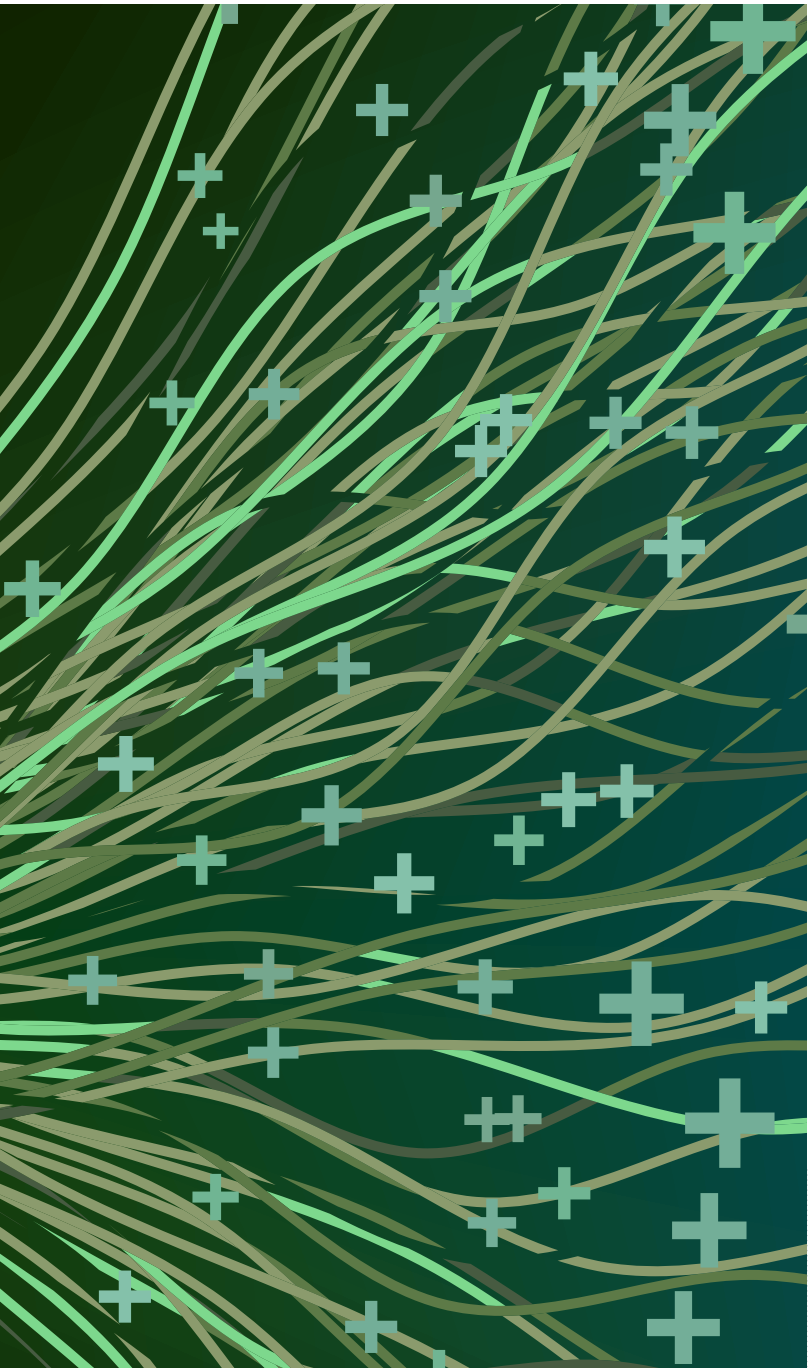
SearchBusinessAnalytics



E-Guide

HADOOP MYTHS BUSTED

TechTarget

SearchBusinessAnalytics

IN MANY ORGANIZATIONS, the growing volume and increasing complexity of data are straining performance and highlighting the limits of the traditional data warehouse. Today, Hadoop systems and related big data technologies are popping up alongside data warehouses to manage the flow of unstructured and semi-structured data. However, as Hadoop becomes a household name, it is also taking on a certain mythological form. In this E-Guide, learn some of the most common Hadoop myths.

TDWI KICKS 12 COMMON HADOOP PROJECT MYTHS TO THE CURB

SAN DIEGO -- By now, everyone's heard of Apache Hadoop. Created by Doug Cutting during his tenure at Yahoo and named after his son's stuffed elephant, Hadoop is a library of open source software used to create a distributed computing environment. Today, it is touted as one of the newer -- and perhaps one of the best -- technologies designed to extract value out of "big data."

But as Hadoop becomes a household name, it is also taking on a certain mythological form. Philip Russom, industry analyst and director of research for The Data Warehousing Institute (TDWI) in Renton, Wash., wants to bust that thinking wide open. At last week's TDWI Solution Summit titled "Big Data Analytics for Real-Time Business Advantage," Russom presented 12 facts about Hadoop in the hopes of dispelling some of the common myths circulating throughout the industry.

Fact 1: Hadoop consists of multiple products. People may talk about Hadoop as if it's this enormous, singular thing, but it's actually made up of multiple products, Russom said.

"Hadoop is the brand name of a family of open source products," Russom said. "Those products are incubated and administered by Apache software."

When people typically think of Hadoop, they think of its Hadoop Distributed File System, or HDFS, which Russom calls a foundation to layer over other products -- like MapReduce.

Fact 2: Apache Hadoop is open source but it's available from proprietary vendors, too. While the software is open source and can be downloaded for free, vendors like IBM, Cloudera and EMC Greenplum have also made Hadoop available through special distribution, Russom said.

Those distributions tend to come with added features such as administrative tools not offered by Apache Hadoop as well as support and maintenance. Some may scoff at that: Why pay for support when the open source community is free? But Russom said the distributions are making HDFS more powerful for businesses with established IT departments.

Fact 3: Hadoop is an ecosystem, not a single product. The products, which help extend the technology, are being developed by the open source market as well as vendors. Specifically, Russom notes that vendors are providing new products to help make Hadoop look more relational and structured.

"We have this long history of having reporting platforms or data integration

platforms and providing interfaces to the newest platforms," Russom said. "We're seeing a similar thing right now with Hadoop."

Fact 4: HDFS is a file system, not a database management system. That lapse in semantics is one of Russom's biggest pet peeves. While it can manage collections of data, certain database management system attributes are absent in Hadoop.

"Like the ability to randomly access data thanks to query indexes," he said. "We expect structure, which is typically missing from the kind of data types Hadoop deals with."

Fact 5: Hive resembles SQL, but it's not standard SQL. Russom said that fact can be a little unnerving because businesses -- and the tools they use to access data -- tend to be SQL-based. Instead, Hadoop uses Apache Hive and HiveQL, a SQL-like language.

"I've heard people say, 'It's so easy to learn Hive. Just learn Hive,' " Russom said. "But that doesn't solve the real problem of compatibility with SQL-based tools."

Russom believes the compatibility issue is a short-term problem, but one that acts as a barrier to mainstreaming Hadoop.

Fact 6: Hadoop and MapReduce are related, but they don't require

each other. MapReduce was developed by Google before HDFS existed, Russom said. Plus, he added, some vendors such as MapR are peddling variations of MapReduce that do not need HDFS.

Russom, though, considers the duo a good combination. Most of the value in HDFS, he said, lies in the tools that can be layered over the distributed file system.

Fact 7: MapReduce provides control for analytics, not analytics per se. MapReduce is a general-purpose execution engine, Russom said. It's conducive to big data analytics because it can take hand-coded data, automatically process it in parallel and then map the results into a single set. But MapReduce doesn't actually do the analytics itself.

"This is basic MPP [massively parallel processing] architecture but generalized so that you can throw any code at it imaginable and it just has this talent of making it parallel," Russom said. "That's very powerful."

Fact 8: Hadoop is about data diversity, not just data volume. Some have pigeonholed Hadoop as technology designed for high volumes of data, but Hadoop's real value is in the way it can handle diverse data, Russom said.

"That can include the stuff most of our data warehouses were not designed to handle," he said. "Things like semi-structured and fully nonstructured data."

Fact 9: Hadoop complements a data warehouse; it's rarely a replacement. Managing diverse data types has induced comments that data warehouses are dying, but Russom cautions against these sweeping statements.

"How often do people replace things in IT?" he asked. "Almost never."

Data warehouses still do the work they were built to do well, he said, and Hadoop will complement the data warehouse by becoming "an edge system."

"We see data warehouses and architecture getting more and more distributed with more and more pieces added to it," he said.

Fact 10: Hadoop enables many types of analytics, not just Web analytics. Hadoop is sometimes seen as technology for Internet giants, which raises the question of whether it will go mainstream. Russom believes it will partly because it can handle broader analytics.

Railroad companies are, for example, using sensors to detect unusually high temperatures on rail cars, which can signal an impeding failure, said Russom, who also cited additional examples from the robotics and retail industries.

Although he sees a promising future for Hadoop, Russom said mainstream adoption will take years.

Fact 11: Big data does not require Hadoop. The two have become synonymous, but Russom said Hadoop isn't the only answer. Specifically, he

mentioned products from Teradata, Sybase IQ (now owned by SAP) and Vertica (now owned by Hewlett-Packard).

Plus, some companies have been working with big data longer than Hadoop has been in existence -- for example, the telecommunications industry with its call detail records, Russom said.

Fact 12: Hadoop is not free. While the software is open source, the cost for deploying Hadoop is not. Russom said the lack of features such as administrative tools and support can create additional costs. But it also lacks an optimizer and will require professionals -- who make upward of $200,000 -- to hand code within the environment.

That doesn't include the hardware costs of a Hadoop cluster or the real estate and the power it takes to make that cluster operational.

"Don't go thinking Hadoop is free or even cheap," he said. "There are a lot of costs that go with it."

SearchBusinessAnalytics

HADOOP HELPS BRING BIG DATA INTO A DATA WAREHOUSE ENVIRONMENT

In many organizations, the growing volume and increasing complexity of data are straining performance and highlighting the limits of the traditional data warehouse. IT and data management professionals can respond by tweaking and tuning existing system implementations, but the rush to incorporate a variety of unstructured information into the data warehouse environment may call for new technologies that help power big data analytics applications.

In particular, Hadoop systems and related big data technologies are popping up alongside data warehouses to manage the flow of unstructured and semi-structured data, including Web server and other system and network log files, text data, sensor readings and social network activity logs. Hadoop and cohorts such as MapReduce and NoSQL databases can complement data warehouse systems in such cases, creating what analysts describe as a logical or hybrid data warehouse architecture that puts processing workloads on the platforms best able to handle them.

The available building blocks for high-performance business

intelligence (BI) and data warehouse environments also include a selection of other technologies that can fill specific roles -- for example, data warehouse appliances, columnar databases and in-memory databases. Used together, the various tools can boost warehousing speed, but they also challenge an organization's data architecture and integration skills.

"Architecture is becoming increasingly important. You thread everything together with architecture," said William McKnight, president of McKnight Consulting Group in Plano, Texas. Companies need to think about pushing some data warehouse workloads out to technologies that can better handle them, especially when unstructured or semi-structured data is involved, he said.

Most companies do see a need for warehousing speed, according to a report published by The Data Warehousing Institute (TDWI) in October 2012. Sixty-six percent of 278 IT professionals, business users and consultants surveyed by TDWI said getting high levels of performance from data warehouses and related platforms was "extremely important." Only 6% said performance wasn't a pressing issue for them.

Topping the list of processes that respondents thought would benefit the most from high-performance data warehousing were advanced analytics, cited

by 62% of the people surveyed, and the use of big data for analytics, chosen by 40%.

GOING TO EXTREMES WITH HADOOP, BIG DATA

Report author Philip Russom, research director for data management at TDWI in Renton, Wash., wrote that Hadoop has come into prominence in no small part due to its ability to manage and process the extremes of big data. Massively parallel Hadoop clusters can scale out to meet the demands of ever-larger workloads, Russom said, adding that what he described as Hadoop's "data-type-agnostic file system" makes it a better fit for unstructured and semi-structured data than relational databases are.

Yet Hadoop big data systems should be viewed as part of a larger picture, he asserted. "Hadoop is a wonderful complement to the data warehouse, but no one that has worked their way through it would see it as a replacement for the data warehouse," Russom wrote in the report.

"Until Hadoop came along there really wasn't a good way to handle unstructured data," said Wayne Eckerson, director of the BI Leadership Research unit at TechTarget Inc., the Newton, Mass., parent company of SearchDataManagement.com. Organizations had to use text mining tools to parse the data into

rows and columns and then load that into fields in a data warehouse, Eckerson said. But, he added, "it was a two-step process, and a lot of people just didn't use it."

Hadoop, MapReduce and related tools enable developers to automate the data parsing process, according to Eckerson and other consultants. In addition, a variety of Hadoop, data warehouse and data integration vendors have released software connectors that make it easier to transfer data between Hadoop and data warehouse systems.

Ben Harden, a managing director at consultancy CapTech Ventures Inc. in Richmond, Va., sees Web server logs as a good example of data that's best channeled to Hadoop to offload processing from conventional systems and improve the overall performance of a data warehouse environment.

SIDE-BY-SIDE ON BIG DATA

Instead of loading Web logs directly into a data warehouse, they can be stored on a Hadoop system and crunched there, Harden said. Aggregated results then can be fed into a relational model in the data warehouse for analysis by business users, he said. Again, that scenario places upstart Hadoop alongside the venerable data warehouse. "The relational database doesn't go away," Harden

said, adding that the "hardcore processing" of BI and analytics data still has to be done there.

"Everyone is suddenly very log-happy. That's where Hadoop comes in: We need a place to put this stuff, then we have to make sense of it," said Joe Caserta, president of Caserta Concepts LLC, a New York-based data warehouse consulting and training company. He is also co-author -- with BI and data warehousing consultant Ralph Kimball -- of The Data Warehouse ETL Toolkit.

Caserta and other consultants caution that there are still barriers to wider Hadoop use in data warehouse architectures. The open source technology requires advanced programming skills and can benefit from the addition of custom-built tools and functionality, they said. Moreover, Hadoop is a batch-oriented technology that doesn't intrinsically lend itself toreal-time processing of big data. That has led to the use of a variety of advanced messaging and event-oriented technologies to help Hadoop systems keep up with the rapid velocity of data updates, Caserta said.

Overall, though, the pieces are available to extend a data warehouse environment to deal with big data, said Colin White, president and founder of consultancy BI Research in Ashland, Ore. Nowadays, "I don't think it's practical to put everything in the data warehouse," he said. "The key will be to make all

SearchBusinessAnalytics

the different pieces work together."

SearchBusinessAnalytics

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.

SearchBusinessAnalytics

RELATED TECHTARGET WEBSITES

BeyeNETWORK

SearchDataManagement