Best practices for managing the data warehouse to support Big Data

The new challenge for IT and data warehousing teams is how to leverage existing technology investments along with emerging tools and techniques to manage this vast amount of information – or “big data”. But “big data” makes the divide between collecting data and leveraging data more complex than ever before. Before leveraging big data, companies must put in a solid data warehousing foundation. This E-Guide focuses on how to develop that foundation, including a discussion of when to update older technologies or modify a current data warehouse architecture to match today’s new requirements.

Sponsored By: SAP
Best practices for managing the data warehouse to support Big Data

Table of Contents

'Big data' sprawl forces new thinking on data warehouse strategies

Data warehousing best practices to support 'big data': No small task

Resources from SAP
‘Big data’ sprawl forces new thinking on data warehouse strategies

By Alan R. Earls, SearchDataManagement.com Contributor

It’s an old human problem: too much information and too little capacity to process and understand it. Computer technology has helped, of course, by providing automated systems and tools for organizing and analyzing ever-larger volumes of information. But now transaction systems and other data sources, such as websites, sensors and mobile devices, are producing a veritable flood of data that is overwhelming – or bypassing – the data warehouse frameworks built up in the past.

The new challenge for IT and data warehousing teams is how to leverage existing technology investments along with emerging tools and techniques to manage this tsunami of information – or “big data,” as it’s being referred to within data management circles.

At a minimum, says Forrester Research Inc. analyst Brian Hopkins, big data will require some rethinking by organizations with traditional data warehouses in place. For instance, Hopkins noted that integrating some forms of big data into a centralized data warehouse with hub-and-spoke connections to separate data marts will be a challenge. Data warehouses are primarily focused on structured data, he said. But much of what is categorized as big data is unstructured or semi-structured information.

"In a sense, big data turns data warehousing assumptions upside down," Hopkins said. In traditional data warehousing and business intelligence environments focused on answering specific questions from business users, data is cleansed, put through an extract, transform and load process, and ultimately turned into reports or made available for analysis. Typically, less than 5% of an organization’s available data is used, “sometimes significantly less,” according to Hopkins.

**Big data’s new spin on data warehousing strategies**

By contrast, big-data strategies often focus on a broader swath of information. In addition, “the notion of a big database in the sky with a completely consistent 360-degree version of
the truth wholly evaporates,” Hopkins said. Targeted data stores and so-called analytical sandboxes are common in big-data environments, which can add management complexity for IT and data warehousing teams and require heavy-duty processing power.

“With big data, there are bound to be more patterns and anomalies that are interesting,” said Wayne Eckerson, research director for TechTarget Inc.’s business applications and architecture media group. “But computationally, it’s a lot harder to analyze big data because there’s so much of it. It gets expensive.” That’s why many organizations are looking beyond traditional data warehouse technology and considering emerging software alternatives, such as open source Hadoop and MapReduce, Eckerson added.

Richard Winter, president of Winter Corp., a Cambridge, Mass.-based consulting firm that focuses on data warehousing, said big data presents a variety of opportunities for using analytics to gain business insights that previously would have been difficult to uncover. He cited the example of a new “smart” inhaler developed for the treatment of asthma; when the inhaler is used, built-in wireless technology sends data such as the identity of the patient and the time and location to a database.

“It may be possible to combine that information with information on, say, when a load of potentially allergenic soy products was being unloaded nearby, putting soy dust in the air,” Winter said. Such correlations could help medical researchers to better understand the characteristics of asthma, he added.

**Big-data challenges: Volume and more**

But there are potential data management challenges: If the devices become widespread, Winter noted, tens of millions of people worldwide could begin using them and generating data that would need to be “stored, curated to some extent and made available for analysis over an extended period of time.”

IT and data warehousing professionals also need to understand that big data isn’t only about volume, said Philip Russom, research director for data management at The Data Warehousing Institute (TDWI). “The other attributes are just as important,” he said – for example, the fact that big data can be highly diverse, including Web clickstream data, call
detail records, point-of-sale data, text from social network posts and various other types of information.

When confronted with the challenges of managing big-data installations, many organizations end up suffering from what Russom calls the “deer in the headlights effect,” where they recognize the potential value of the information but are daunted by the difficulties of getting a handle on it – especially when much of the data doesn’t lend itself to traditional data warehousing approaches.

To avoid that kind of paralysis, Russom’s advice is to narrow the focus of a big-data management initiative to a high-payback area – customer behavior, for example – and then look at ways to apply both traditional and emerging tools to help gain control over the information. “Pick a topic that will lead you to something useful,” he said. “Don’t assume that you need to pursue every topic.”
Data warehousing best practices to support ‘big data’: No small task

By Alan R. Earls, SearchDataManagement.com Contributor

Wayne Eckerson, research director for TechTarget Inc.’s business applications and architecture media group, puts it in simple terms: If you’re going to be successful in working with “big data,” you need the right culture, the right people, the right data and the right tools. If only it were that simple to put all those elements in place.

Getting to that point requires careful planning and a clear understanding of the potential opportunities and challenges presented by big-data management technologies and processes, according to Eckerson and other analysts.

For starters, Eckerson said, “you need people at the top of the organization willing to invest” in the required technologies – and committed to instilling an analytics-oriented culture to ensure that the company will use the information “and not just revert to relying on spreadsheets” for data analysis. He added that as organizations look at how to respond to the challenges of storing and managing big data, they need to be open to the possibility of moving to more purpose-built data warehouse platforms. Such offerings can provide processing performance that’s “an order of magnitude better” than what general-purpose databases support, Eckerson said.

However, Richard Winter, president of Cambridge, Mass.-based consulting firm Winter Corp, warned that emerging technologies such as Hadoop and MapReduce aren’t the solution to every big-data management problem. Organizations need to take care not to “throw out the baby with the bath water,” Winter said. “Some people think they can now do everything in Hadoop and they can stop investing in [traditional] data warehouse technology – but that would be a terrible mistake for most enterprises.”

Winter recommended looking at applications individually and assessing which platform is best for a particular set of big data. Two key factors are how long the data will be retained and how it will be used, he said. Core transaction data belongs in a data warehouse, where it can be systematically managed for long-term usefulness and value, Winter said. On the
other hand, clickstream data, social network posts showing customer sentiment and other types of less structured data might be a good fit for a Hadoop cluster, especially if the information won’t be kept for as long as transaction data typically is. How broadly data needs to be accessed within an organization should also influence the choice of a technology platform, he said.

Volume isn’t the only characteristic of big data, according to separate but similar definitions of the term by Forrester Research Inc. and Gartner Inc.; both firms also take into account attributes such as variety and variability (or complexity, in Gartner’s model). But Forrester analyst James Kobielus said that in practice, preparing a data warehouse to handle big data is still fundamentally about scalability – and he offered three sets of tips on data warehousing best practices aimed at helping organizations to deliver more powerful and scalable systems.

**Big-data decision point: Scale up or scale out?**

First, consider upgrading and potentially building parallelism into your data warehouse architecture. Possible steps could include scaling up data warehouse server nodes based on shared-memory symmetric multiprocessors or scaling out by using server clusters or shared-nothing massively parallel processing systems, Kobielus said. Partitioning MPP installations into hub, staging and query tiers is another option. But Kobielus warned that attempting to make such changes without proper attention to the underlying technology infrastructure is likely to lead to disappointing results. For instance, he noted that single-core CPUs probably won’t measure up to MPP requirements and that storage I/O bandwidth typically must be increased to support the increased processing capabilities.

Second, Kobielus advised organizations to look at adopting data warehouse appliances in cases in which the hardware and software bundles can address specific performance issues or pain points. And third, he recommended that companies work to optimize the data management and storage layers of their data warehouses to help boost performance. That could include compressing data for maximum efficiency, improving database schemas, joins and partitions, and using nontraditional database technologies such as columnar and in-memory software “as needed to achieve specific goals,” he said.
Lyndsay Wise, president and founder of Toronto-based consulting firm WiseAnalytics Inc., said the end goals of big-data projects are often much the same as those of traditional data warehousing initiatives – for example, providing information that can help business users identify customer buying patterns or aid in fraud protection efforts. The challenges are similar, too: “There may be different nuances in terms of what you’re trying to get from the data, but the results may still depend on integration and data quality issues or on the challenges of data management and data governance,” she said.

Wise added, though, that the degree of difficulty on those challenges can be heightened by the amount of data that needs to be managed as well as its complexity, especially if a big-data project involves pulling together information from multiple data sources. As a result, companies need to honestly assess their capabilities, she advised. “Organizations want to say they have great IT people, but unless their DBAs and developers are really well-versed in data warehousing and specific big-data technologies, it almost pays to invest in outside [help] to really develop a strong platform,” Wise said.

With big data, it’s also critical to be able to frame what you want to achieve from an analytics standpoint and to determine up front what information is needed and what kind of hurdles might be involved in pulling it all together, according to Wise. “It’s really important to understand how everything interacts,” she said.

*Alan R. Earls is a Boston-area freelance writer focused on business and technology.*
Resources from SAP

SAP HANA: Helping Businesses Run Better – in Real-Time

Real-Time Business: Playing to win in the new Global marketplace

In-Memory Computing: Realized Business Benefits and The Road Ahead

About SAP

SAP’s vision is for companies of all sizes to become best-run businesses. Best-run businesses transform rigid value chains into dynamic business networks of customers, partners, and suppliers. They close the loop between strategy and execution, help individuals work more productively, and leverage technology for sustainable, profitable growth. This vision is in keeping with SAP’s mission to accelerate business innovation for companies and industries worldwide - contributing to economic development on a grand scale.