data can be offloaded from warehouse and application databases using information integration software and tools. Data can then be federated with data warehouse data for federated queries.

In the **ad hoc analysis** approach, stream computing enables analytics on data in motion, often enabling organizations to perform analytics that might have previously been done in the warehouse, thus optimizing the warehouse and enabling new types of analysis. Different data types (structured, unstructured, streaming) can be combined with warehouse data, enabling deep analytics to provide insights not previously possible. Plus, stream computing can act as an analysis filter to find the high-value nuggets of data that then can be stored in the data warehouse.

What are the benefits of data warehouse augmentation?

There are many benefits to a big data warehouse augmentation. Organizations can reduce TCO, including licensing and maintenance costs. They can optimize the size and performance of the DW by optimizing volume across data warehouse and big data technologies. They can reduce storage through smarter processing of streaming data. They also gain business insights by leveraging structured, semi-structured, and unstructured data sources for deep analytics and by being able to analyze data in motion.

If an organization wants to augment its data warehouse, what is needed? For example, do they need to have a specific data warehouse environment?

You can start with whatever data warehouse you currently have from any vendor. No specific warehouse is needed to enjoy the benefits of a data warehouse augmentation. IBM big data technologies are designed to plug into any existing data warehouse environment.

What is the best way to get started augmenting a data warehouse?

Start by first understanding the key challenge you are currently facing. For example, if your problem is that you have too much data in your existing data warehouse and you can't simply delete or tape-archive some of the data, then look at using an enterprise-grade Hadoop product to offload data. If your current problem is that you need to analyze or filter data that is being continuously generated perhaps for real-time decision making, then start with stream computing. If you simply need a landing or holding zone for all of your new data so it doesn't fall on the floor, then look at augmenting your data warehouse with enterprise-grade Hadoop.

BIG DATA CHALLENGING THE STATUS QUO

By Stephen Swoyer

The exhibit hall at last year's Strata + Hadoop World conference looked like a surrealistic tableau, with its juxtaposition of seemingly incongruous objects—or, in this case, vendors.

Strata + Hadoop World gave us old-school vendors such as Versant, the venerable object database company, along with wunderkind upstarts such as SiSense; established players such as Simba Technologies, the ODBC and JDBC connectivity specialist, shared floor space with machine data powerhouse Splunk. Even Microsoft was there.

Big data is, in a way, rebooting the status quo—time-tested assumptions are being tested, evaluated, and (in some cases) discarded. In the data warehousing (DW) world, this means that the arrangements or practices that worked so well for so long—or that seemed so self-evident or common sense—are now being called into question. These practices might not themselves disappear (or become extinct), but at the very least they'll be reevaluated, especially if they're products of the SQL-centrism of the DW world.

For a company such as Versant, which has no business intelligence (BI) or data warehousing (DW) presence of which to speak, this means highlighting the way object support is implemented in most SQL-centric relational database platforms.

"The analogy I've used in the past is if you have what is predominantly a relational model and then you have certain types of data that don't fit into that, what does an RDBMS do? Instead of putting your car in the garage, you put a note in [the garage] saying that your car is parked outside," says Dr. Robert Brammer, Ph.D., independent director of Versant.

Brammer is referring to BLOBs, or binary large objects, the standard way of accommodating unstructured data, such as photo or video objects, in an RDBMS.

"You have to put some kind of pointer out with some kind of kludge as to how you're going to handle this. Some of these national intelligence systems, they're all BLOBs," he says.

Brammer knows about national security. Before joining Versant, he was CTO of Northrop Grumman Information Systems. He's also a principal with Brammer Technology, a Boston-based information management consultancy.

Neither Brammer nor Paul McCullugh, Versant's executive vice president of sales, wants to champion an object-oriented data warehouse. Brammer, in fact, rolls his eyes at the notion.

At the same time, he argues, BLOBs *are* kludgy. They're kludgy because—in the relational era—kludgy was acceptable, at least for unstructured information. From the perspective of SQL-centric data management (DM) teams, unstructured information was viewed as incidental, as

peripheral; as a second-class citizen in comparison with the elegant structure of relational information. Think of it as a kind of information apartheid, in which structured SQL was accorded certain privileges—vested in the SQL-native data warehouse—that were denied to unstructured data. What's most intriguing about "big data," says Brammer, is that, as an industry-wide phenomenon, it has the potential to explode information apartheid.

It has the potential to break the power, the exclusivity, of the data warehouse.

Take Hadoop. It's a *polyglot*: in addition to SQL, Hadoop speaks (or can be made to speak) Perl, Python, Java, C++, and a host of other languages. Unlike the data warehouse, which wants to speak only SQL, Hadoop is *inclusive*: its motto could easily be "there's a library or a project or a framework for that." In a series of interviews at Strata + Hadoop World, several attendees contrasted Hadoop's inclusivity—its essential pragmatism—with what they described as the exclusivity (the essential recalcitrance) of the data management world.

A representative with a prominent data integration vendor told TDWI about a customer that wants to use Hadoop to bypass its DM team altogether.

"It's almost like a coup d'etat for them. They asked [their application developers] how long it would take to develop this [specific] application. Eight days, they were told. That's great, but when they asked how long it would take to get [data] source connectivity for these [applications]? Nine months. They were told *nine months*. Now they just want to go around [the data management group]," this representative said. "Basically, what they want is an ETL layer for all source data and they basically want Hadoop to be their new massive data warehouse."

As the teeming exhibit hall at Strata + Hadoop World demonstrated, big data is about mixing, mashing, and commingling all kinds of (seemingly unrelated) information together in a single big bowl. On the show floor, for example, some of the best-known names in the BI and DW space—Greenplum (a division of EMC), IBM, Kognitio, Oracle, ParAccel, Pervasive Software, SAP AG, SAS Institute, Talend, and Teradata—were mixing and matching with open source software (OSS) services vendors, application development tools vendors, hardware and hosting vendors, and others.

One upshot of this is that many age-old distinctions (e.g., the explicit and often antagonistic separation between application development and data management) are starting to become less important, less definitive, and less exclusive. The data warehouse will still have its place, just not *the* place. It will be one (important) source among many. First among equals, so to speak.

"I think a lot of what is becoming more widely appreciated now is that the earlier data models are being stretched past the point of breaking. This is encouraging us to look more closely [at these earlier models]. You used to have a relational model, perhaps with some sort of object-relational data mappings [i.e., BLOBs], and this was acceptable. It wasn't ideal; it was acceptable. It isn't acceptable anymore," says Brammer.

What's driving this, he continues, isn't so much the *problem* of big data (i.e., the so-called three Vs [see http://tdwi.org/articles/2012/07/24/big-data-4th-v.aspx]) but the *opportunities* presented by big data-powered products, services, and processes.

"If you look at the new generation of 'Smart Grid' problems, there are people who want to put a lot more advanced information technology into the power grid to make it much more efficient and environmentally friendly and more cost-effective, but that means they're going to put a lot more instrumentation out there that's going to collect a lot more data, much of which they're not currently dealing with at all today. This is very likely a growth area for Versant, as a high-performance [object-oriented] engine for this [kind of use case]."

Stephen Swoyer is a contributing editor for TDWI.

Q&A: GETTING A HANDLE ON BIG DATA AND HADOOP

By Linda Briggs

Cirro CEO Mark Theissen discusses the challenges of successfully working with big data, including the fast-moving market, lack of use cases, and dealing with Hadoop.

Cirro CEO Mark Theissen has spent more than 22 years in the BI, analytics, and data warehousing industry in a variety of roles. Previously, he was worldwide data warehousing technical lead at Microsoft, following its acquisition of DATAllegro in 2008, where Theissen served as COO and a member of the board of directors. Prior to that, he was a VP and research lead at META Group (acquired by Gartner Group), covering data warehousing, BI, and data integration markets.

In this interview, Theissen discusses the challenges of successfully working with big data, including the fast-moving market, lack of use cases, and dealing with Hadoop.

TDWI: Is getting a handle on big data proving to be a challenge for some companies, and why?

Mark Theissen: There are several challenges involved. First, there's no long history of use cases and successes with big data, so the people reaping advantages from big data are really spearheading the movement and driving business value.

Another challenge for companies is that when you talk about big data, you pretty much include Hadoop in the conversation. Hadoop represents a different set of skills for people who are

traditionally responsible for analytics within an organization. Typically, the people responsible for analytics are more DBA types—people who have been working on data warehouses. They tend to be SQL experts. They have some good BI tools, data visualization tool capabilities, and they know how to design and do their models, but big data is a different set of skills. It's Java. It's Hadoop. It's other NoSQL data sources. There are cloud technologies involved. It's a whole new world for them.

Yet Hadoop is not an easy tool to work with, and Hadoop skills can be hard to find.

Right. Hadoop is a processing framework, and there's a lot of work to be done to make that processing framework do the things you want to do.

Let's talk about vendors. How are they doing in terms of dealing with big data, from your vantage point?

Part of what I see is some of the same challenges customers face, including the one that you brought up, which is that finding really good resources with Hadoop skills is a challenge not only for customers but also for vendors. You can invest as a vendor and you can invest as a customer to create those skills within the company, of course, but then retaining those people is always going to be a challenge as well.

The other aspect for vendors is that this is a hypermarket, if you will. This market is moving faster than any market I've seen in the past. It's a combination of things. Take data warehousing and analytics or BI markets. Those are strong markets that are growing quite nicely in their own right, but then add big data and you have multiple markets converging on big data.

Big data itself includes aspects of cloud computing and mobile computing—all of it seems to be feeding on each other and accelerating the market to a pace that we've not seen before.

We can add social media into that mix as well.

Exactly. It's not just one thing. There are so many moving parts right now—they feed on each other. It all creates what I see as hyper-acceleration in terms of the speed at which things are moving.

When you talk to customers, what do companies typically want to do with big data—and what are they doing?

Their first challenge often lies in just being able to explore big data. It's not just, "Oh, this is Twitter data and I want to look at Twitter data," but being able to explore that data and provide context. They might have semi-structured or unstructured data in Hadoop, for example, that they want to explore. Typically, one approach is to join that data to other data sources, often in data marts or data warehouses. Combined, those sources can provide some context to what you're trying to explore. That's one of the first issues that people run into—where am I going to get the value out of my Hadoop implementation?

The other thing companies struggle with are technology decisions. Are you going to be able to use your existing BI and data visualization tools and extend them to be used against big data, or do you have to buy a completely new set of tools that can do analytics on just Hadoop? Customers struggle with that because they've made significant investments in different platforms and technologies, and they are looking to leverage those investments as they go after the riches of big data. They don't want to be told they have to completely retool to be able to get value out of big data.

In terms of use cases, we see some common patterns. For example, you might have data in a data warehouse—customer data, sales data—and have other information in Hadoop, maybe customer survey data, or data from multiple pharmacy outlets. The point is, if you have that data in Hadoop, you want to be able to combine it with the data in your data warehouse. You really don't want to have to move all the data from Hadoop into your data warehouse to be able to query it.

I think people look at that and they say, "Well, can you do it?" "Yes." "Can you do it in a timely and efficient manner?" "No." Most of the time, the customer will say, "We're not even sure if we have to move it all there anyway because we're not exactly sure which queries are the high-value ones that we want to run."

What other ways do you see customers using Hadoop?

In some cases, we see customers who want to use Hadoop as an operational data store (ODS). That's a case where people are running queries and they want to place those results into Hadoop, another relational database, a BI server, or something similar. In other cases, people want to make Hadoop a destination platform, but they still have a requirement to join to other data sources. That's usually where the challenge lies for them. It's a challenge even to do joins sometimes between multiple Hadoop clusters or between Hadoop and HBase. Making Hadoop as a destination platform—that's definitely a use case.

There's also exploration and analysis—you have data, you want to combine things, and you want to explore it. You also want to share results downstream, whether you're sharing the query so other people can also run it or you're sharing the results of those queries.

Another use case is application enrichment. You might say, "I'd like to do some processing with Hadoop. I'd like to combine the data with other data, then pump that back into something, maybe HBase, where real-time applications can access that information"—all to improve the customer experience.

Linda Briggs is a contributing editor for TDWI.

www.ibm.com

Big data is more than a matter of size; it is an opportunity to uncover insights in new and emerging types of data and enterprise content to shape business strategy. IBM's comprehensive big data platform uses sophisticated technologies, including patented advanced analytics, to solve "big data" problems and equip your organization with a competitive edge. Designed to complement your existing information infrastructure, IBM's enterprise-class big data platform lets you get started quickly today and easily expand to address more complex problems tomorrow.

- www.ibm.com/bigdata

tdwi.org

TDWI, a division of 1105 Media, Inc., is the premier provider of in-depth, high-quality education and research in the business intelligence and data warehousing industry. TDWI is dedicated to educating business and information technology professionals about the best practices, strategies, techniques, and tools required to successfully design, build, maintain, and enhance business intelligence and data warehousing solutions. TDWI also fosters the advancement of business intelligence and data warehousing research and contributes to knowledge transfer and the professional development of its members. TDWI offers a worldwide membership program, five major educational conferences, topical educational seminars, role-based training, on-site courses, certification, solution provider partnerships, an awards program for best practices, live Webinars, resourceful publications, an in-depth research program, and a comprehensive website, tdwi.org.