Volume 5 Number 1 ■ SPRING 2019

BDQ
BIG DATA QUARTERLY

THE CHANGING ROLE of the DBA in the NEW CLOUD WORLD

8

WWW.DBTA.COM

USE CODE **BDQ19** TO **SAVE $100 OFF** THE COST OF YOUR CONFERENCE PASS TODAY!

DATA SUMMIT

UNLEASH THE POWER OF YOUR DATA

FEATURING THESE SPECIAL EVENTS

- Cognitive Computing & AI Summit
- DATA LAKE BOOT CAMP
- DATAOPS BOOT CAMP

MAY 21–22, 2019

PRECONFERENCE WORKSHOPS
MONDAY, MAY 20

HOT TOPICS INCLUDE:

- Becoming an Insights-Driven Enterprise
- Moving to a Modern Data Architecture
- Unlocking the Power of Data Science
- Adopting a DataOps Strategy
- Building a Data Lake for the Enterprise
- Taking Advantage of Machine Learning
- Navigating the Cloud Landscape
- Enabling Real-Time Analytics
- Modernizing Security and Governance
- Tapping Into the Internet of Things
- Future-Proofing Data Warehousing
- Supercharging Customer Experiences

TUESDAY **MAY 21**

9:00 a.m. – 9:45 a.m.

WELCOME & KEYNOTE

Big Data, Technological Disruption, and the 800-Pound Gorilla in the Corner

Michael Stonebraker, Adjunct Professor, MIT, & Co-Founder/CTO, Tamr

9:45 a.m. – 10:00 a.m.

SPONSORED KEYNOTE

ORACLE

WEDNESDAY **MAY 22**

8:45 a.m. – 9:30 a.m.

OPENING KEYNOTE

Digital Transformation Is Business Transformation: How to Incorporate AI Technology Into a 130-Year-Old Company

Michelle L. Gregory, SVP, Data Science, Elsevier

4:00 p.m. – 5:00 p.m.

CLOSING KEYNOTE

Bring It Home: How to Advance Your Analytics Strategies

John O'Brien, Principal Advisor & Chief Researcher, Radiant Advisors

HYATT REGENCY BOSTON

BOSTON, MA

dbta.com/datasummit

DIAMOND RESEARCH & KEYNOTE SPONSOR: Pythian love your data

DIAMOND SPONSOR: ORACLE

PLATINUM SPONSORS: Actian activate your data; ATTUNITY; CAMBRIDGE SEMANTICS THE SMART DATA COMPANY; CloverDX; DATA KITCHEN; Quest; SQREAM; VERTICA

ASSOCIATION SPONSORS: COGNITIVE COMPUTING; IOUG

BROUGHT TO YOU BY: database TRENDS AND APPLICATIONS; BDQ BIG DATA QUARTERLY

MEDIA SPONSORS: BESTSEOs; BDQ; CRM; CrowdReviews Buyers Guide Based On Client Reviews; database TRENDS AND APPLICATIONS; DATAFLOQ; DMF; KMWorld; TOPSEOs; Visibility

ORGANIZED AND PRODUCED BY: Information Today, Inc.

CONNECT: #DataSummit

BDQ
BIG DATA QUARTERLY

PUBLISHED BY Unisphere Media—a Division of Information Today, Inc.

EDITORIAL & SALES OFFICE 121 Chanlon Road, New Providence, NJ 07974

CORPORATE HEADQUARTERS 143 Old Marlton Pike, Medford, NJ 08055

BIG DATA QUARTERLY (ISBN: 2376-7383) is published quarterly (Spring, Summer, Fall, and Winter) by Unisphere Media, a division of Information Today, Inc.

POSTMASTER
Send all address changes to:
Big Data Quarterly, 143 Old Marlton Pike, Medford, NJ 08055

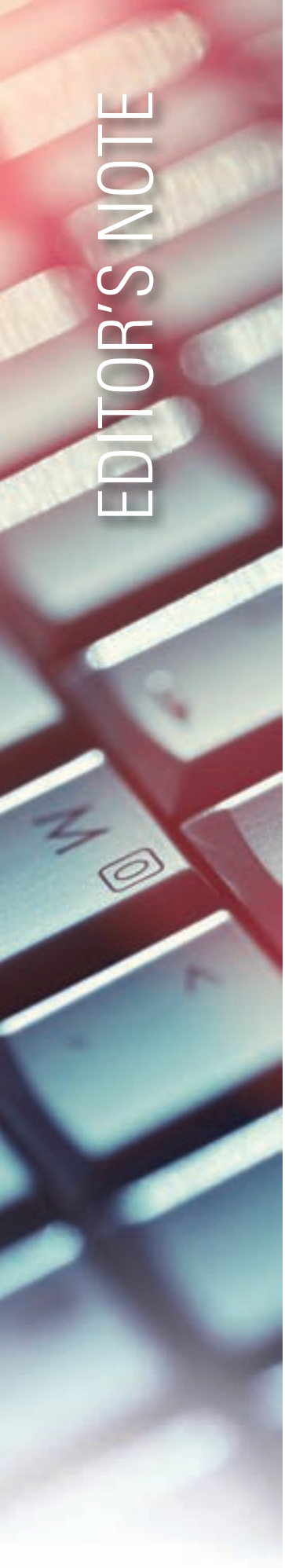
Big Data Quarterly is a resource for IT managers and professionals providing information on the enterprise and technology issues surrounding the 'big data' phenomenon and the need to better manage and extract value from large quantities of structured, unstructured and semi-structured data. *Big Data Quarterly* provides in-depth articles on the expanding range of NewSQL, NoSQL, Hadoop, and private/public/hybrid cloud technologies, as well as new capabilities for traditional data management systems. Articles cover business- and technology-related topics, including business intelligence and advanced analytics, data security and governance, data integration, data quality and master data management, social media analytics, and data warehousing.

SUBSCRIPTION INFORMATION
Subscriptions to *Big Data Quarterly* are available at the following rates (per year):
Subscribers in the U.S. —$97.95; Single issue price: $25

Information Today, Inc.

BIG DATA QUARTERLY
SPRING 2019

CONTENTS



EDITOR'S NOTE

Big Data Trends in 2019

By Joyce Wells

The inexorable forces of cloud and automation are increasingly impacting data management. Already, one-fourth of corporate data is being maintained by cloud providers, and data managers expect to move as much of their data environments to the cloud as they can, as soon as possible, according to the "2019 IOUG Databases in the Cloud Survey" report, produced by Unisphere Research, a division of Information Today, Inc. and sponsored by Amazon Web Services.

The most significant trend in databases over the next 3 years will be the transition from on-premise systems to hybrid environments, according to Unisphere Research analyst Joe McKendrick. "At the same time, close to half of new database projects are going to public cloud providers. While most database projects are still on-premise, these will soon be in the minority."

In this issue of *Big Data Quarterly*, McKendrick looks at the trend toward cloud deployments from another angle: the impact that the shift to cloud is having on DBAs. The transition to cloud data environments and more automation is altering DBAs' roles from "hands-on database overseers" to "value-drivers" for their organizations, notes McKendrick, who sees an unprecedented opportunity for DBAs in terms of career advancement. "The growing role of cloud-based data and databases frees up DBAs from more mundane, day-to-day tasks, providing more time to work more directly with business leaders, managers, and customers on ways data can be applied to business problems," he adds.

Storing and accessing data in the cloud is not the only mega trend shaping data management today. Increasingly, data security is under attack from a range of predators, and the penalties posed by a growing array of regulations have never been greater. In this issue, LicenseFortress' Michael Corey and VMware's Don Sullivan look at the breach involving the Starwood guest reservation database—known widely as the Marriott breach, though it occurred years before Marriott acquired Starwood. Corey and Sullivan weigh in on the current state of data security and suggest it is time to ask this question: "Why is there not legislation similar to the Sarbanes-Oxley Act established as a set of federal requirements for the U.S.?"

And, there are many additional articles on new directions in data management.

Elsevier's Jabe Wilson explains how bias can ruin an AI project and lead to damaging conclusions; MapR's Jim Scott describes the impact of Kubeflow on making machine learning workflows easier to build; and SAS's Anne Buff considers the value a data lake can add to an enterprise data architecture. Also, in a Q&A interview with Kelly Stirman, the Dremio executive highlights the virtues of data as a service for access to data no matter where it resides.

Read on for more articles by our subject matter experts who reflect on the technologies and trends affecting enterprise data management.

We will continue exploring these topics and more at Data Summit 2019, which is returning to the Hyatt Regency Boston, May 21–22, 2019, with preconference workshops on May 20.

Go to www.dbta.com/DataSummit/2019 for more information.

BIG DATA BRIEFING

Key news on big data product launches, partnerships, and acquisitions

MongoDB has released the beta offering of its native data visualization tool, **MongoDB Charts**, on its cloud database-as-a-service, MongoDB Atlas, which runs on Amazon Web Services, Microsoft Azure, and Google Cloud Platform. www.mongodb.com

MapR Technologies, provider of a data platform for AI and analytics, has announced support for Apache Drill 1.15. The release offers new enhancements to conduct queries on complex nested data structures, including files, MapR JSON database tables, and cloud data sources specifically for S3 (Amazon Simple Storage Service). https://mapr.com

Attunity, a provider of data integration and big data management software solutions, has announced two new solutions: Attunity for Data Lakes on Microsoft Azure, designed to automate streaming data pipelines, and Attunity Compose for Microsoft Azure SQL Data Warehouse, designed to enable data warehouse automation for Azure SQL Data Warehouse. www.attunity.com

Collibra, a provider of enterprise data governance and catalog software, has announced $100 million in series E funding led by CapitalG at a post-money valuation of more than $1 billion. New funding brings the company's total venture funding to $233 million. www.collibra.com

Birst, an Infor company and a provider of cloud business intelligence and analytics for the enterprise, is introducing its Smart Analytics family of solutions, powered by Infor Coleman AI. **Birst Smart Analytics** is a new set of AI-enabled capabilities that elevates organizations above traditional reports and dashboards, using machine learning algorithms to power intelligent insights not previously available to business users. www.birst.com

At IBM Think 2019 in San Francisco, IBM unveiled enhancements to **IBM Cloud Private** to deliver integrated platform management and orchestration capabilities to help enable a secure private cloud by running the entire private cloud infrastructure on IBM Z. www.ibm.com

Oracle has added new features to its blockchain platform to help users speed up the development, integration, and deployment of new applications. According to the company, while blockchain can streamline many existing processes surrounding supply chain, identity, cross-border payments, and fraud detection, businesses are struggling to implement blockchain networks within their existing ecosystems. www.oracle.com

Quest Software, a global systems management, data protection, and security software provider, has joined the Veeam Alliance Partner Program. Additionally, Quest QoreStor, the company's software-defined secondary storage platform with deduplication technology, has been verified as a Veeam Ready Repository. www.quest.com

Cisco has announced a new architecture to extend the data center to wherever that data lives and everywhere applications are deployed. To support this "data center anywhere" initiative, Cisco is introducing a range of innovations across networking, hyperconvergence, security, and automation. www.cisco.com

Google Cloud has announced the general availability of Cloud Firestore, a serverless, NoSQL document database, which is available in 10 new locations to complement the existing three, with a significant price reduction for regional instances, and that enables integration with Stackdriver for monitoring. https://cloud.google.com

Alfresco, an open source provider of process automation, content management, and information governance software, is offering an updated version of its Application Development Framework. ADF 3.0 extends Alfresco's content market leadership with significant performance enhancements. www.alfresco.com

MariaDB is releasing MariaDB Platform X3, an open source database that unites transactional and analytical workloads at scale, and introducing a new MariaDB Managed Service supporting public and hybrid cloud deployments. www.mariadb.com

SolarWinds, a provider of IT management software, has rolled out the SolarWinds Flow Tool Bundle, a set of free analysis tools designed for greater visibility into network infrastructure. www.solarwinds.com

Kelly Stirman, VP Strategy
Dremio

Making Data Accessible

For many companies, designing and implementing a data platform for analytics is a critical task. With data growing internally at a rapid pace, along with the challenges of mergers and acquisitions adding new systems and data silos, accessing data for exploration and insight can be a problem.

Recently, Kelly Stirman, vice president, strategy, of Dremio, a VC-backed firm which emerged from stealth in 2017, discussed how using open source projects, open standards, and cloud services, companies can deliver data as a service (DaaS) to their data consumers across critical lines of business. By combining capabilities and technologies into a solution that enables access, transformation, security, and governance, Stirman contends, DaaS represents a new approach to vexing analytics challenges, delivering data at scale with high performance.

Where is data as a service most useful?

It is targeted at data consumers—people who are dependent on data to do their jobs effectively. These are analysts, data scientists, users of BI tools and—if you think about it—that is a lot of people. Most days, if you have a question about the world around you—if you want to find a restaurant, or know what the weather is going to be over the next few days, you can do that instantly, but when you are at work and you have basic questions, there is no simple way to answer them. Instead, it can take weeks and months to get an answer. It is completely different from your experience in your personal life.

What does it enable?

Data as a service is an underlying movement to change people's relationship to data and enable data consumers to get what they need just as easily as they can in their personal lives. They are not beholden to IT and they are not waiting for their turn to get IT to do something on their behalf.

What type of organization is it targeted at?

I think it applies to every company, no matter how big or small. It would be hard to find any company that would say: "Oh, we have plenty of folks in IT. IT is very responsive." Everyone is short-staffed and waiting on IT to help them get their jobs done. One of the big trends over the past decade has been how lines of business have come to own the bulk of the budget because companies have realized that if IT owns the budget, things just never get done. So, data as a service applies to virtually all companies. Even for small companies, as they get larger, the proliferation of data in different systems and different fiefdoms across the business grows.

Is it necessary to move data to take advantage of data as a service?

The answer to the problem of having data in different systems has been to try to just copy it all into a new system. But then you create a new silo and you never finish consolidating all the data into this new system—and that is the history of data warehouses, data lakes, and even the cloud. You are never going to have all of your data in one place. If your data is in the data warehouse or the data lake or one of your ERP systems, a supply chain system, or a cloud service, it should not matter. You should be able to access and work with the data wherever it is, no matter what the underlying technology is, and—no matter how big it is—data as a service needs to make it work the way it already is.

What level of skills or expertise is necessary?

Skills are on the order of Office 365 or Google Docs: Most people who use data to do their jobs have some sort of tool like a Tableau or a Power BI, or, if you are a data scientist, something like Python or Jupyter Notebook. Those are tools that people like and are good at using, and that is ultimately how they take data and visualize it or plug data into a predictive model or recommendation system. That part of the equation is not broken. It is getting the data into those tools and making it accessible—searching and finding datasets, and blending or curating data from different sources together to get the data that you need to perform your analysis—that is the really hard problem, and that is what Dremio is focused on.

How is it accessed?

The experience of someone using Dremio is logging in through a browser, doing a search, finding different datasets, and then clicking around, sharing data with other people. There is no coding involved, and if you can use something like Office 365 or Google Docs, you can use Dremio. Once you find or build the dataset that you need, you click a button and launch your favorite tool and it connects to that dataset.

Data as a service sounds similar to data virtualization.

There are some things that feel and sound similar, but data virtualization was never a tool for the data consumer. First, data virtualization was always something that required programming and complex APIs and kind of low-level machinery designed for an IT user. Data as a service in contrast is focused on the data consumer. It is about removing the complexity—streamlining access, making work collaborative, and using existing tools. The second big difference is that data virtualization never really solved the problem that a lot of companies have today which is centered around performance. It was really designed for smaller datasets. Part of the underlying challenge for any company dealing with data is making it perform at a speed that lets people do their jobs.

Why is open source critical to data as a service?

It is a reflection of modern IT and data infrastructure that newer projects are predominantly based on open source technology. In our experience, companies expect mission-critical infrastructure to be open source. That allows us to build on efforts that span hundreds of companies that are collaborating on the development of these core building blocks, and to also benefit from that to bring a more robust product to market more quickly.

There is some very key open source technology that is part of the Dremio stack. The most important is Apache Arrow, a product we helped start 3 years ago that is the standard for how people today do analytics. It has become the cornerstone for how newer systems are built.

Apache Parquet is another, and, if you look back over the past 10 years or so as companies have embraced data lakes, overwhelmingly the standard that they have used to store their data for analytics is Parquet. We build on that in our own product, so that if you already have your data in Parquet, Dremio is the fastest way to access that data.

What else?

There are a number of other open source projects that we build on in our product, but actually Dremio itself is open source because we want it to be something everyone can access whether you are a small startup or a Fortune 10 company. We monetize Dremio by selling subscriptions that include an enterprise version with advanced features in security and maintenance beyond what is available in open source. But the open source version of Dremio is in use by thousands of companies and in 100-plus countries.

Where does privacy and security fit into the platform?

Every company is focused on security, and, if they are not, they should be. However, one of the fundamental causes of security vulnerability is this friction between IT and legacy systems. People are going to find a way to get their jobs done. If you make the process of accessing data too hard, too cumbersome, or too slow, they will take matters into their own hands by moving data into Dropbox, spreadsheets, emailing information to other people, and moving out of any kind of controlled environment. One of the philosophies of data as a service is that you attract bees with honey. We integrate with existing security standards. Data as a service is about keeping data access in a controlled, protected environment by letting people be more productive and removing the temptation for people to take matters into their own hands.

Looking to the future, what are the challenges you see on the horizon?

There is an interesting transition coming up as organizations move part, or all, of their infrastructure to the cloud. Whether you are 100% in the cloud or you have a mix of things in your data center and in the cloud, security is one of the challenges, but data access in general is an issue because the services that cloud providers make available to their customers are different than what people are using in their own data center. That should not matter and it should not matter what the underlying technology is. And data consumers should not have to care that over the past 3 months their company has moved pieces of their data center to the cloud. As companies make the critical transition to the cloud, data as a service lets them do that in a much more seamless way to their data consumers.

This interview was edited and condensed by Joyce Wells.

There's a Ghost in the Machine—Our Own Biases

By Jabe Wilson

AI is not a panacea, and it has the potential to make huge miscalculations if given incorrect or biased data.

THIS DECADE HAS seen AI take on a somewhat similar role to that of the internet in the 1990s. Everyone knows that it's set to become ubiquitous, major organizations are shifting toward it, and yet very few people fully understand it. This is a problem because, as AI becomes increasingly prevalent, it's critical that we do not treat it as simply a magic box to which we can pose any question and it will give us the "right" answer.

Computing pioneer Charles Babbage was once asked, "If you give your computer the wrong data, how does it arrive at the right result?" He responded that he simply couldn't comprehend the level of confusion that would lead someone to ask such a question. As an anecdote, it's a perfect reminder that any computer or AI is only ever as good as two things—the algorithms that underpin it and data it is given to work with. The crucial point is that data is delivered by humans, and algorithms are built by humans, who can introduce all sorts of issues, not the least of which is their own unconscious biases.

As the algorithms we're using become more sophisticated, bias can become harder to spot. Researchers who work with AI have a responsibility to constantly guard against this and keep their work as impartial as possible. This is especially important in industries such as pharmaceuticals, where poor calculations could have huge healthcare ramifications, including slowing down the progress of drugs to market or causing reduced efficacy for new drugs.

When Things Go Wrong

Bias—whether consciously or unconsciously introduced—can ruin any project and lead to extremely damaging conclusions if inaccurate or incomplete data is fed in. We've already seen multiple examples of engineers being unable to prevent AI networks from accidentally becoming racist or sexist, and it's not hard to see how bias can lead to many worrying outcomes.

For example, to take a case of fairly obvious bias, what could happen if a new drug were only ever tested on Caucasian men? Other genders and ethnicities could have entirely different, more negative, reactions. It's hardly a theoretical problem either—in 2014, a study found that 86% of the clinical trial population was white and two-thirds of the population was male. The more we learn about the genetic code, the more we are realizing the safety and efficacy of any given drug will vary greatly depending on the patients' genetic profile. Nor is this simply a theoretical problem. Across the industry, a staggering 93% of patients fail to make it through clinical trials. This rate of dropout means researchers can't gather fresh data each time and instead are relying on historical data, which can be heavily weighted toward particular ethnicities or genders.

Jabe Wilson is consulting director, Text and Data Analytics, at Elsevier's R&D Solutions (www.elsevier.com/rd-solutions).

To make matters even more complicated, there are less obvious sources of bias that can also creep in. For example, if a company is conducting pharmacovigilance by using AI to monitor social media posts for any mention of adverse reactions to a particular drug, it could easily fail to account for cultural differences between different groups or demographics. This could lead to the company to erroneously believe that it is only having side effects for one group of patients and not others.

Removing the Bias?

So how can researchers combat this problem? Unfortunately, it's impossible to fully remove bias from an experiment. No dataset is ever totally comprehensive and so results will always be skewed to a certain degree. On top of that, not all information is created equal. Some datasets are more reliable and objective than others.

One partial solution is to ensure that the widest and most diverse range of data is being used so that bias is minimized. If AI is able to gather and contextualize data from multiple sources, the outputs will inevitably be richer and more comprehensive. However, this is a big data management challenge as firms need to collate and harmonize data from both a number of different internal company data silos as well as external sources, such as published literature or patent data. This mandates the need for robust platforms with advanced computing power and intelligent algorithms built by expert scientists in their field.

In fact, smart researchers will not only look to offset any potential bias in their AI systems but have the systems hunt for it themselves by flagging issues such as a dataset overwhelmingly biased toward white men. This allows researchers to easily spot certain forms of bias and make sure they're accounted for.

The Human Touch

Beyond using the broadest possible datasets, first-class tools, and best practices for data management, another key factor in offsetting bias is the knowledge of experienced researchers. Having experts look over the findings from AI systems can often highlight results that could have occurred due to bias in the data. Scientists working with AI systems must constantly interrogate and sanity-check the answers they're given. For example, in our pharmacovigilance case above, AI might tell researchers that patients in a few select countries are having side effects that aren't being detected elsewhere on social media. Rather than assuming the problem is limited to those countries, they would need to ask why those countries and not others? More importantly, they would need to ask if there are factors involved that the AI algorithm wouldn't pick up, such as cultural differences.

If AI is able to gather and contextualize data from multiple sources, the outputs will inevitably be richer and more comprehensive. However, this is a big data management challenge. …

The problem of bias is only increasing in importance as AI moves into the mainstream. Experts have predicted that, by 2030, AI will add more value globally than the outputs of China and India combined. It has an exceptional capacity to enhance our lives by aiding in scientific discovery. Yet this is only going to be possible if it's used intelligently and in conjunction with expert human researchers.

AI is not a panacea, and it has the potential to make huge miscalculations if given incorrect or biased data. The modern researcher should heed the words of Charles Babbage and remember to constantly question and interrogate the results they are given, no matter how advanced a computer they are working with.

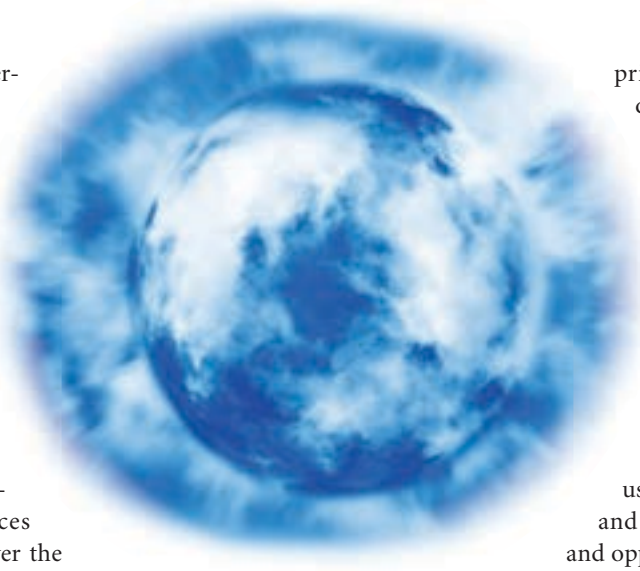
THE CHANGING ROLE of the DBA in the NEW CLOUD WORLD

THE CHANGING ROLE OF THE DBA IN THE NEW CLOUD WORLD

By Joe McKendrick

Cloud computing—and everything that goes with it—is dramatically changing the roles and aspirations of database administrators. No longer do DBAs need to be chained to their databases, wrestling with managing updates, applying security patches, and dealing with capacity issues. Moving to a cloud data environment is steadily shifting DBAs' roles from hands-on database overseers to value-drivers for their businesses—and enabling a range of career advancement opportunities not seen since the dawn of relational databases. ▸

THE CHANGING ROLE OF THE DBA IN THE NEW CLOUD WORLD



Overall, DBAs and their enterprises are embracing cloud computing in a big way. A recent survey conducted by Unisphere Research, a division of Information Today, Inc., in partnership with Amazon Web Services, found that, on average, 25% of organizations' critical enterprise data is now managed in public clouds. In addition, 60% of data managers indicate their use of public cloud-based data resources and platforms has increased over the past year. Close to one-third anticipate growth exceeding 10% over the coming year ("2019 IOUG Databases in the Cloud Survey").

Leading vendors in the database space are increasingly promoting cloud-based approaches that will transfer many DBA functions from on-premise data centers to cloud providers. This is the goal of Oracle, which sees cloud-based "self-driving databases" doing much of the heavy lifting of enterprises, said Steve Daheb, senior vice president of Oracle's Cloud Business Group.

At the same time, DBAs will be busier than ever, engaging in designing and delivering data-driven capabilities to their businesses. "The average DBA spends 90% of their time in maintenance, managing 50 databases each," he said. "They're shifting now to higher-value tasks, from tuning and provisioning to focusing on business analytics. They see it as an opportunity." The rise of cloud-based databases is also helping DBAs guide their organizations into new technology realms, such as blockchain, the Internet of Things (IoT), and AI, he added. "It's all cloud-based. For example, cloud is an enabler for IoT, because if you think about it, where would you store all that data streaming in?" Oracle also sees greater engagement with non-technical business users as well, he said.

> 'The more autonomous databases become, the more events that consume precious DBA time can be rectified automatically.'

Cloud means a number of changes to DBAs' jobs, including the following.

MORE HIGH-VALUE DEPLOYMENTS, FEWER LOST WEEKENDS

The growing importance of cloud-based data and databases frees up DBAs from mundane, day-to-day tasks, providing more time to work directly with business leaders, managers, and customers on ways data can be applied to business problems. Data clouds take out much of the grunt work involved in setting up, operating, and scaling enterprise databases, whether for production, development, or testing purposes. Cloud services offer flexible capacity that can be expanded or reduced on demand and provide automated management of lower-level tasks such as security patching, server provisioning, and backups. DBAs can then focus on solutions and innovations to boost the efforts of their IT teams and business users to apply data analytics, AI, and machine learning to problems and opportunities.

MORE AUTONOMY, LESS HANDS-ON MAINTENANCE

DBAs are seeing their job roles—and perhaps even titles—change. Emerging roles, including that of enterprise data architects, data stewards, and data engineers, provide new career paths for today's data professionals. In addition, the cloud opens the way to DevOps, in which the work of developers and production teams is in sync, delivering quality software at a pace the business requires. This is being made possible through increasing database automation, with almost completely autonomous databases on the horizon. There will be less of a requirement for DBAs to constantly be on the lookout for events—be they minor glitches or patching requirements. The more autonomous databases become, the more events that can be rectified automatically. "The average enterprise gets 17,000 alerts a week, and only 19% are reliable and only 4% are ever investigated," Daheb illustrated. "So you just have all of this noise." A more autonomous database employing AI will catch many of the false positive alerts and enable DBAs to focus on issues that are real and of material importance to the business.

MORE DATA ENRICHMENT, LESS DATABASE

The always-on, data-driven enterprise doesn't function on internal databases alone—competing in today's economy means leveraging a range of data sources beyond the firewall. As a result, DBAs need to focus on the value that the data is bringing to their businesses, versus simply running the relational database environment on a day-to-day basis.

"Data has too long been viewed as something to just be managed by technology or systems—instead we truly need to make data ubiquitous throughout an organization," said Katie Fabiszak, chief marketing officer of Riversand, a provider of master data management and product information management solutions. "For the past decade we have been talking about data being a strategic asset but we haven't really arrived there yet. Making data the responsibility of everyone is the way we can finally deliver on the notion that data is a strategic asset. In and of itself, data is not all that important or compelling. It's when we take great data and make it part of our business processes that allows magic to happen."

Cloud-based data environments enable DBAs to concentrate on bigger questions, Daheb remarked. "How do I architect the database? How do I take advantage of things like blockchain? How do I think about digital interactions or IoT? It becomes much more about what's possible now. You have this infrastructure to make it feasible to run everything from engagement to product development to running financials for the business."

With the increasing reliance on data-driven capabilities such as AI and analytics, "organizations need to truly understand how they need to use data to reach their full potential," Fabiszak added. "For starters, data needs to be viewed as an organization-wide responsibility. The actual data itself must be shared and leveraged to fulfill a particular business purpose. Data can absolutely live up to its promise of being a strategic asset only when companies—and people—stop thinking of it purely as something to manage and instead think of data as being just as essential to an organization as its people and its revenue."

'DBAs can't be DBAs anymore—they must be data professionals.'

MORE GROWTH, FEWER TECHNICAL RESTRAINTS

The need for greater capabilities means heavier emphasis on data, regardless of format or origin. "Modern applications, such as social media, investment, and fantasy gaming apps need five or six-nines availability, along with worldwide accessibility," said Gaurav Yadav, founding engineer and product manager at Hedvig, provider of a distributed storage platform. "They are best suited for entire database systems in the cloud. There are multiple ways to consume databases in the cloud such as virtual machines with databases installed, database schema as a service, or scalable database as a service solutions. These approaches also provide a quick way to bring up database infrastructure instead of building one from the ground-up."

MORE PLATFORM THINKING, LESS STANDALONE ENVIRONMENTS

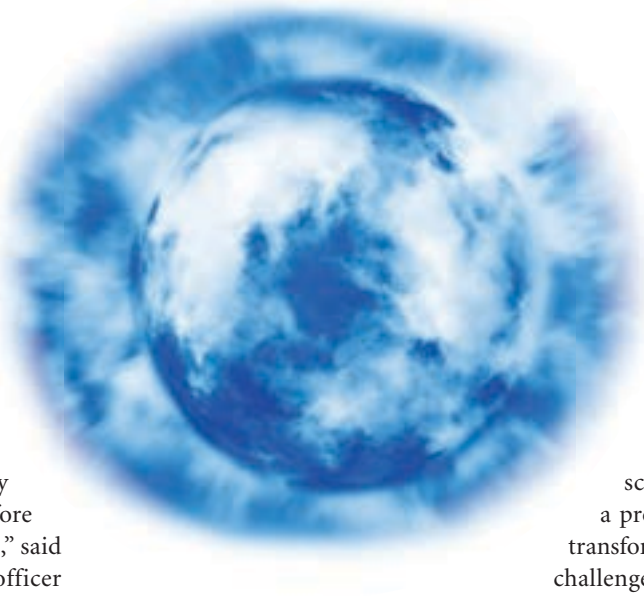
With cloud, today's databases are less likely to be solely servicing limited groups of internal users and are more likely serving wider audiences beyond corporate firewalls. The rise of IoT, for example, means data coming in from devices and systems anywhere across the globe. This requires the ability to securely acquire, replicate, and assure the availability of data beyond the relational database management systems within the local corporate data center. The DBA's role is evolving from a sole, specialized database operator to a maestro capable of coordinating a symphony of data environments.

"Many hybrid data challenges come down to SaaS and PaaS platform abstractions; in traditional database management and administration, the DBA held the keys to the kingdom of the data environment," said Kaus Phaltankar, president and CTO of Caveonix, which provides a risk management platform for the hybrid cloud. "In the cloud, especially with SaaS and PaaS offerings, these super users' privileges are sometimes not present. The cloud environment requires more coding and command-line knowledge, as the game is much more software-oriented. DBAs can't be DBAs anymore—they must be data professionals. It's also important for DBAs to work with DevOps engineers for data environment deployments in the public cloud, and, in many instances, on-premises installations as well. Being able to embrace infrastructure-as-code is becoming more and more important in the ever-changing data center environment."

MORE RELIANCE ON DBAS AS TRANSFORMATION AGENTS, LESS AS LEGACY DATABASE OPERATORS

As enterprises seek to compete in this new era, they are leaning heavily on managers and professionals to help make the move into the cloud realm. This ▶

THE CHANGING ROLE OF THE DBA IN THE NEW CLOUD WORLD



requires DBAs to be knowledgeable about existing systems and application requirements. Business and application teams are looking to their DBA partners to identify valuable data and work with the business to guide them on their cloud journeys.

"Legacy applications and monolithic applications based on a single-tier architecture may require significant refactoring before they can be migrated to the cloud," said Casi Johnson, chief operations officer and innovations leader at M3, a provider of accounting and analytics software for hospitality management. "Any newer applications that can take advantage of the cloud-hosting benefits—scalability, monitoring—or would only require a small amount of modified code can be considered ideal for cloud migrations."

This extends to managing the costs and resource requirements involved in such migrations, said Anupam Singh, general manager of analytics at Cloudera. "Managing cost for operational expenditure of the cloud will become harder than the savings of capital expenditure on the cloud," he said. "Data locality and multiple data stores can lead to performance challenges. A lot of big data benefits were from taking compute to the data. But, as S3 [AWS Simple Storage Service] and ADLS [Azure Data Lake Storage] become popular data stores, the data will have to be moved to compute. Database engines have to be re-designed to work with remote data."

DBAs also will play a greater role in selecting and managing outside cloud service providers. "The main challenge when handling any cloud environment is vendor lock-in," said Hedvig's Yadav. "Extreme care needs to be taken when selecting any cloud provider, because each provider has its own framework to consume resources which makes it harder for consumers to move from one provider to another. A cost analysis is also important, as cloud costs can easily get out of control if the hybrid environment is not designed—because simplicity comes at a cost. A careful budget analysis should be done before deciding on what applications/data should be kept in cloud vs on-premises."

MORE LEADERSHIP FROM DBAS, LESS ORGANIZATIONAL CHAOS

The most prevalent mode of cloud deployments seen these days is "accidental hybrid environments where different groups choose different cloud platforms based on their own criteria and an enterprise ends up supporting a diverse set of clouds," said Ken Rugg, chief product and strategy officer at EnterpriseDB. "Since many companies are still in the experimental phases of cloud adoption, this can actually work well since it gives them experience in many different environments to help inform future choices. However, this creates a new legacy since applications may take advantage of proprietary features of a particular cloud, making it difficult to move in the future." DBAs are increasingly being called to step up and provide clarity and guidance to manage cloud chaos.

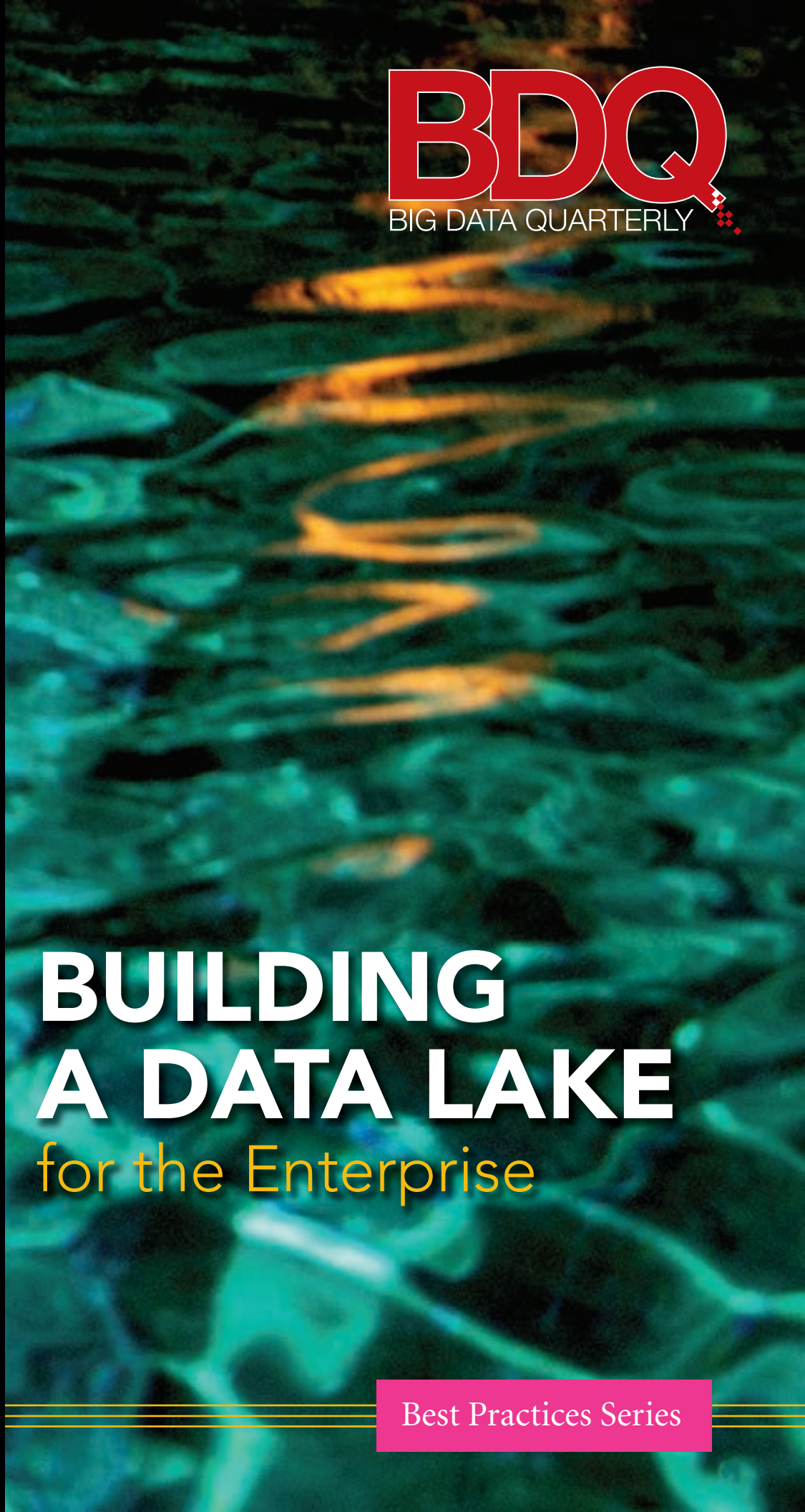
DBAs will also assume roles as educators, as the greatest barriers to the adoption of cloud-based data environments are knowledge and experience. "Adopters need to learn the capabilities of new technologies as well as how they'll fit into existing DevOps practices, monitoring, and automation," said Kyle Clubb, analytics and data science prinicipal at Quisitive, a provider of solutions for digital transformation. "These are behavioral challenges, not technical ones." DBAs will be called upon to address these issues easily by supplying their teams with adequate training and tools.

DBAs are uniquely positioned to bridge the worlds of data management and business, paving the way to successfully developing data-driven enterprises.

Ultimately, the movement to the cloud provides DBAs with two-fold opportunities: first to help guide the business as it makes the shift to cloud, and, ultimately, allowing them to devote more time to the things that matter such as building the business and meeting the needs of customers.

"Data won't be managed by a single person or group of people," said Riversand's Fabiszak. "It is woven throughout an organization and feeds business processes that span the organization. People must become savvier and understand what data is needed to solve specific business problems. No longer will we have siloed data practices and management, but instead we will have shared data disciplines across the entire organization. Almost like crowdsourced data management—we may change the way a data element is described or used based on new inputs or information we learn from each other. At the end of the day, this will result in much more successful efforts to create, consume, and manage data."

BDQ
BIG DATA QUARTERLY



BUILDING A DATA LAKE
for the Enterprise

Best Practices Series

Data executives, managers, and professionals face a lot of pressures these days. Not too long ago, their main concerns were database performance, database security, and integration with on-premise ERP-type applications—concerns mainly confined to the data center.

Now, there's a much greater need to focus on the business side itself and to harness the power of data to play a role in addressing innovation and reaching customers in more profound ways. The information that businesses need still moves much too slowly. Yet, at the same time, there are firehoses of information, and data managers are often saddled with trying to sort out what is and what is not of material importance to their businesses and end users.

> Business requirements need to be the first consideration, above and beyond any technology issues.

What is needed is a way to have the data ready and waiting when and where it is required. That's where data lakes come in. Adoption of these environments—data repositories in which data is ingested and maintained in its original state, without the overhead of cleansing, ETL processes, or building models—has been rising. Overall, 38% of organizations responding to a Unisphere Research survey of 300 database managers are employing data lakes as part of their data architectures, up from 20% in the 2016 survey. Another 15% are currently considering adoption. Data lakes are growing to impressive proportions as well. Close to one-third, 32%, now support more than 100TB of data ("2018 Next-Generation Data Deployment Strategies Report," sponsored by Oracle).

The swing to data lakes is part of the larger evolution of data management from relational to frameworks such as Hadoop and Spark. The goal of these new forms of data architecture is to more efficiently handle the variety and the volume of data surging through enterprises in a faster and

more efficient manner than is possible using the techniques that have been deployed over the past 3 decades. This not only requires a change in architectural approach, but a change in thinking.

The following are steps to adopting a data lake foundation:

Think About the Business—First and Always—and Long-Term Requirements.

As with any major enterprise technology initiative, the first question to ask with data lakes is, "What's in it for the business?" Business requirements need to be the first consideration, above and beyond any technology issues. While the value of the data lake may be seen in the ability to quickly design and answer business queries, the long-term value is even more compelling: maintaining information for which applications have not yet been designed but are on the horizon.

Now, some cloud platforms are seen as more secure than on-premise corporate systems, bringing unlimited capacity and flexibility to ever-expanding data repositories.

Borrow From the Data Warehouse Experience.

While data warehouses have been successful up to this point, their roles have proven to be too limited for today's real-time enterprise. But there are some very important lessons from the data warehouse experience. For starters, don't try to boil the ocean. In other words, don't try to roll out a gigantic data environment and hope members of the enterprise will start figuring out how to use it. Start small, working on specific pain points with selected business units. As others see the success of these initial projects, they will want in as well.

Determine Whether the Data Lake Should Be in the Cloud or Remain On-Premise.

Until recently, amid concerns about cloud data security, it was natural to build data lakes on-premise. Now, some cloud platforms are seen as more secure than on-premise corporate systems, bringing unlimited capacity and flexibility to ever-expanding data repositories. Still, the pros and cons of both on-premise and cloud-based options need to be weighed. "On-premises storage and processing offer tighter control over data security and data privacy, whereas public cloud systems offer highly scalable and elastic storage and computing resources to meet enterprises' need for large-scale processing and data storage without having the overheads of provisioning and maintaining expensive infrastructure," explains Ben Sharma in *Architecting Data Lakes* (O'Reilly Media, 2018). A hybrid environment may see data lakes employed as testing environments for tools and applications. The key for both options is to "put a robust data management structure in place, one that provides complete metadata management, [so] you can enable any combination of on-premises storage, cloud storage, and multicloud storage easily."

Set the Ground Rules for Data Governance and Security.

Data governance has always been an important component of data management, and data lakes are no exception. The same policies and procedures that involve data owners, stewards, and specialists need to be extended to data lakes. With so much data flowing into a lake relatively unfettered, there must be a framework that applies the organization's data and information policies. Such governance is important to creating a common framework of what data is permissible and how it will be used. In addition, it's critical that security be baked into the data lake architecture as well, and it is consistent or even more robust than all other currently available data security policies.

Make Data Discovery Easy.

Establishing a metadata layer and enabling governed data discovery will also be paramount to address data accessibility and how insights, context, and analytics will be delivered to the enterprise. Plus, it's essential to know exactly what data is available with the data lake. Otherwise, data lakes will suffer from the same inertia that hamstrung data silos and proprietary databases. Nobody will know or understand what data resources are available to them. User awareness and acceptance is critical to the success of a data lake.

THE CHALLENGE OF DISPARATE DATA SOURCES

Data lakes help address the greatest challenge for many enterprises today, which is overcoming disparate and siloed data sources, along with the bottlenecks and inertia they create within enterprises. Data lakes bring data into a single enterprise location, in its original state, so as business questions arise, users can access the appropriate data, and run it through their platforms or applications of choice. Ultimately, data lakes help provide greater choice.

—Joe McKendrick

IRI Voracity
An Insatiable Appetite for Data

Stock, Govern & Fish the Data Lake in a Bigger, Better Boat

Whether on-premise or in the cloud, data lakes all present the same challenges: how to most efficiently populate them, govern them for quality and security, and best leverage them for analytic value. The IRI Voracity data management platform—powered by CoSort or Hadoop engines and built on Eclipse—provides a uniquely fast, versatile and affordable environment for doing it all.

STOCK (POPULATE)

Getting the right data into your data lake will affect its storage footprint, downstream analytic potential, and how data owners perceive your intentions.

Thorough profiling processes that involve source and target stakeholders, and tools that stratify, classify, and search through data, you can determine the nature and fitness of data for purpose, and inspire confidence. These activities are also necessary prerequisites for metadata management, data integration, quality, security, and analytic activities.

To replace or mimic a data warehouse, consolidated high-speed ETL operations can populate a data lake with the right raw, and refined, data. Diagrams and metadata from these jobs can also lend structure to what can otherwise turn into a forgotten "data dumping ground."

To speed both discovery and integration, use fast acquisition and transformation engines like those in Voracity: IRI Fast Extract and IRI CoSort or Hadoop which do not tax databases and applications, or require Java steps that legacy ETL tools see crawling or crashing in volume. Ensure your data mapping engine also handles legacy and semi/unstructured data, S3 and HDFS, URLs and brokered data streams.

GOVERN (CLEAN, PROTECT, DOCUMENT)

Without the structure developed from prior profiling and ETL steps, there is no understanding. And without understanding, there can be no trust in the quality or security of data in the lake, either.

It makes sense therefore to use a common layout format—and metadata management infrastructure—in the data lake. Voracity uses the same simple, shared 4GL to define and document data cleansing and masking jobs as it does for ETL operations.

Voracity can also combine all these operations in the same job script and I/O pass, and produce visual workflows and transform mapping diagrams to reflect everything you do. So, add to your ETL jobs data cleansing functions that:

- Filter and de-duplicate
- Validate and replace
- Synthesize and enrich
- Unify and standardize

And to those, add the simultaneous or separate ability to mask PII in structured, semi-structured, and unstructured sources. Voracity includes deterministic, secure, and reversible (or not) functions to:

- Encrypt, hash, and/or tokenize
- Pseudonymize
- Redact or randomize
- Blur or bucket to anonymize
- Encode or scramble

Voracity can also score the risk of re-identification from quasi-identifiers, and create compliance logs.

FISH (ANALYZE)

Once data is wrangled through the separate or combined processes of ETL, cleansing, and masking, it is ready to analyze through any models you need and visualize in any dashboarding system you prefer, including reporting and statistical analysis that can be performed during the aforementioned preparatory work. Such consolidation led Dr. Barry Devlin in 2018 to label Voracity a "Production Analytic Platform."

Voracity supports these analytic options for data lake users:

- Embedded BI—report and analyze *while* blending
- BIRT & KNIME—use Voracity data from APIs in the same Eclipse IDE
- Data Wrangling—Give display-ready subsets to Power BI, Qlik, R, Spotfire, Tableau, et al.

LEARN MORE AT
iri.com/voracity

ARCADIA DATA

Three Ways to Win With Data Lakes

The tide has turned for data lakes. Once thought of as "data swamps" with little value for anyone beyond data scientists, a 2018 study of 238 organizations with data lakes in production by Eckerson Group titled "Data Lakes for Business Users" showed that 72% of respondents say their data lake "fosters better decisions and actions by business users."

So, what changed?

Part of the answer may be the methods agile businesses are using. In the December 2018 report "The Three Part Guide to BI Modernization," Forrester Research found a correlation between agile business intelligence (BI) methodologies and high-growth firms. Namely, 83% of analytics decision makers at these firms agreed that their BI development and deployment is based on agile methodologies. Data lakes are one such approach that enables business agility to answer unexpected questions, enabling organizations to respond quickly to ever-changing consumer sentiment or prevent cybersecurity attacks.

Why are some struggling with getting insights and value quickly from data lakes while others thrive? There are three key attributes (or "Pro Tips") seen in organizations succeeding with their data lake projects.

PRO TIP #1: GIVE BUSINESS USERS THE TOOLS THEY LOVE

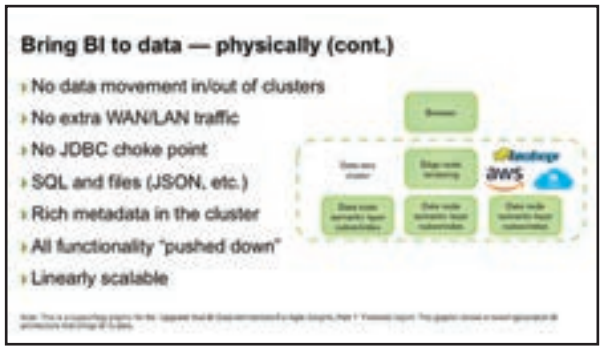
Initially, data lakes were the playground for data scientists to develop predictive models or train machine learning. Specialized skills were required. Today, organizations have given business users direct access. The same Eckerson study shows that nearly two-thirds of organizations agree that, "business users can explore data to get the views they want," and one-third of organizations have more than 250 users accessing their data lake. But how? By giving business users the tools they love—familiar BI interfaces that enable point-and-click dashboards and analytics for faster insights. In fact, Eckerson found that 50% use BI tools to explore data in the lake with only 25% using straight SQL and 25% using other methods. As tools have evolved, support for data lake environments have improved, enabling self-service BI for a broader group of users.

PRO TIP #2: INCORPORATE CLOUD IN YOUR DATA LAKE STRATEGY

Similar to the agility of BI tools, cloud services provide agility of another kind. Elastic scalability and fast provisioning enable faster application delivery and usage-based pricing which supports agility for decision support. While production data lakes are mainly on-premises, Eckerson found nearly a 40% growth in cloud-based deployments between organizations that deployed three or more years ago (18% of data lakes in the cloud) when compared to data lakes deployed in the past 1-2 years (25%). Experts expect this trend to continue as organizations reduce cost and time-to-deployment with cloud-based BI and data lakes.

PRO TIP #3: MODERNIZE BI ARCHITECTURE—BRING BI TO THE DATA

Organizations are choosing a new BI standard for their data lake based on a modern, distributed processing architecture for scale and performance. Traditional BI tools require data to be moved to their processing tier and modeled for high performance. This works fine for small data sets, but as data increases within data lakes, moving data to the BI tool requires data extracts and duplication while reducing the granularity of analysis possible. Forrester has recognized a new category of in-data-lake BI platform, which resides with the data (bringing BI processing to the data) and scales linearly to reduce network traffic as well as performance bottlenecks. In-data-lake BI also gives users more real-time access to data in its native state to unlock insights without waiting for data modeling.



In-Data-Lake BI Architecture Brings BI to the Data for Faster and Deeper Insights

UP YOUR GAME FOR DATA LAKE ANALYTICS TODAY

Winning companies are adopting modern, agile BI processes and technologies for their data lake environments. Whether you start with new cloud-based data lakes or testing modern analytics and architectures, the hardest part is getting started. Don't get left behind. Read Forrester's research on modernizing BI architecture or learn more about winning companies, their use cases, and how they have applied modern BI to data lakes at www.arcadiadata.com.

ARCADIA DATA
www.arcadiadata.com

SLAMDATA

Supercharge your Data Lake for Everyone

Empower analysts and data scientists to curate complex JSON into analytics-ready tables in minutes

Data lakes are increasingly popular for two primary reasons. First is the availability of cheap, large-scale cloud storage such as AWS S3, Google Cloud Store, and Azure Blob store. The second is the ability of these storage services to hold a wide variety of data without enforcing fixed schema or data models. This allows companies of all sizes to load huge amounts of data from a wide variety of sources quickly and easily into these services and create "Data Lakes." It's such an efficient process that even Data Warehouse companies like Snowflake use S3 as their primary place to store customer data.

So far so good, but the usefulness of the data starts to quickly break down once it enters the Data Lake. The variable nature of the data, including flexible schema data models like JSON and XML, make using the data for analytics and data science very challenging.

Enter the army of Data Integration Engineers. Much like the rise of the Data Scientist, the Data Integration Engineer has become the "rock star" of the Data Lake by cranking out tons of custom code to transform the Data Lake into analytics-ready data. Unfortunately, this approach can be slow, costly and anything but Agile in terms of curating fresh data in a timely manner.

While traditional ETL/ELT tools provide some help in extracting and transforming the well-structured data from the Data Lake, they provide little help transforming the messy JSON or XML data in the lake. The alternative is writing custom code, which requires a Data Integration Engineer. Until the Engineer completes his work, Analysts and Data Scientists are essentially locked out! All of this has contributed to the failure of many Data Lake projects.

REFORM

Well NOT anymore! With SlamData REFORM, Analysts and Data Scientists can quickly and easily curate fresh data sets from the most complex JSON data in minutes. The data can be stored in AWS S3, Azure, Google Cloud, MongoDB and more.

With SlamData REFORM now any user can:

- Easily browse any JSON data and discover the information they need
- Stream JSON data as tables into any Data Warehouse, RDBMS or ML solution
- Access JSON using standard SQL to perform JOINs, GROUP BYs and other operations
- Produce analytics from JSON with Tableau, Power BI, Python, R, Looker and more
- Stay up-to-date with the latest JSON data even as schemas change
- Painlessly make updates and check details
- Acheive 10X performance increase over manually written Python

No other solution can make complex JSON transformation so simple. In fact, REFORM is built from the ground up to work only on JSON data. Other solutions were built to deal with normal structured data, and limited JSON support was added as an afterthought. REFORM was purpose-built, and emphasizes ease of use. Don't take our word for it, give it a try.

SlamData REFORM is a software-based solution, not SaaS, which means there are no additional compliance or security issues with your data. It's a quick and simple install on any operating system via **Docker**, or as an AMI from the **AWS Marketplace**.

One installed, Analysts, Data Scientists or any Data Integration Practitioner can connect to their Data Lake and visually build analytics-ready data sets in minutes. The power behind REFORM that makes this possible is an advanced linear algebra called MRA. It understands JSON data in ways that traditional ETL and Data Prep tools simply can't, and makes working with this complex data simple, visual and painless.

If you want to empower your team and get more value from your Data Lake or Data Warehouse investments, then REFORM is a tool that can help.

SLAMDATA
https://slamdata.com/

melissa®

Prevent the Big Data Dump (Garbage In, Garbage Out) with Unison

Big data can lead to big insights for data-driven organizations. However, without cleansing the data flowing through your business, the potential for garbage in means a high potential for garbage out. Not what you'd hope for with analytics driving business decisions? A rigorous data quality process is your best defense against GIGO: Garbage In, Garbage Out.

Melissa can help prevent your big data from becoming a big data dump with Unison, an enterprise-level solution for deploying and maintaining contact data quality. Unison compiles industry-leading data quality APIs into one centralized platform for the essential speed, scalability and security that data stewards need to effectively clean data, while eliminating development time and the need for programming.

GLOBAL CONTACT DATA QUALITY SOLUTIONS

Unison currently features name, phone, email and global address validation with geocoding, streamlining data prep for ETL endeavors. A SERP and CASS Certified™ address engine performs aggressive corrections on U.S. and Canadian addresses. It also fixes spellings and naming mistakes for cities and streets, then adds the correct street name suffix, prefix and ZIP+4 information.

Address validation also appends latitude and longitude coordinates with U.S. Census data and Congressional districts to improve communication with your customers or constituents. Unison validates and standardizes email addresses and phone numbers, plus validates and parses through full names.

PROJECT COLLABORATION AND SCHEDULING AUTOMATION

Building the right community of data practitioners can be the difference between success and failure. Unison offers a lightweight, revolutionary approach to project collaboration with automated scheduling of jobs. Connect to your existing LDAP system, or quickly create a new one with multiple user accounts and access levels. Anyone permitted can link new databases, schedule a job for later, or process one instantly. Even the most stringent data privacy organizations are no longer shackled with having to trust one steward.

Additionally, no programming skills are required. Instead of building out a project then executing using an external scheduler, Unison allows you to do both within minutes then schedule the project using the tool itself. Once your ETL process is in place, Unison handles the rest!

SCALABILITY AND SPEED

The first thing you'll need to know about speed is that development time is completely eliminated. This platform supports Oracle, MS SQL, and MySQL database management systems right out of the box. It is central to the architecture's goal because the only bottleneck to rendering your data quality is the speed of the network.

The benefits of disseminating data governance across a wider scope of people in the enterprise extend to the benefits of Docker Swarm, as well. Unison has been clocked at 50 million records an hour, so if you're looking for the cutting edge of rapid results, Unison is it.

REPORTING AND ANALYTICS

As with project collaboration, a better approach to reporting and analytics is to involve people from various departments in the process for differing views on the same data.

Beautiful, easy-to-read reports are automatically generated per job to provide a high-level overview of how your data improved. This gives you both the high-level overview of operations on your data, as well as a detailed report on how your data changed. Unison's reporting system offers a drill-down at the record level for each section of reports, demonstrating exactly which record sets contain specific result codes.

SECURITY

You might say, "With all these users dipping into my data, how can I be sure that my security concerns won't be exacerbated?" Melissa has a long history of handling sensitive information, and places an emphasis on the security of your data. Unison leverages Docker's containerization and security features on the back end to keep everything locked down and ready for Docker Swarm's lightning fast cleansing. In addition, everything is handled securely, on premise, so you can worry about security less and get down to business.

TAKE ACTION

With the best defense against GIGO available in one centralized platform, preventing big data from dumping on big insights has never been easier. Unison supports your data governance operating model's best practices, plus the essential speed, scalability, and security necessary to enable stewardship across the organization—and effectively wrangle with GIGO.

For more information, visit www.Melissa.com/gigo

Melissa ■ 22382 Avenida Empresa ■ Rancho Santa Margarita, CA 92688-2112
Phone: 949-858-3000 ■ Fax: 949-589-5211 ■ Toll Free: 1-800-635-4772 ■ www.melissa.com

BIG DATA BY

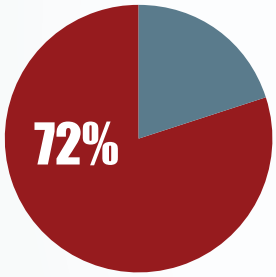
THE FUTURE OF DATA MANAGEMENT

Three intertwined initiatives—automation, DevOps, and cloud—are fueling modern data management opportunities. However, DevOps teams, web product managers (WPMs), and developers still spend most of their days troubleshooting.

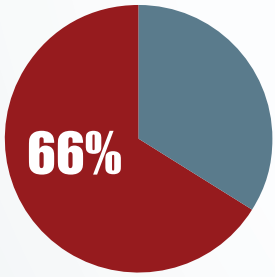
ENTERPRISE APPLICATION TEAMS ARE FACING PRESSURE TO RELEASE APPLICATIONS MORE QUICKLY, BUT MOST ENTERPRISES STILL HAVE A MANUAL PROCESS FOR REVIEWING, VALIDATING, AND DEPLOYING DATABASE CHANGES. THIS CREATES A BOTTLENECK FOR BUSINESS INNOVATION AND IMPROVING CX.

The top 3 benefits of database release automation are:

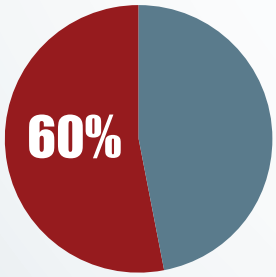
72%
The ability for developers to find and fix errors in database changes more quickly



66%
Reduced application downtime because of fewer bad database changes



60%
Improved application performance



Source: "The State of Database Deployments in Application Delivery," conducted by Dimensional Research, commissioned by Datical

AUTOMATION OF SECURITY POLICY CONFIGURATIONS IS A CRITICAL PRACTICE AT THE HIGHEST LEVELS OF DEVOPS EVOLUTION, BUT EXECUTIVES HAVE A ROSIER VIEW OF THEIR DEVOPS PROGRESS THAN THE TEAMS THEY MANAGE.

64% of C-suite respondents believe that security teams are involved in technology design and deployment versus 39% at the team level

Highly evolved organizations are **24x** more likely to always automate security policy configurations compared to the least evolved organizations

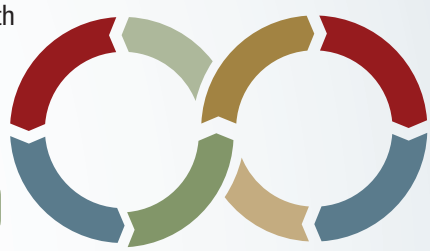
Source: "2018 State of DevOps Report" by Puppet and Splunk

THE NUMBERS

DevOps adoption is increasing in organizations, bringing application and database development closer together and resulting in business benefits and increased compliance with data privacy regulations.

- **85%** of companies have either adopted DevOps or plan to do so in the next 2 years
- The database is increasingly part of DevOps—avoiding issues with traditional approaches such as:
  - Failed deployments **(23%)** and
  - slow development and release cycles **(20%)**

  and having a positive impact on regulatory compliance **(61%)**



- Close to **two-thirds** of respondents described the relationship between developers and DBAs in positive terms, as application and development teams are increasingly working together on DevOps and adopting common practices

Source: "2019 State of Database DevOps" from Redgate Software

Many of the capabilities necessary to support the business—ensuring security, resilience, and support for mission-critical applications—are now available in the cloud, which is quickly becoming a leading choice for new deployments.

- **One-quarter of enterprise data is now managed by public cloud providers**
- **Scalability is the #1 benefit cited for cloud**
- **About half of new database projects are going to public cloud providers**
- **Hybrid arrangements are being embraced as enterprises seek to leverage the best of on-prem and public cloud**

Source: "2019 IOUG Databases in the Cloud Survey," produced by Unisphere Research and sponsored by AWS

Organizations need to provide tech professionals with tools that enable them to spend less time monitoring and troubleshooting and more time innovating to avoid the risk of a demotivated and demoralized DevOps team

- Troubleshooting app issues is the **#1** activity tech pros spend their time on, making it difficult to prioritize business growth and innovation
- **53%** of DevOps team members agree troubleshooting app issues is the top task completed on a given day

- On average, DevOps and WPMs spend less than **25%** of their time proactively optimizing performance of their environments

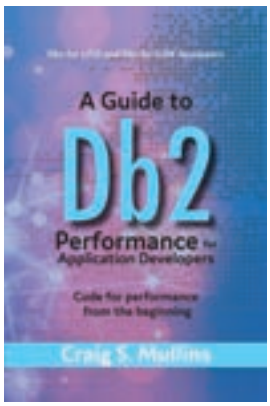
Source: "SolarWinds Cloud Confessions: The Trouble with Troubleshooting"

TRENDING NOW

Improving Db2 Performance

Q&A With Craig S. Mullins

Craig S. Mullins,
President, Mullins Consulting, Inc.



DESPITE THE ATTENTION given to new big data management technologies, Db2 remains one of the most widely used database management systems in the world and is a fundamental component of many enterprise data architectures.

In a new book, titled *A Guide to Db2 Performance for Application Developers*, Craig S. Mullins, president and principal consultant, Mullins Consulting, provides advice and direction to Db2 application developers and programmers on writing efficient, well-performing programs.

Mullins recently reflected on the role of Db2 in a big data world and the need for application code that will perform optimally.

'Data management today is a world of polyglot persistence—the right database systems for the right applications—and Db2 is definitely a part of this world.'

In the new era of big data and cloud, where does Db2 fit in?

Db2 is still one of the most powerful and most used database management systems in the world. It is, for the most part, the only relational DBMS used on mainframes—and mainframes still power most of the Fortune 500, including major financial institutions and banks, airlines, retailers, telcos, and more. It also has a substantial installed base on Linux, UNIX, and Windows platforms. Infoclutch estimates that more than 36,000 companies use IBM's Db2.

From a big data perspective, organizations are looking to run analytical processes on many different types of data. That said, business transactions are usually at the top of the list of the type of big data that executives want to analyze. And Db2 runs the transactions of many of the world's largest businesses.

What else is Db2 optimized for?

It is not just transactions that drive big data. And Db2 is well-equipped to handle, manage, and optimize many disparate types of data, including unstructured and multimedia data, JSON, XML, and even temporal data. Db2 has also been optimized to support very high-speed processing for complex Db2 queries to support business-critical reporting and analytic workloads. The Db2 Analytics Accelerator for Db2 for z/OS and Db2 BLU for Db2 for LUW provide a hybrid transaction and analytic processing (HTAP) environment that drives out cost and complexity and enables analytics on transactional data as it is generated.

And finally, keep in mind that the new NoSQL database systems are not replacing relational systems in existing applications but are augmenting the data processing capabilities of organizations implementing applications that are not traditional relational/SQL applications. So data management today is a world of polyglot persistence—the right database systems for the right applications—and Db2 is definitely a part of this world.

Why is database performance so critical now?

Every database application, at its core, requires three components in order to operate: the system, the database, and the application.

To deliver performance, we must be able to monitor and tune each of these components. This is easier said than done. Of the three, the aspect that causes most performance problems is the application code. As much as 80% of all database performance problems are caused by inefficient application code—and much of that is due to inefficient and improperly coded SQL.

Mullins will be at Data Summit 2019 to present a session, titled "The New World of Database Technologies," on Tuesday, May 21.

To access the book, go to https://store.bookbaby.com//bookshop/book/index.aspx?bookURL=A-Guide-to-Db2-Performance-for-Application-Developers&b=p_bu-ba-or.

Repeatable Machine Learning With Kubeflow

In the last couple of years, we have heard a lot about machine learning and the plethora of tools and frameworks available to support these efforts. People have begun to understand that without data and proper data operations practices, these machine learning tools are very difficult to leverage. While there is an abundance of great tools available, there is one persistent complaint, and that is how to achieve simple repeatability. Workflow management must encompass the data scientists' pipeline for attaining the data and preparing it, building and testing models, and versioning the entire workflow. It must also support versioning the data inputs, outputs, logs, and even performance metrics.

This workflow management task is one of the more critical pieces of infrastructure to support data scientists and the machine learning endeavors of any organization. In the book, *Machine Learning Data Logistics*, published by O'Reilly, it is noted that 90% of the time and effort that go into machine learning by a data scientist is devoted to data logistics. Anything that can be done to reduce that level of effort should be considered very important to the success of machine learning-based programs.

Most recently, we have seen some workflow tools rise to the occasion. They have gained rapid popularity for their flexibility, ease of adoption, simplicity in extending their capabilities, and—at least in some part—because they plug into Kubernetes (k8s). Plugging into k8s enables the workflow to be operated or scheduled by k8s, removing the often mundane and tedious task of identifying the hardware for which to run a workflow. This appeals to many end users because k8s has effectively become the data center resource manager of choice, the golden standard, if you will, for managing and allocating resources for jobs.

Kubeflow is a workflow tool which prides itself on making machine learning workflows simple to build, scalable, and portable. It provides graphical end-user tools to set up and define the steps in a pipeline. Most importantly, as data scientists build out their use cases, they add more and more steps and, when using Kubeflow, they end up with a documented, repeatable process.

Kubeflow delivers on the promise of portability by enabling users to create and run workflows on their desktops. That same workflow can then be moved to another environment, such as production. Abstractions to separate the shape and size of the environment are built into k8s so the user does not need to worry about how to deploy into a test or production environment. This is a big benefit since development, testing, and production environments nearly always look different from one another, ultimately leading to reduced friction for the data scientist.

Since we are talking about repeatability for machine learning, it is important to keep GPUs in mind. The GPU has a place in many of these workflows. A data scientist may use an algorithm that is CPU- or GPU-based. The process gets more complicated when a GPU is required in an environment, when trying to abstract the environment to make the workflow portable. This is a problem that Kubeflow with k8s solves. Kubeflow supports Chainer for performing model training as well as RAPIDS.ai. There are many more popular frameworks that have been integrated into Kubeflow which leverage GPUs. Jupyter notebooks, TensorFlow, MXNet, PyTorch, Katib, Horovod, Istio, Caffe2, and TensortRT are all supported. This is just a short list of the many integrations that exist, and, if nothing else, it should give confidence to anyone considering Kubeflow that it is the workflow tool of choice for ML workloads.

Something can be a great tool, but if the barrier to entry is too high, people won't use it. Fortunately for data scientists, Kubeflow is easy to get started with. Kubernetes is the key requirement and it is easily attainable for the desktop with microk8s or minikube. Aside from that, there is a "quick start" on the kubeflow website which takes minutes to follow through.

While workflow management is one of the most important items to enable machine success, it is also important to differentiate between generic workflow management and machine learning-based workflow management tools. This is the difference between the jack-of-all trades and the master of machine learning workflow management. Having depth baked into a solution supporting machine learning is critical because the needs of data scientists are very different from those of traditional software engineers and even systems administrators.

Jim Scott, VP of enterprise architecture at MapR (www.mapr.com), is the co-founder of the Chicago Hadoop Users Group (CHUG), where he coordinates the Chicago Hadoop community.

Does Your Viz Pass the Eye Candy Test?

By now, the practice of data visualization for analytical insight has become part and parcel to many (if not most) analytics programs. Thanks to the prolificacy of self-service tools and a new visual imperative, coupled with an increase in visualization-capable analysts entering the analytics workforce, everyone, from executives to business users, is taking advantage of the ability to see and understand data using the power of visualization.

The use case for data visualization is simple. Whether it's a single visualization, a dashboard, or even a data story, when layered upon the foundation of careful statistics and good data, a well-balanced data visualization can deliver meaningful, immediate, and actionable insights at a glance. When done correctly, using visualization for insight is intuitive and it's easy; it doesn't just give you the data you need, it shows you what you need to know about it, too. Through the careful distillation of images, color, and design, a well-crafted data visualization can leverage the visual horsepower of the human brain to spur decision-making, improve analytical research, or act as a beacon for effective communication. Good data visualization should be fast, informative, and—above all—valuable. This makes data viz a critical tool in the modern analyst toolkit.

Although based in data and statistics, it's hard not to think visually when it comes to data visualization. Half science and half art, visual information representation is an artful process, with the end results delivered via colorful and content-rich charts and graphs; carefully crafted dashboards; compelling, data-driven narratives; or even astute infographics.

But visualization comes with its own set of risks. Getting too carried away with the artful, visual aspects of data visualization can dilute—or worse, distort—the data's meaning. What could be extremely insightful presentations of data can become little more than pretty colors cleverly disguised as information visualization. While a good data visualization is often an aesthetically pleasing one, beauty itself does not equate to analytical prowess. Thus, any visualization attempt should not forget the cardinal rule of data visualizaton: Above all else, show the data.

Based in the greater New York City area, **Lindy Ryan** researches and teaches business analytics and data communication at a major East Coast university, and is the author of *The Visual Imperative: Creating a Culture of Visual Discovery.* Follow her on Twitter @lindy_ryan.

The key to curating a "well-balanced" data visualization isn't just providing a visual representation of data; it's providing the right kind of visualization for the data and visualizing it with intent by using the art to support the data's story and not overshadow it.

Here are three simple questions to ensure your data visualization passes the eye candy test:

Is it approachable?

First, make sure the visualization is straightforward and easy to understand by its intended audience. Then, capitalize on the fact that people perceive a design as easier to use if it includes design elements (color, shapes, etc.) to make it visually appealing and leverages pre-attentive features. This is visual design, the practice of removing and simplifying things until nothing stands between the message and the audience. In visualization, the best design is the one you don't see because you're too busy looking at the data.

Does it tell a story?

A visualization should tell a story about—or explain—its data. Therefore, visualizations require a compelling narrative to transform data into knowledge and emotion into action. Make sure your visualization has a single story to tell. Too often, people want to present all the data in a single visualization that can answer many questions, but effective visualizations are a one-visualization-to-one-story ratio. Your audience needs answers, not more questions.

Is it actionable?

Does the visualization provide guidance through visual clues that prompt action? Visualizations should leverage visual clues (colors to highlight or alert, or annotations to narrate important information) and dashboards should establish a visual hierarchy to direct the audience's attention. This is the "fitness" test: Before you even know what the numbers say, your audience's eyes should know where to go to decode the information being presented.

If you can answer "yes" to all three questions, chances are good that your visualization is a well-designed, meaningful, non-eye candy data visualization that leverages colors, shapes, and design to not only display but influence the way your audience receives insights into data. If it doesn't, this test should help you identify where you need to go back and spend a little more time perfecting your viz.

The Real Dragon in the Room

What happened to the 500 million datapoints from the Starwood data breach?

By now it is old news. Everyone has heard about the Marriott data breach. The headlines told the story: "GDPR May Add Up to $915M Marriott's Data Breach Expenses," proclaimed *Forbes* in a headline; "New Year, new tactics to keep your personal info safe after Marriott," said the *Los Angeles Times*; "Marriott: Hackers accessed more than 5 million passport numbers," stated *The Washington Post*.

And, it wasn't really even Marriott—at least, not when it started. However, even suggesting that we can identify when the breach was initiated is a tenuous supposition. Perhaps the most frightening aspect of the entire debauched state of affairs is that no one truly knows who has the data and what they intend to do with it.

Yet, almost everyone leads with Marriott Hotels since it makes for a better soundbite. To be clear, it was Starwood's guest reservation database that exposed up to 500 million records, which included the most personal and horrifically useful datapoints. Marriott inherited this mess when it acquired Starwood Hotels & Resorts Worldwide. The Marriott Hotels reservation system was not affected by the breach.

This breach of Starwood's reservation system included names, addresses, phone numbers, passport numbers, birthdates, and genders in Starwood's Loyalty Program account information. Starwood brands include St. Regis Hotels & Resorts, Sheraton Hotels & Resorts, Westin Hotels & Resorts, Le Meridien Hotels & Resorts, and W Hotels.

What Happened to the Data?

As the horror gradually revealed itself similar to a slowly opening barn door on a moonlit evening in a Stephen King short story, it became clear that the breach had been lurking within the system for an indeterminate number of years. As per its main focus of news-entertainment, much of the media has focused on the more sensational sound bites and dramatic visuals as opposed to taking the time to ask the really important questions, such as: What happened to the data?

According to Bob Sullivan, an independent journalist and 20-year veteran of MSNBC.com/NBC news, "This is no ordinary credit card data heist. If the criminals were using card accounts stolen in this incident, banks would have figured out where the stolen cards had come from long ago. I doubt it's an identity theft ring. There's no way some kind of casual prankster or amateur would have kept up this effort for four years. Something more serious is going on here" (https://bobsullivan.net/cybercrime/starwood-breach-what-should-you-do-that-depends-on-who-and-why).

It was clear when the story first broke that Sullivan was one of the few with the insight and interest to consider the more pertinent ramifications of what happened. Very few outlets were addressing the real problems we should all be concerned about: Who did it? What happened with the data? What do they plan to do with the data? For 4 years the hackers had all this data, yet not a single credit card number was sold on the dark web.

We agree with Sullivan's assessment that something more serious is going on. Our guess is that it is a state actor with a 100-year plan, such as China or Iran, but of course we are guilty of gross speculation. One may want to consider what other mega-actors, with 100-year plans, would hold onto the personal records of nearly every business traveler on earth for an indefinite amount of time. Who else has the resources and long-term determination with no apparent need to monetize this fortune?

We reached out to Sullivan, who graciously agreed to let us interview him and gather his thoughts on all things cyber and more specifically the Marriott subsidiary Starwood's data breach.

The Sad State of Cybersecurity Today

When we asked Sullivan his thoughts on the state of cybersecurity today, his words were very chilling. "In almost all cases a company has to hire an outside company before they release how big a deal the break-in really is," he said. Somehow, it is legal that companies get to hire lawyers and PR firms before they reveal to the general public that they have been robbed.

According to the official Marriott International News Center, "On September 8, 2018, Marriott received an alert from an internal security tool regarding an attempt to access the Starwood guest reservation database in the United States. Marriott quickly engaged leading security experts to help deter- ▶

Michael Corey, co-founder of LicenseFortress, was recognized in 2017 as one of the top 100 people who influence the cloud. He is a Microsoft Data Platform MVP, Oracle ACE, VMware vExpert, and a past president of the IOUG. Check out his blog at http://michaelcorey.com.

Don Sullivan has been with VMware (www.vmware.com) since 2010 and is the product line marketing manager for Business Critical Applications and Databases with the Cloud Platform Business Unit.

mine what occurred" (http://news.marriott.com/2018/11/marriott-announces-starwood-guest-reservation-database-security-incident).

With all the money and resources available to a company such as Marriott, even it had to bring in outside experts to help determine just how big the data breach was. This makes Sullivan's words even more chilling.

Marriott hired a new security officer in January 2018 and on its recent Form 10-K report identified a number of business risks around security. "Cyber-attacks could have a disruptive effect on our business," the report noted. It added, "Changes in privacy and data security laws could increase our operating costs, increase exposure to fines and litigation." In Marriott's 2018 proxy statement (https://marriott.gcs-web.com/static-files/e8be6c13-f70c-4d5a-8d56-a46992c7edb8), it noted that the board of directors reviews the company's cyber risk profile and is informed of specifics to Marriott's cybersecurity risk program.

This should be of no surprise given that it is a well-known fact that hospitality is the third most frequently targeted industry, after retail and finance. Yet, with all of Marriott's efforts, money, and resources, its subsidiary Starwood Hotels failed to protect its customers' information.

The GDPR Effect

In May 2018, the EU's General Data Protection Regulation (GDPR) took effect. This is something we have been writing about for years. Marriott could become the first major test case for Europe's new stringent data laws. GDPR fines can be substantial for a serious violation—up to 4% of annual revenues. With revenues of more than $22 billion, a 4% fine could be levied if it's found in violation of the new EU law.

Legislation similar to GDPR is long overdue in the U.S. A number of prominent CEOs have had to face the U.S. Congress over privacy transgressions. In 2017, the CEO of Equifax was put in front of the congressional cameras. In 2018, Mark Zuckerberg, the CEO of Facebook, was grilled by both House and Senate committees. Unfortunately, despite the seemingly genuine interest and effort of the congressional leaders, they were clearly ill-prepared to draw any substantial conclusions. So, nothing changes. With fines up to 4% of global revenues, we may see the effect of GDPR even in the U.S. Possibly major corporations will now at least consider privacy and security more seriously, given the fact the it will finally hurt the bottom line.

What's Ahead: Sarbanes-Oxley Meets Cybersecurity

It is time to ask the question: Why is there not legislation similar to the Sarbanes-Oxley Act established as a set of federal requirements for U.S. corporations as it pertains to privacy and security? As a nation, we need to create a system of legitimate accountability for officers of corporate America. Too often, companies casually use personal data, especially when there is even a modicum of marginal profit. Frequently, security breaches are the result of negligence or simple indifference. And, in many cases, breaches occur from an obvious vulnerability that a company knew about many months prior but it neglected to take reasonable and necessary actions to address in a timely manner. Why is that? Is it because they simply don't care enough?

The consequence for such negligence is typically a one-time write-down or momentary public exposure in front of American media outlets that are mostly looking for a dramatic story to lead with on the nightly news. In other words, it's no big deal.

Often, corporate officers will be dragged in front of cameras and even get hauled before Congress to testify. Sadly, it's cheaper many times for the company to pay the fines than to actually fix the root cause of the problem. This is not a new behavior from U.S. corporations where they compare the cost of fixing a product to the cost of lawsuits if they

don't. What is clear is that if we want companies to keep our data secure, then we must hold the corporate officers responsible. We need to make sure the penalties are severe for mishandling our data and treating our personal lives in the most negligent of manners.

The U.S. Border Needs to Get a New Type of Wall

It is apparent that U.S. corporations need to be held to a higher standard. It is also time to expect more from Uncle Sam. "Is it fair that Marriott has to defend itself from a state actor?" asked Sullivan. Let's face it, we are at war. There is an old saying, "If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck."

While we should expect more from U.S. corporations, the U.S. government needs to own the strategic solution to this problem. We need more programs similar to the FBI InfraGard Program (www.infragard.org). This is a program in which the FBI partners with corporations on cybersecurity. When a major event happens, the Department of Homeland Security contributes resources to address, minimize, understand, and fix the problem. Why not an agency similar to FEMA (How about Cyber Attack Task System [CATS])? While Trump fights for his physical wall, Congress needs to enact laws to hold corporations responsible and to also put the necessary funding, resources in place to build a cyberwall that keeps American citizens' data safe.

When this information is used against us in the future, it will be 100 times more costly to fix than it would have been to put preventive capabilities in place.

Insurance Companies' Exclusions

Today, it is commonplace for companies to buy cyberinsurance. Yet, if you read the fine print, insurance companies are not responsible for acts of war. It's just a matter of time before one of the big insurance companies invokes its exclusion and decides not to pay a claim.

'It is apparent that U.S. corporations need to be held to a higher standard. It is also time to expect more from Uncle Sam.'

The Real Questions

Too often, news outlets focus on collecting the most dramatic details and not on providing us with accurate information about real problems. The key questions regarding the Marriott subsidiary Starwood's data breach are: Who did it? What happened with the data? What do they plan to do with the data?

If we want our privacy and data to be respected, then we need to pass a law similar to Sarbanes-Oxley for cybersecurity. We need to create a system of legitimate accountability for the officers of corporate America. We need to impose fines that really matter to a corporation's bottom line. We need to expect more from our government. The FBI program InfraGard is a step in the right direction. But no corporation has the skills and resources to protect itself from a large state actor. Corporations such as Equifax, Target, Marriott, and others need to know the U.S. government has their backs. We need to build another type of wall, a "cyberwall," that helps keep our data secure.

In 1987, in a speech given in West Berlin, President Ronald Reagan said, "Tear down this wall," calling for Mikhail Gorbachev to open up the barrier separating the east from the west. Our words to President Donald Trump are: "Build us a cyberwall." This will be much more valuable to Americans than any other wall currently under consideration.

And, we need our CATS to extinguish the dragons—but most importantly we need more journalists such as Bob Sullivan asking the questions that really matter.

Data Lake Lessons From the Coffee Can

Considering a data lake? Or reconsidering one? You are not alone. The topic of the value a data lake can add to an enterprise data architecture remains evergreen. However, the verdict on the value it delivers remains divided. And value is not the only division of opinion; so is the definition of a data lake, its purpose, how it is managed, and how it is used.

The challenge is that the definition, purpose, use, management, and value of a data lake are all dependent on several factors, including who you ask and what they are trying to accomplish. It is very contextual. To bypass the subjectivity of context in understanding the value and use of data lake, perhaps an analogy is better than expert opinions. Your mileage may vary, but there is an allegorical household phenomenon that provides some lessons in understanding the data lake—the coffee can.

It is not just any coffee can. It is the coffee can that ends up in the garage or tool space of a home to collect all the stray, old-project hardware someone in the household cannot bear to part with. They may say something like, "We shouldn't throw these out. We might need them some day." Or, they may say, "You never know when these might come in handy." And no matter how valuable these extra pieces and parts may be one day, to this person, they are not worthy of the organization required so that they may one day be easily found again. Oh no—they are simply added to the coffee can where other nuts and bolts from long-ago projects have been brought together to one day serve their not-yet-defined purpose.

Sound familiar to your data world? Have you heard comments such as, "If we can just get all of our data in one place, then we can manage it"? Or, have you heard, "Since storage costs are no longer a concern with our open source storage options, we should collect and store all the data we can now. We can analyze it later"? In fact, some approach the data lake idea with a "store everything" mentality, resulting in data hoarding behavior. It is similar to the coffee can: Don't throw it out—we might need it someday. But, even the coffee can has rules and limitations. And, when used appropriately, value. And so should your data lake.

First, let's take a closer look at the rules, limitations, and value of the coffee can.

1. **Rules.** Albeit, the rules of the coffee can are unwritten, but understood—perhaps passed down from previous generations of tool-bench-coffee-can owners. The purpose of the coffee can is to collect fastener hardware in case a fastener is needed for a future unexpected project. Screws, nuts, bolts, nails … you name it; if it's left over and it can be used to fasten something later down the road, it goes in the coffee can. But, it's not for tools. Hex keys or Allen wrenches do not belong, no matter how small they are. Drawer pulls? Nope. Pegboard hooks? Nope.

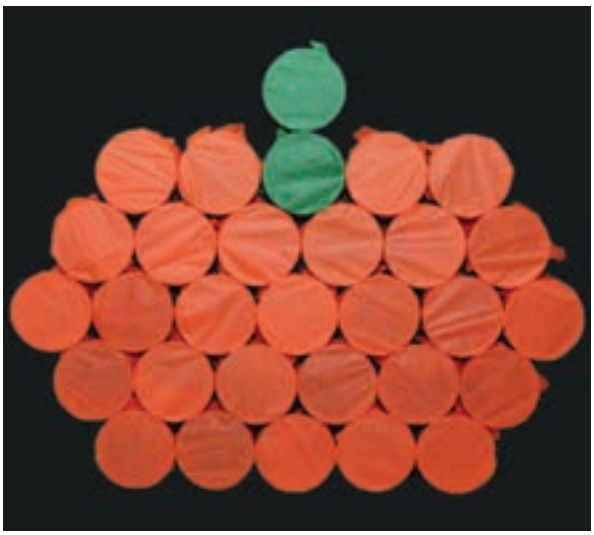
2. **Limitations.** The coffee can is about 7 inches tall and just more than 6 inches round. At some point, it will get full. But it can never get completely full, because you still need room for rummaging when needed. The can can't be exposed to the elements or many of your fasteners will rust. It needs to be readily accessible, yet out of reach from those who could hurt themselves. (I love the data lake corollary here. But, hold tight—we will get to that.) Ideally, the can would have a lid on it to prevent some of these issues, but that is fundamentally against one of the unwritten tool-bench-coffee-can rules. #NoLids

3. **Value.** In addition to tool-time pride, the coffee can of collected orphaned hardware delivers value in a couple of ways.
   a. *The one-off/ad hoc project.* A project that occurs once or requires similar but not identical hardware. These are not large-scale projects and once complete are not likely to be repeated. Take, for instance, a trundle drawer pull that loses a screw and the handle hangs loose—a perfect job to dig through the coffee can to find a screw good enough to fix the handle. The fastener specifications don't have to be precise if the drawer handle is ultimately secured. It could be a Phillips head screw, a flat head screw, a carriage bolt, a wood screw, or a machine screw. As long as it meets a

Anne Buff is a business solutions manager and strategic advisor at SAS Institute (www.sas.com). In her current role, she leverages her training, consulting experience, and data-savviness to help guide executives and their companies to realize their full data and analytics potential.

few needs (width, thread size, and length), the project is complete, and no other coffee can diving is necessary.

b. The prep work for a large-scale project. Ultimately, you have big plans for an extensive project, maybe even one that will be ongoing. The decisions you make must be scalable and sustainable, but at the outset you are not exactly sure what you need. You don't know what type of fastener is best. An example here is the Poke-A-Pumpkin game, the Pinterest project of all Pinterest projects. It's a fun fall game where players get to poke through a tissue covered cup to grab a hidden surprise. What's really the surprise is how tricky it is to build. We can see a fastener is needed for each cup, but without instructions (who follows Pinterest instructions, anyway?) the specific width, length, and weight (a surprisingly important factor) are completely unknown. The coffee can affords you the opportunity to dig through its contents and try many types, sizes, and weights to determine what is best before you scale up and out. This testing-ground or sandbox environment allows you to test small and scale large with confidence.



Now, with a better understanding of the tool bench coffee can, let's consider the data lake.

1. Rules. Written rules are better in business environments than rules assumed and passed down through tribal knowledge. But the unwritten rules of the coffee can are still better than the no-rules approach typically applied to data lakes. Data lakes cannot be a free-for-all data dumping ground. What goes in the data lake should be predicated on what is expected to come out of the lake and how the lake is to be used. Sure, you can store data of all types, structured, semi-structured, and unstructured, but do so with purpose.

2. Limitations. A typical driver for data lake consideration is reduction in storage costs. While "the more data, the better" sounds wonderful, there are always limits. And, the limits become greater the more specific the business use case becomes. Processing, response time, access, and security are typical project needs where a data lake is likely to fall short. As a project becomes more defined, do not be afraid to move from the lake to another environment. Just as with the coffee can, the data lake cannot be exposed to the elements as it puts the contents at risk. And, similar to digging through the coffee can without gloves, digging in the data lake without protection hurts too. Most certainly, the data lake should not be accessible to those that could hurt themselves (or the company).

3. Value. The data lake serves as a fantastic storage area for data with potential use and purpose. However, it does not deliver value until the data is used. For ad hoc projects, the data lake data may be used in a once-and-done fashion and deliver immediate value. In these instances, even good enough fits the bill.

But, in extensive projects where good-enough does not work, where security, scalability, and repeatability matter, the data lake provides value as a sandbox environment or testing ground, preventing costly mistakes if deployed at scale. Once specific data needs are determined, scaling and operationalizing these projects should then be executed in more appropriate, sustainable, and secure data environments.

As you consider your data lake options, consider these lessons from the tool bench coffee can. A controlled, managed environment to store data "that might come in handy one day" might not be such a bad idea after all.

Preprocessing Data for Analytics: A Review

Data is the key ingredient for any analytical exercise. It is of the utmost importance to thoroughly consider and list all data sources that are of potential interest and relevant before starting the analysis. Large experiments, as well as our own experience in different fields, indicate that when it comes to data, bigger is better. However, real-life data can be dirty because of inconsistencies, incompleteness, duplication, merging, and many other problems. Throughout the analytical modeling steps, various data preprocessing checks are applied to clean up and reduce the data to a manageable and relevant size. Worth mentioning here is the garbage in, garbage out (GIGO) principle, which essentially states that messy data yields messy analytical models. Hence, it is of critical importance that every data preprocessing step is carefully justified, carried out, validated, and documented before proceeding with further analysis. Even the slightest mistake can make the data unusable for further analysis and the results invalid and of no use. Let's look at some of the most important data preprocessing activities.

Data Preprocessing Activities

The application of analytics typically requires or presumes the data will be presented in a single table containing and representing all the data in a structured way. A structured data table enables straightforward processing and analysis. Typically, the rows of a data table represent the basic entities to which the analysis applies (e.g., customers, transactions, enterprises, claims, cases, etc.). The rows are also called instances, observations, or lines. The columns in the data table contain information about the basic entities. Plenty of synonyms are used to denote the columns of the data table, such as (explanatory) variables, fields, characteristics, indicators, features, etc.

Denormalization refers to the merging of several normalized source data tables into an aggregated, denormalized data table. Merging tables involves selecting information from different tables related to an individual entity and copying it to the aggregated data table. The individual entity can be recognized and selected in these tables by making use of (primary) keys, which have been included in the table to allow identifying and relating observations from different source tables pertaining to the same entity. Figure 1 on the next page illustrates the process of merging two tables (i.e., transaction data and customer data) into a single denormalized data table by using the key attribute type ID that connects observations in the transactions table with observations in the customer table. The same approach can be followed to merge as many tables as required, but the more tables are merged, the more duplicate data might be included in the resulting table, as a consequence of the denormalization. It is crucial that no errors are introduced during this process, so checks should be applied to control the resulting table and to make sure that all information is correctly integrated.

Sampling takes a subset of historical data (e.g., past transactions) and uses that to build an analytical model. A first obvious question that comes to mind concerns the need for sampling. It is true that, with the availability of high performance computing facilities (e.g., grid and cloud computing), one could also try to directly analyze the full dataset. However, a key requirement for a good sample is that it be representative of the future entities on which the analytical model will be run. The timing becomes important since transactions of today are more similar to the transactions of tomorrow than the transactions of yesterday. Choosing the optimal time window of the sample involves a trade-off between lots of data (and hence a more robust analytical model) and recent data (which may be more representative). The sample should also be taken from an average business period to get an accurate picture of the target population.

Exploratory analysis is a very important part of getting to know your data in an "informal" way. It allows gaining some initial insights into the data which can be usefully adopted throughout the analytical modeling stage. Different plots/graphs can be useful here, such as bar charts, pie charts, histograms, and scatter plots.

The next step is to summarize the data by using some descriptive statistics which provide information regarding a particular characteristic of the data. Basic descriptive statistics are the mean and median value of continuous variables, with the median value less sensitive to extreme values but not providing as much information with respect to the full distribution. Complementary to the mean value, the variation or the standard deviation provides insight to how much the data is spread around the mean value. Likewise,

Bart Baesens is a professor at KU Leuven (Belgium) and the University of Southampton (U.K.) doing research on big data and analytics, web analytics, fraud detection, and credit risk management. See dataminingapps.com for an overview of his research.

Transactions		
ID	Date	Amount
XWV	2/01/2015	$52
XWV	6/02/2015	$21
XWV	3/03/2015	$13
BBC	17/02/2015	$45
BBC	1/03/2015	$75
VVQ	2/03/2015	$56

Customer Data		
ID	Age	Start Date
XWV	31	1/01/2015
BBC	49	10/02/2015
VVQ	21	15/02/2015

Transactions				
ID	Date	Amount	Age	Start Date
XWV	2/01/2015	$52	31	1/01/2015
XWV	6/02/2015	$21	31	1/01/2015
XWV	3/03/2015	$13	31	1/01/2015
BBC	17/02/2015	$45	49	10/02/2015
BBC	1/03/2015	$75	49	10/02/2015
VVQ	2/03/2015	$56	21	15/02/2015

Figure 1: Aggregating normalized data tables into a non-normalized data table.

percentile values, such as the 10th, 25th, 75th, and 90th percentile, provide further information about the distribution and are complementary to the median value. For categorical variables, other measures need to be considered, such as the mode or most frequently occurring value. It is important to note that all these descriptive statistics should be assessed together (i.e., in support and completion of each other). For example, comparing the mean and median can give insight into the skewness of the distribution and outliers.

Missing values can occur for various reasons. The information can be non-applicable. For example, when modeling the amount of fraud, this information is only available for the fraudulent accounts and not for the non-fraudulent accounts since it is not applicable there. The information can also be undisclosed, such as a customer who has decided not to disclose his or her income because of privacy. Missing data can also originate from an error during merging (e.g., typos in name or ID). Missing values can be very meaningful from an analytical perspective because they may indicate a pattern. As an example, a missing value for income could imply unemployment, which may be related to default. Some analytical techniques (e.g., decision trees) can directly deal with missing values. Other techniques need additional preprocessing. Popular missing value handling schemes are removal of the observation or variable and replacement (e.g., by the mean/median for continuous variables and by the mode for categorical variables).

Outliers are extreme observations that are very dissimilar to the rest of the population. Two types of outliers should be considered: valid observations (e.g., the CEO's salary is $1,000,000) and invalid observations (e.g., age is 300 years). Two important steps in dealing with outliers are detection and treatment. A first check for outliers is to calculate the minimum and maximum values for each of the data elements. Various graphical tools can also be used to detect outliers, such as histograms, box plots, and scatter plots. Some analytical techniques, such as decision trees, are robust with respect to outliers. Others, such as linear/logistic regression, are more sensitive to them. Various schemes exist to deal with outliers. It depends upon whether the outlier represents a valid or invalid observation. For invalid observations (e.g., age is 300 years), one could treat the outlier as a missing value by using any of the schemes (i.e., removal or replacement) discussed in the previous section. For valid observations (e.g., income is $1,000,000), other schemes are needed, such as capping, in which a lower and upper limit are defined for each data element.

These are some of the key activities when preprocessing data for analytics. Appropriately preprocessing data is of high importance in building powerful analytical models.

Working Backward (Or: What IoT Can Learn From Steve Jobs)

The complexity in the domain of the Internet of Things (IoT) is staggering. Gaining greater insight into your business by learning how operations deep at the heart of the organization are really performing requires a level of analytical proficiency that (until recently) was only found in either very large or very specialized organizations. In order to notch the analytics up to the next level in your organization, you will need to double up on your efforts. All IoT use cases—and especially Industrial IoT—are doomed if the organization fails to get to grips with the increased complexity.

Take, for example, predictive maintenance. This is a use case that is on everyone's mind, especially if you have any form of assets in your organization that support your primary processes. Ask yourself: How many different assets do you have? Do you have historic data available on those assets? Do you know the normal behavior of those assets? How will you solve the challenges of identifying abnormal behavior?

But before you get bogged down in the details, it is helpful to realize that there is a good approach to address those challenges. It is called "working backward," an approach made famous by Steve Jobs in the context of CX. (Borrowing from businessman and author Stephen Covey, this is habit number four in my continuing series on the seven "habits" that successful IoT projects have in common.)

The idea behind it is simple. Just as when you were a kid and you figured out that it was easier to solve a maze by starting at the destination and then working your way backward, you will see that it is easier to solve these types of complex cases by tracking back from a desired outcome. If this sounds similar to good detective work, it is!

Let's take the predictive maintenance case. The first question you must ask is: What outcome is it that any organization would normally like to achieve? This answer is probably already in your mind, but it never hurts to make sure all your colleagues are on the same page. The answer might be as simple as preventing a machine or asset from failing by detecting early signs of problems (and taking an appropriate action).

Now, the tougher questions need to be addressed. This shouldn't be a challenge if you are working your way through our seven habits because you will have your multi-disciplinary team in place to come up with an answer to the question of what kind of analytical model could give you such an insight. The data scientist in your team might tell you to set up a model to calculate remaining lifetime. The data scientist probably asked you to join a meeting with maintenance guys, and together, you would have a conversation on how, for example, increased vibration, temperature, and energy consumption are good signs for the particular asset you have in mind.

"Fantastic," you think, "are we there yet?" Well, not completely. You now want the engineers in your team to figure out how that behavior can be measured. The engineers with whom you discuss the devices that could measure vibration might conclude that vibration can be measured if they have sensors that are able to measure acceleration and velocity. Once you have gotten to this point, there is light at the end of the tunnel, as you now can connect the dots end-to-end. You are ready to align all topics on your journey back from the sensor to the original goal of being able to do predictive maintenance. Where acceleration and velocity sensors help you to collect vibration data, the data scientist can then analyze it and fit this data into a remaining life model. This will help you to validate the assumptions of the maintenance guys and allow you to assess potential early warning indicators that are expected in predictive maintenance. (For simplicity's sake, I left out the fact that an enterprise architect might point out that connection to some administrative case systems might be necessary too, but—hey—he is on your team too, so no sweat!)

The good part of this practical approach is that it works in many situations—not only IoT—and many leaders have advocated it. So when you preach it to your team, you will be comfortable in saying, "Guys this is common knowledge, even Steve Jobs did it" (www.imore.com/steve-jobs-you-have-start-customer-experience-and-work-backwards-technology).

For previous articles in this series, go to www.dbta.com/BigDataQuarterly.

Bart Schouw VP of technology and digital alliances, Software AG (www.softwareag.com).

EARLY BIRD PRICING

NOW AVAILABLE!

Cognitive Computing & AI Summit

MAY 21-22 2019

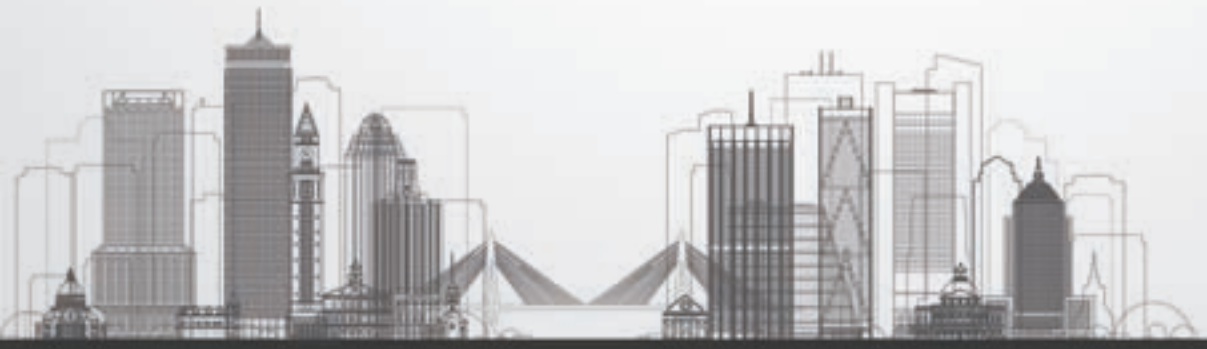
A featured event at

DATA SUMMIT
UNLEASH THE POWER OF YOUR DATA

HYATT REGENCY BOSTON | BOSTON, MA

A new era of cognitive computing has already begun, and its impact is being felt across industries, from healthcare and financial services to manufacturing and education. However, building cognitive systems and applications that can perform specific, humanlike tasks in an intelligent way is far from easy. The Cognitive Computing & AI Summit is an intense two-day immersion into the leading cognitive computing and AI use cases, strategies, and technologies that every organization should know about. Whether you are a developer, engineer, executive, entrepreneur, or product manager, if you are on the front lines of AI and cognitive computing, this summit is for you. **Reserve your seat today to join your peers in Boston!**

dbta.com/cognitivecomputingsummit



REGISTER TODAY!

Address the ELEPHANT IN THE ROOM

Bad address data costs you money, customers and insight.

Melissa's 30+ years of domain experience in address management, patented fuzzy matching and multi-sourced reference datasets power the global data quality tools you need to keep addresses clean, correct and current. The result? Trusted information that improves customer communication, fraud prevention, predictive analytics, and the bottom line.

- Global Address Verification
- Digital Identity Verification
- Email & Phone Verification
- Location Intelligence
- Single Customer View

See the Elephant in Your Business -

Name it and Tame it!

melissa

www.Melissa.com | 1-800-MELISSA

Free Trials, Free Data Quality Audit & Professional Services.