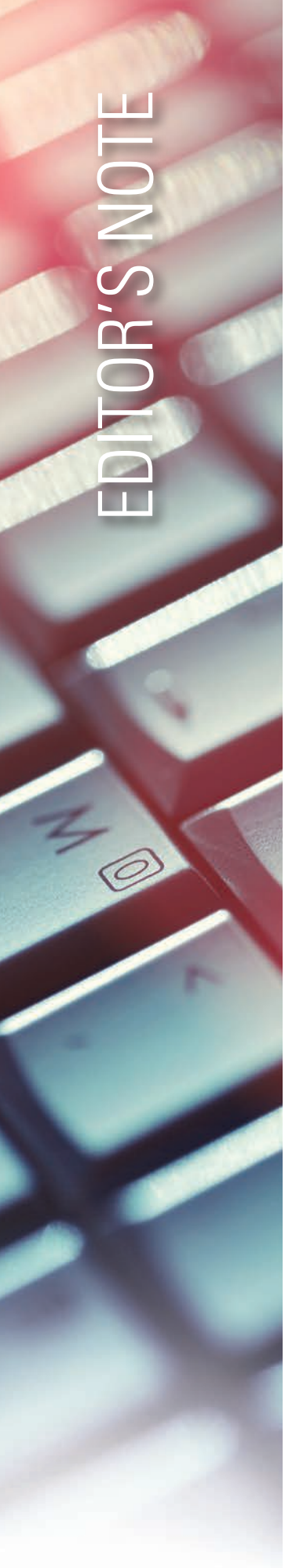
Volume 1 Number 2 ■ SUMMER 2015

BDQ
BIG DATA QUARTERLY

NoSQL, NewSQL, AND THE EMERGING BLENDED ENTERPRISE DATA ENVIRONMENT

12

WWW.DBTA.COM



EDITOR'S NOTE

The Expanding World of Data Management

By Joyce Wells

We are just beginning to understand what big data can mean for organizations across sectors as diverse as healthcare, finance, marketing, and manufacturing, to name just a few. Well-known web-based companies are already far along in their use of cutting-edge technologies to provide fast answers about people and processes, but many other organizations have just started their big data journey.

In this issue of *Big Data Quarterly,* we explore how emerging technologies such as NoSQL, NewSQL, Hadoop, cloud, and virtualization are reshaping the data landscape while co-existing with other more established approaches.

It is clear that changes are coming, and now is the time to become informed. A recent Unisphere Research survey sponsored by Dell Software finds that many organizations are already managing a high volume of data. In fact, close to 30% are managing more than 500TB. Structured data still represents the lion's share of data under management, and relational systems continue to do much of the work. New data technologies are making forays into respondents' data environments, however.

Nearly 10% of the 300 respondents to the survey are now using a NoSQL database, and about one-third say they are currently deploying it, or plan to do so in the next 1–3-plus years. As the pressure to deliver fast-paced data and reporting intensifies, enterprises need to tap into, integrate, and explore the vast array of unstructured data flowing through their systems, observes Joe McKendrick in his cover article. NoSQL and NewSQL are among the technologies that are providing new opportunities.

The most common motivational force for layering in new database systems, according to the Unisphere-Dell survey, is support for new analytical applications, a scenario that augurs well for Hadoop. However, beyond its ability to scale and wield impressive computational power, Hadoop also provides a critical benefit, notes RedPoint Global's George Corugedo in an interview with *Big Data Quarterly.* Because the original data is preserved, it remains useful for analysts and statisticians, as well as for projects that may emerge in the future or change in scope. That's the advantage of the data lake, Corugedo says—the principle that you never destroy the raw data.

'The shift is already happening.'

Big Data Quarterly's cadre of subject matter experts also probe a range of big data topics and trends in this issue. Be sure to check out Michael Corey and Don Sullivan's column on the "virtual infrastructure war," Scott Zoldi's column on the pressing need for big data governance, as well as Jim Scott's column on the Zeta Architecture, "an enterprise architecture built with big data and real time in mind."

And there are many more thought-provoking articles to help get you abreast of the latest big data developments. Whether you believe that the influx of big data and related technologies will dramatically change every industry or think it is just hype is practically irrelevant at this point, says Trifacta's Tye Rattenbury in his article about the emergence of the data preparation market. "The shift is already happening."

Leveraging the Magic of Hadoop

George Corugedo, CTO & Co-Founder, RedPoint Global

RedPoint Global was founded in 2006 by Dale Renner, Lewis Clemmens, and George Corugedo, who previously had worked together at Accenture. Based in Wellesley, Mass., RedPoint collaborates with clients around the world in 11 different verticals. "We have always been very focused on the data, and recognize that a lot of business problems live and die by the quality of the data," says Corugedo.

'That is the magic of the data lake ... that you never destroy the raw data.'

What does big data mean from a business perspective?

Big data, unfortunately, is a term that has been so overhyped that it is kind of meaningless. I think what it is trying to capture is that there are opportunities to know your customers better, to really refine the way you talk to people, the way you understand and predict behaviors, whether it is manufacturing a hard drive or whether it is chasing terrorists or trying to predict the financial markets. Analytics traditionally could only go so far because the data storage was expensive, and the analytics tools and the computational power to derive insights were limited. Typically, what an analyst had to do in the past was take a sample of the data, then build a probabilistic model, and push it out to everything.

What has changed?

Data storage is now much cheaper and much more cost-effective. And the computational power of something like Hadoop is remarkable. Now, instead of taking a sample of the data, you can take all of the data from all of these sources that wouldn't necessarily fit nicely into a structured database. You can bring all that together and have a much higher resolution on a behavior, a moment, a context of when that behavior is going to happen—with more accuracy about that context and that moment than we ever could before. That is what big data provides—that opportunity. To take advantage of it, we think you need three core elements. You have to be able to manage the data; you have to be able to derive insights in the data in the appropriate manner that solves the problem; and you have to be able to act on it. Those three capabilities are what we think of as the big data monetization cycle.

What does RedPoint provide?

RedPoint provides the RedPoint Data Management application. It works seamlessly across complex and hybrid data environments to do a number of very important tasks prior to being able to do accurate analysis. First, it captures data from every source—any source format, structure, lack of structure, any cadence, any web service flavor. If it is in any way exposed or available, we can capture that data. And second, we bring all that data together and integrate it but the way we integrate it and the reason we are so good at integrating it is because of the data quality. We take the data, we clean it, correct it, format it, and standardize it. Then, we deduplicate the data and that is really important because you don't want to have repetitive data being counted in your reporting.

How?

Using probabilistic and heuristic matching techniques, our software allows you to combine records and have a complete view of a customer. We not only deduplicate it but also stitch together identities across different sources. An obvious use is that, as a marketer, you want to know every time a person contacts a brand, whether it is on a mobile application, via telephone or email. There are lots of places where this is really important and we do all of that across all traditional databases—and we can now do it entirely in Hadoop, which opens up another world of possibilities when you think of data lakes and the flexibility of Hadoop in terms of formats of data. One of the things that makes us really unique is that we have a very

native connection to Hadoop, while other companies have to go through Hive, then Tez, then Yarn, then HDFS. We bypass all that and talk natively to Yarn. What that means is that we have a level of control over what the core HDFS is doing.

Why is that important?

First, that means that all of our functionality can work entirely the same way either outside of Hadoop on a traditional database, or inside of Hadoop. No. 2, in one implementation, you can work 100% client-server, 100% Hadoop, or in a hybrid process. To us, Hadoop just becomes a scaling and computational platform. No. 3, our application has been in the market for many years and what we have been able to do is optimize the way we process data. Because of our ability to work in a native way through Yarn we can take that optimization and actually push it into HDFS. And then a final piece is that one of the biggest challenges and probably the single most limiting factor in the global adoption of a technology such as Hadoop is the skill set needed to operate Hadoop. Because our environment is entirely drag-and-drop, we eliminate the need for code-based solutions.

How do you view the data lake?

We are a big fan of the Hadoop data lake because if you think about where an enterprise data warehouse would go awry in the past, it is that either the IT department that was building the data warehouse didn't talk to the business users, or if they did, by the time the data warehouse was built, the business users' requirements had changed. Most enterprise data warehouses go through a process where they aggregate and summarize data, and the original data is destroyed and so it is very hard to realign that enterprise data warehouse with the business users' needs. MDM projects also suffer from the same problems that enterprise data warehouses suffer from. That is the magic of a data lake—it is this idea, this principle, that you never destroy the raw data.

What does that allow?

That gives you the opportunity to always go back and realign with your business community because at the end of the day, it is the business community that is going to create value—or monetize—your data investments. You can always go back and incrementally adjust, change, tweak and give the business users what they need. The other byproduct of having the granular data is that statisticians that are building predictive models can't and won't use summarized or aggregated data. It does them no good because the behavioral predictors are all in the granular data. So, by keeping that data around you are empowering your statisticians to build predictive models that are very accurate and they can always go back and mix and match data in different ways. We are big believers in the data lake approach and being able to provide that for analysts and statisticians—but there is a catch.

What is it?

That catch is that whether you are an analyst, a statistician, or a regular user of the data—that data has to be clean and it has to be deduplicated. Where we fit into that picture is that we can actually turn your data lake into your data quality and master data management platform. Hadoop is an enormously powerful computational platform and if you can make that data lake your central point of data ingestion and do all of your data quality and all of your master data management in there, you are going to be able to do a number of things.

'Latency is the killer of accuracy in modeling.'

Such as?

You will have an inexpensive platform that can scale to give you insane performance in terms of this computationally intensive process of doing data quality, deduplication, and master data management of your data. If you bring all that processing into Hadoop, you will save money, increase the performance dramatically, and then be able to feed minimal amounts of data to other databases as they need it and enable immediate access to data for modeling and predicting different types of behaviors. The importance there is that latency is the killer of accuracy in modeling.

If you could give one piece of advice to someone getting started with big data, what would it be?

Take a specific use case and use the technology to solve that use case and then once that is done and successful, take two more use cases and solve them because the question that everyone is going to face when they are looking for funding for a big data project is: What business problem are you solving for me and how much is it worth? That is going to determine whether you get the funding for it or not. There is magic in the data if you are ambitious about it, but you have to define it and show proof that you solved it. If you build the blocks that way, the argument is inescapable that you should go forward.

Interview conducted, condensed, and edited by Joyce Wells.