FEATURE

8 big trends in big data analytics

Big data technologies and practices are moving quickly. Here's what you need to know to stay ahead of the game.

By Robert L. Mitchell

Computerworld |
23 OCTOBER 2014 16:00 IST

Bill Loconzolo, vice president of data engineering at Intuit, jumped into a data lake with both feet. Dean Abbott, chief data scientist at Smarter Remarketer, made a beeline for the cloud. The leading edge of big data and analytics, which includes data lakes for holding vast stores of data in its native format and, of course, cloud computing, is a moving target, both say. And while the technology options are far from mature, waiting simply isn't an option.

"The reality is that the tools are still emerging, and the promise of the [Hadoop] platform is not at the level it needs to be for business to rely on it," says Loconzolo. But the disciplines of big data and analytics are evolving so quickly that businesses need to wade in or risk being left behind. "In the past, emerging technologies might have taken years to mature," he says. "Now people iterate and drive solutions in a matter of months — or weeks." So what are the top emerging technologies and trends that should be on your watch list — or in your test lab? Computerworld asked IT leaders, consultants and industry analysts to weigh in. Here's their list.

1. Big data analytics in the cloud

Hadoop, a framework and set of tools for processing very large data sets, was originally designed to work on clusters of physical machines. That has changed. "Now an increasing number of technologies are available for processing data in the cloud," says Brian Hopkins, an analyst at Forrester Research. Examples include Amazon's

Redshift hosted BI data warehouse, Google's BigQuery data analytics service, IBM's Bluemix cloud platform and Amazon's Kinesis data processing service. "The future state of big data will be a hybrid of on-premises and cloud," he says.

Smarter Remarketer, a provider of SaaS-based retail analytics, segmentation and marketing services, recently moved from an in-house Hadoop and MongoDB database infrastructure to the Amazon Redshift, a cloud-based data warehouse. The Indianapolis-based company collects online and brick-and-mortar retail sales and customer demographic data, as well as real-time behavioral data and then analyzes that information to help retailers create targeted messaging to elicit a desired response on the part of shoppers, in some cases in real time.

Redshift was more cost-effective for Smart Remarketer's data needs, Abbott says, especially since it has extensive reporting capabilities for structured data. And as a hosted offering, it's both scalable and relatively easy to use. "It's cheaper to expand on virtual machines than buy physical machines to manage ourselves," he says.

For its part, Mountain View, Calif.-based Intuit has moved cautiously toward cloud analytics because it needs a secure, stable and auditable environment. For now, the financial software company is keeping everything within its private Intuit Analytics Cloud. "We're partnering with Amazon and Cloudera on how to have a public-private, highly available and secure analytic cloud that can span both worlds, but no one has solved this yet," says Loconzolo. However, a move to the cloud is inevitable for a company like Intuit that sells products that run in the cloud. "It will get to a point where it will be cost-prohibitive to move all of that data to a private cloud," he says.

2. Hadoop: The new enterprise data operating system

Distributed analytic frameworks, such as MapReduce, are evolving into distributed resource managers that are gradually turning Hadoop into a general-purpose data operating system, says Hopkins. With these systems, he says, "you can perform many different data manipulations and analytics operations by plugging them into Hadoop as the distributed file storage system."

What does this mean for the enterprise? As SQL, MapReduce, in-memory, stream processing, graph analytics and other types of workloads are able to run on Hadoop with adequate performance, more businesses will use Hadoop as an enterprise data

hub. “The ability to run many different kinds of [queries and data operations] against data in Hadoop will make it a low-cost, general-purpose place to put data that you want to be able to analyze,” Hopkins says.

Intuit is already building on its Hadoop foundation. “Our strategy is to leverage the Hadoop Distributed File System, which works closely with MapReduce and Hadoop, as a long-term strategy to enable all types of interactions with people and products,” says Loconzolo.

3. Big data lakes

Traditional database theory dictates that you design the data set before entering any data. A data lake, also called an enterprise data lake or enterprise data hub, turns that model on its head, says Chris Curran, principal and chief technologist in PricewaterhouseCoopers’ U.S. advisory practice. “It says we’ll take these data sources and dump them all into a big Hadoop repository, and we won’t try to design a data model beforehand,” he says. Instead, it provides tools for people to analyze the data, along with a high-level definition of what data exists in the lake. “People build the views into the data as they go along. It’s a very incremental, organic model for building a large-scale database,” Curran says. On the downside, the people who use it must be highly skilled.

"People build the views into the data as they go along. It's a very incremental, organic model for building a large-scale database," says PwC's Chris Curran.

As part of its Intuit Analytics Cloud, Intuit has a data lake that includes clickstream user data and enterprise and third-party data, says Loconzolo, but the focus is on "democratizing" the tools surrounding it to enable business people to use it effectively. Loconzolo says one of his concerns with building a data lake in Hadoop is that the platform isn't really enterprise-ready. "We want the capabilities that traditional enterprise databases have had for decades — monitoring access control, encryption, securing the data and tracing the lineage of data from source to destination," he says.

4. More predictive analytics

With big data, analysts have not only more data to work with, but also the processing power to handle large numbers of records with many attributes, Hopkins says. Traditional machine learning uses statistical analysis based on a sample of a total data set. "You now have the ability to do very large numbers of records and very large numbers of attributes per record" and that increases predictability, he says.

The combination of big data and compute power also lets analysts explore new behavioral data throughout the day, such as websites visited or location. Hopkins calls that "sparse data," because to find something of interest you must wade through a lot of data that doesn't matter. "Trying to use traditional machine-learning algorithms against this type of data was computationally impossible. Now we can bring cheap computational power to the problem," he says. "You formulate problems completely differently when speed and memory cease being critical issues," Abbott says. "Now you can find which variables are best analytically by thrusting huge computing resources at the problem. It really is a game changer."

"To enable real-time analysis and predictive modeling out of the same Hadoop core, that's where the interest is for us," says Loconzolo. The problem has been speed, with Hadoop taking up to 20 times longer to get questions answered than did more established technologies. So Intuit is testing Apache Spark, a large-scale data processing engine, and its associated SQL query tool, Spark SQL. "Spark has this fast interactive query as well as graph services and streaming capabilities. It is keeping the data within Hadoop, but giving enough performance to close the gap for us," Loconzolo says.

5. SQL on Hadoop: Faster, better

If you're a smart coder and mathematician, you can drop data in and do an analysis on anything in Hadoop. That's the promise — and the problem, says Mark Beyer, an analyst at Gartner. "I need someone to put it into a format and language structure that I'm familiar with," he says. That's where SQL for Hadoop products come in, although any familiar language could work, says Beyer. Tools that support SQL-like querying let business users who already understand SQL apply similar techniques to that data. SQL on Hadoop "opens the door to Hadoop in the enterprise," Hopkins says, because businesses don't need to make an investment in high-end data scientists and business analysts who can write scripts using Java, JavaScript and Python — something Hadoop users have traditionally needed to do.

These tools are nothing new. Apache Hive has offered a structured a structured, SQL-like query language for Hadoop for some time. But commercial alternatives from Cloudera, Pivotal Software, IBM and other vendors not only offer much higher performance, but also are getting faster all the time. That makes the technology a good fit for "iterative analytics," where an analyst asks one question, receives an

answer, and then asks another one. That type of work has traditionally required building a data warehouse. SQL on Hadoop isn't going to replace data warehouses, at least not anytime soon, says Hopkins, "but it does offer alternatives to more costly software and appliances for certain types of analytics."

6. More, better NoSQL

Alternatives to traditional SQL-based relational databases, called NoSQL (short for "Not Only SQL") databases, are rapidly gaining popularity as tools for use in specific kinds of analytic applications, and that momentum will continue to grow, says Curran. He estimates that there are 15 to 20 open-source NoSQL databases out there, each with its own specialization. For example, a NoSQL product with graph database capability, such as ArangoDB, offers a faster, more direct way to analyze the network of relationships between customers or salespeople than does a relational database.

Open-source SQL databases "have been around for a while, but they're picking up steam because of the kinds of analyses people need," Curran says. One PwC client in an emerging market has placed sensors on store shelving to monitor what products are there, how long customers handle them and how long shoppers stand in front of particular shelves. "These sensors are spewing off streams of data that will grow exponentially," Curran says. "A NoSQL key-value pair database is the place to go for this because it's special-purpose, high-performance and lightweight."

7. Deep learning

Deep learning, a set of machine-learning techniques based on neural networking, is still evolving but shows great potential for solving business problems, says Hopkins. "Deep learning . . . enables computers to recognize items of interest in large quantities of unstructured and binary data, and to deduce relationships without needing specific models or programming instructions," he says.

In one example, a deep learning algorithm that examined data from Wikipedia learned on its own that California and Texas are both states in the U.S. "It doesn't have to be modeled to understand the concept of a state and country, and that's a big difference between older machine learning and emerging deep learning methods," Hopkins says.

“Big data will do things with lots of diverse and unstructured text using advanced analytic techniques like deep learning to help in ways that we only now are beginning to understand,” Hopkins says. For example, it could be used to recognize many different kinds of data, such as the shapes, colors and objects in a video — or even the presence of a cat within images, as a neural network built by Google famously did in 2012. “This notion of cognitive engagement, advanced analytics and the things it implies . . . are an important future trend,” Hopkins says.

8. In-memory analytics

The use of in-memory databases to speed up analytic processing is increasingly popular and highly beneficial in the right setting, says Beyer. In fact, many businesses are already leveraging hybrid transaction/analytical processing (HTAP) — allowing transactions and analytic processing to reside in the same in-memory database.

But there’s a lot of hype around HTAP, and businesses have been overusing it, Beyer says. For systems where the user needs to see the same data in the same way many times during the day — and there’s no significant change in the data — in-memory is a waste of money.

And while you can perform analytics faster with HTAP, all of the transactions must reside within the same database. The problem, says Beyer, is that most analytics efforts today are about putting transactions from many different systems together. “Just putting it all on one database goes back to this disproven belief that if you want to use HTAP for all of your analytics, it requires all of your transactions to be in one place,” he says. “You still have to integrate diverse data.”

Moreover, bringing in an in-memory database means there’s another product to manage, secure, and figure out how to integrate and scale.

For Intuit, the use of Spark has taken away some of the urge to embrace in-memory databases. “If we can solve 70% of our use cases with Spark infrastructure and an in-memory system could solve 100%, we’ll go with the 70% in our analytic cloud,” Loconzolo says. “So we will prototype, see if it’s ready and pause on in-memory systems internally right now.”

Staying one step ahead

With so many emerging trends around big data and analytics, IT organizations need to create conditions that will allow analysts and data scientists to experiment. “You need a way to evaluate, prototype and eventually integrate some of these technologies into the business,” says Curran.

“IT managers and implementers cannot use lack of maturity as an excuse to halt experimentation,” says Beyer. Initially, only a few people — the most skilled analysts and data scientists — need to experiment. Then those advanced users and IT should jointly determine when to deliver new resources to the rest of the organization. And IT shouldn’t necessarily rein in analysts who want to move ahead full-throttle. Rather, Beyer says, IT needs to work with analysts to “put a variable-speed throttle on these new high-powered tools.”

Robert L. Mitchell writes on a wide range of topics, including analytics, emerging technologies, green IT and data centers.

Follow

7 inconvenient truths about the hybrid work trend

SHOP TECH PRODUCTS AT AMAZON