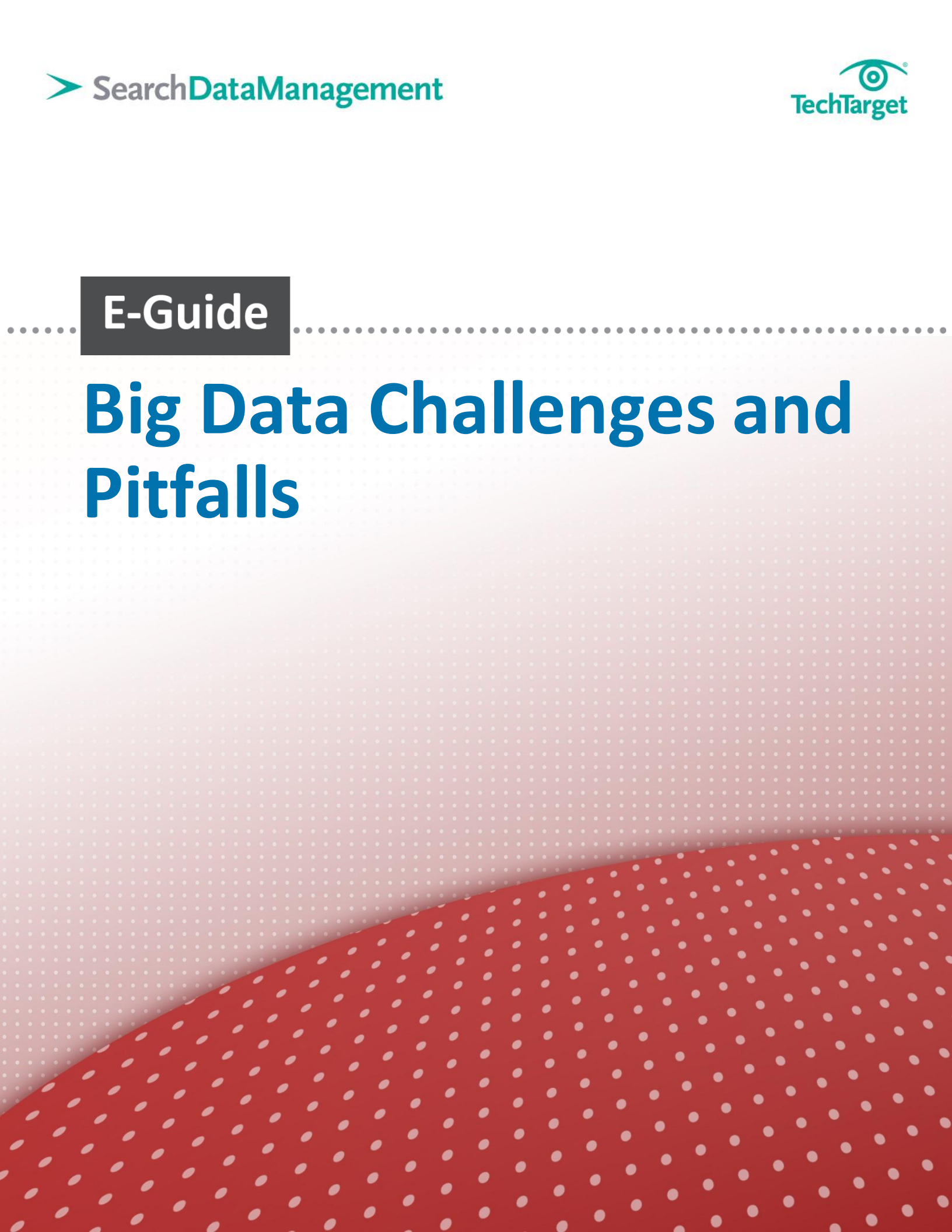
SearchDataManagement

TechTarget

E-Guide

Big Data Challenges and Pitfalls



"Big data" has already arrived in many organizations – for many others, it's only a matter of time. But like any new technology opportunity, big data comes with a raft of potential problems and issues that IT and data warehousing teams need to address. This E-Guide examines key big data challenges and potential pitfalls that organizations face as they look to store and manage large volumes of data.

Contents

Dealing with 'Big Data' Challenges: Real-World Strategies and Advice

By: Mark Brunelli, News Editor

For Eric Williams, the recipe for caapitalizing on "big data" – and avoiding the data management problems it can lead to – reads like this: Start small, quickly demonstrate business value and keep in close contact with the end users who are running analytical queries against all that information.

Williams is executive vice president and chief information officer at Catalina Marketing Corp., a St. Petersburg, Fla.-based company that uses information gleaned from retailer loyalty cards to both track and predict the shopping habits of individual consumers in countries around the world. Thanks to a combination of data warehouse appliances and predictive analytics software, Catalina has been managing, and making sense of, enormous data sets since long before the phrase *big data* entered the IT vernacular.

On a typical day, the company receives up to 525 million pieces of data from U.S.-based retailers alone. In its systems, Catalina stores about 800 billion rows of customer data comprising the purchasing history of 200 million Americans over the past three years.

Williams' advice to organizations that are launching big-data management and analytics strategies is straightforward: Avoid the temptation to gather every available piece of information and simply throw it into a data

Contents

warehouse for business users or analytics professionals to contend with. Instead, as you load your data warehouse or other databases with large volumes of information, start the analytics process by analyzing a subset of key business data for meaningful patterns and trends to prove the value of the big-data approach and gain experience in overcoming big-data challenges.

"Take a sampling of your information from a limited time period or a limited set of products and get a person on board – you probably have one already – who can help with some of the analytics," Williams said. "It doesn't necessarily require a Ph.D. person to be able to do this. Much of it is just providing insights to somebody that is already making business decisions."

Big-data management has become one of the most talked-about trends in the IT industry, as organizations look to cope with the challenges of storing large data sets and mining them for nuggets of information that potentially can provide significant competitive advantages. Complicating matters is the fact that a big-data installation might include both structured transactional data from in-house systems and unstructured information from a variety of sources, including system logs, call detail records and social media sites such as Facebook and Twitter.

Distributed computing's role in big-data management

For example, clickstream data lets companies track what people are doing on the Web, from PCs as well as mobile devices. That produces huge amounts of data, said Tony Iams, a vice president and senior analyst at Ideas International, a Rye Brook, N.Y.-based IT research firm. The benefit, Iams said, is that organizations can use that data to "create potentially a much more accurate picture of user behavior than ever before." But the data needs to be properly structured and managed to make that possible.

Jill Dyche, a partner at Baseline Consulting Group in Sherman Oaks, Calif., said classifying data is a key first step in the big-data management process. "When we talk about it with clients, we very quickly move on to data classification," Dyche said at the Pacific Northwest BI Summit 2011 in Grants

Contents

Pass, Ore. "So they're not just forklifting data onto a data warehouse platform or data marts but really looking at what the data is and how it's used."

Often, one of the defining characteristics of big data is that it's too large for a standalone database server to process efficiently. In addition, nontransactional data types such as Web logs and social media interactions – "the other big data," in the words of Gartner Inc. analyst Merv Adrian – aren't always a good fit for traditional relational databases. As a result, many user organizations engaged in big-data management employ a distributed computing, or scale-out, model, often built around open source technologies such as Hadoop, MapReduce and NoSQL data stores.

The distributed approach has worked well for Catalina Marketing, according to Williams. "This whole idea of grid computing or connecting standardized PC-type appliances and making them work in concert made all the sense in the world," he said. "That's really what has allowed us to scale to the size that we are and to do that very cost-effectively and efficiently."

Another strategy that Williams put in place is holding a monthly user group meeting designed partly to help Catalina keep its data warehouse appliances performing at an optimum level. Williams said the meetings are critical because they allow the IT staff to see how the needs of business users – and the queries they're looking to run – change over time.

"We work with them to understand how they operate, what they're running and what their analytics are showing," he said, adding that the process enabled his team to recognize that the existing data structure and query parameters "weren't optimized to accommodate what [the users] needed." The data structure has now been modified to accommodate new types of queries, Williams said.

Big-data challenges call for management oversight

For some organizations, one of the biggest challenges associated with managing and analyzing super-large data sets is finding valuable information that can yield business benefits – and deciding what data can be jettisoned.

Contents

For example, UPMC, a Pittsburgh-based health care network with 20-plus hospitals and more than 50,000 employees, has seen its data stores grow by leaps and bounds in recent years, largely because workers are afraid to delete any information, according to William Costantini, associate director of the company's integrated operations center.

"The biggest issue right now is [figuring out] what do you purge and what can't you purge, because everybody is afraid of liability and being sued," Costantini said. "Everybody is afraid to throw anything out or get rid of it. At the same time, everybody wants to be budget-conscious and keep the size down."

Adding to the big-data challenges facing organizations is the increasing popularity of "data sandboxes" that enable data analysts to explore and experiment on subsets of information, typically outside of a data warehouse. Companies need to keep a close watch on sandboxes to make sure that they don't end up with inconsistent stovepipes of data, analysts said.

In addition, the databases and Hadoop installations used to store nontransactional forms of big data are often set up by application developers working independently of the IT department. "This is being done by people outside the usual IT focus, with different tools," Adrian said at the Pacific Northwest BI Summit. "*Managed* is probably too generous a term."

Gartner's take, he added, is that organizations able to integrate those data types into a coherent information management infrastructure will outperform businesses that can't.

Trouble Spots: 'Big Data' Pitfalls in the Data Warehousing Process

By: Alan R. Earls, Contributor

"Big data" has already arrived in many organizations; in others, it's coming. And like any new technology opportunity, big data comes with a raft of

potential problems and issues that IT and data warehousing teams should approach with caution.

Contents

For example, Forrester Research Inc. analyst Brian Hopkins said that before organizations jump into big-data management, they need to figure out whether traditional data warehouse strategies and techniques will work for them in the context of information that often is unstructured and potentially not a good fit for mainstream relational databases. And if a traditional data warehousing process isn't the answer for managing big data, companies might need to "get comfortable with open source technology," such as Hadoop, MapReduce and NoSQL databases, Hopkins said.

David Menninger, an analyst at Ventana Research Inc., noted that the people within an organization can make or break a big-data initiative, particularly if new technologies are involved. Ventana recently surveyed 163 IT and business professionals from various countries on Hadoop adoption and other issues related to managing and using big data. "The participants told us the biggest obstacle for them is most often staffing and training, because these technologies are different enough and require further training and different training than what people have studied in school," Menninger said.

If you understand how processing is distributed across a Hadoop cluster, for example, you can avoid moving excessive amounts of data around and potentially get better and faster results on analytical queries, according to Menninger. But first, he said, "you need people who know how to do these things."

Menninger said the Ventana survey showed that relational databases are still dominant, even for big-data management. About 90% of the respondents said their companies use relational software overall, and 75% percent were using it as their primary technology for supporting big data. On the other hand, more than half said they were evaluating Hadoop, while 22% already were in production with the open source technology and 12% planned to begin using it within the next year. Based on the survey, Menninger said, Hadoop most often is being used to store "loosely structured data -- log and event data and to a lesser extent text and social media data."

Contents

Surprisingly to Menninger, the survey found that flat files were the second most popular big-data management technology, employed by about 70% of the respondents. "I think that is in part what leads to Hadoop," he said. "If you're working with flat files, it isn't hard to consider using Hadoop. Hadoop is really all about flat files; it's more sophisticated than that, but if you squint it's sort of the same thing."

Menninger added that companies should also watch out for two other potential big-data pitfalls as part of the data warehousing process: software licensing costs that potentially can soar along with data volumes, and inadequate integration between big-data technologies and business intelligence (BI) tools.

Big data: Too much information?

Avanade Inc., a Seattle-based IT consulting and professional services firm, also recently released a study on big-data trends and challenges, based on survey responses from 543 C-level executives and IT decision makers in 17 countries. Markus Sprenger, global director of Avanade's BI and collaboration practices, said the survey showed that one of the primary hurdles of managing big data is simply figuring out what is worth keeping and what isn't. "We found it is a question of how the business can identify relevant data and then apply it to a decision-making process," he said.

Echoing one of Menninger's points, Sprenger added that there aren't enough IT and data warehousing workers available with experience in managing nontransactional forms of big data, both within the surveyed businesses and in the job market as a whole. Organizations typically have mature processes for handling structured transaction data, he said, but most are just starting to learn how to manage large quantities of unstructured and semi-structured data in an organized way.

That's reflected on the technical side, where many of Sprenger's clients are still struggling to understand what to do with big-data installations based on Hadoop and MapReduce -- assuming that the IT and data warehousing

teams within organizations even know about the deployments and have responsibility for managing them, which often isn't the case.

"We're not at the point yet where IT can provide service-level agreements around these things -- that will probably take another year or two," Sprenger said. "It is still more of an experiment for most organizations."

Contents

TechTarget

Contents

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.

Related TechTarget Websites

SearchContentManagement

SearchBusinessAnalytics

SearchManufacturingERP

SearchSQLServer

SearchOracle

SearchCRM