Volume 6 Number 2 ■ Summer 2020

BDQ
BIG DATA QUARTERLY

REVERSING the 80/20 RATIO in DATA ANALYTICS

Artificial Intelligence Grows Up In 2020

Security Factors to Take Into Consideration in a Multi-Cloud World

Pandemics Happen—AI and Machine Learning Can Provide the Cures

WWW.DBTA.COM

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

Thomas Hogan Jr., Group Publisher
609-654-6266; thoganjr@infotoday

Joyce Wells, Editor-in-Chief
908-795-3704; Joyce@dbta.com

Joseph McKendrick,
Contributing Editor; Joseph@dbta.com

Adam Shepherd,
Advertising and Sales Coordinator
908-795-3705; ashepherd@dbta.com

Stephanie Simone, Managing Editor
908-795-3520; ssimone@dbta.com

Don Zayacz, Advertising Sales Assistant
908-795-3703; dzayacz@dbta.com

Lauree Padgett,
Editorial Services

Tiffany Chamenko,
Production Manager

Erica Pannella,
Senior Graphic Designer

Jackie Crawford,
Ad Trafficking Coordinator

Sheila Willison, Marketing Manager,
Events and Circulation
859-278-2223; sheila@infotoday.com

DawnEl Harris, Director of Web Events;
dawnel@infotoday.com

ADVERTISING

Stephen Faig, Business Development Manager, 908-795-3702; Stephen@dbta.com

INFORMATION TODAY, INC. EXECUTIVE MANAGEMENT

Thomas H. Hogan, President and CEO

Roger R. Bilboul,
Chairman of the Board

Mike Flaherty, CFO

Thomas Hogan Jr., Vice President,
Marketing and Business Development

Bill Spence, Vice President,
Information Technology

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
Copyright 2020, Information Today, Inc. All rights reserved.

PRINTED IN THE UNITED STATES OF AMERICA

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.

Acceptance of advertisement does not imply an endorsement by *Big Data Quarterly. Big Data Quarterly* disclaims responsibility for the statements, either of fact or opinion, advanced by the contributors and/or authors.

The views in this publication are those of the authors and do not necessarily reflect the views of Information Today, Inc. (ITI) or the editors.

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
SUMMER 2020

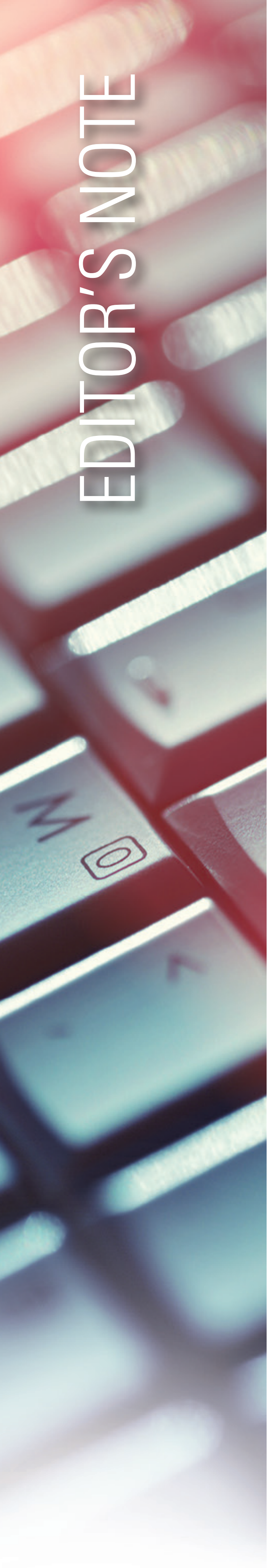
CONTENTS

editor's note | *Joyce Wells*

departments

features

columns



It All Comes Down to the Data

By Joyce Wells

Today, whether it is company leaders dealing with customer and business concerns or public health experts talking about the COVID-19 pandemic, what you hear again and again is that they are relying heavily on data. And, in this issue, we look at the range of data management challenges and opportunities.

Preparing data for analysis remains a problem. While certainly not new, it is one that is becoming increasingly difficult to deal with due to the vast quantities of data being created and stored and the variety of types and sources.

In our cover story, *BDQ* writer Joe McKendrick looks at the challenges of data prep for integration and analysis and shares insights from a wide range of industry executives on the topic. "Even the most ambitious data analytics initiatives tend to get buried by the 80/20 rule—with data analysts or scientists only able to devote 20% of their time to actual business analysis, while the rest is spent simply finding, cleansing, and organizing data," McKendrick observes. "This is unsustainable, as the pressure to deliver insights in a rapid manner is increasing."

And, there are other hurdles organizations must overcome in their quest to become data-driven. Businesses are under tremendous pressure to achieve high levels of data processing performance and scalability, contends GridGain's Nikita Ivanov, who writes about the need for in-memory computing, while Datical's Dion Cornett shares why it is critical today to include the database as a key player in DevOps processes.

Acumatica's Jon Roskill also covers the issues companies should consider as they embark on cloud migration in his article on choosing the right cloud solutions. "Unfortunately, it's fairly common for cloud-based business application vendors to carry out business practices and end-user license agreements that are misleading and border on the unscrupulous," he warns.

Another key theme in this issue is the use of new technologies, such as AI and machine learning, which offer great promise but must be carefully applied. Exploring this topic in an interview, Fractal Analytics' Suraj Amonkar offers information on proposed legislation in the U.S. for dealing with the ethical use of facial recognition technology. And, in his article on AI coming of age, FICO's Scott Zoldi adds further perspective on the need for international standards for AI use.

But despite the risks that new approaches may present, they also offer unparalleled opportunity. In their latest article, License-Fortress' Michael Corey and VMware's Don Sullivan call for a "modern Manhattan Project- or moon landing-scale effort" to leverage AI, machine learning, and the power of graphical processing units (GPUs) to solve the problem of the COVID-19. "The bureaucracy and the human trials, when combined with the discovery process which involves endless hours of iterative testing, should be the perfect target for the use of the modeling capability of machine learning and the inference capability coming from new AI algorithms," they note.

That view is shared by NVIDIA's Jim Scott, who writes about the requirement for greater unity of effort in responding to COVID-19 and future disasters. "The pandemic is exposing the fact that emergency management and public health agencies are behind the curve or underutilizing data science, open source software, and high-performance computing resources," he says.

And there are many more noteworthy articles in this issue on the changing world of data management and analytics. To stay on top of the latest trends, research, white papers, and industry news, be sure to visit www.dbta.com/bigdataquarterly, and tune, in for weekly webinars featuring industry experts at www.dbta.com/Webinars.

BIG DATA BRIEFING

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

Neo4j, a provider of graph technology, is launching Neo4j for Graph Data Science, a data science environment built to harness the power of relationships for enterprise deployments. The goal is to help data scientists leverage highly predictive, yet largely underutilized, relationships and network structures to answer unwieldy problems. https://neo4j.com

Talend, a provider of in-cloud data integration and data integrity, is bolstering its partnership with **Databricks**. With the Winter '20 release of Talend Data Fabric, including Stitch Data Loader for data ingest, Talend now supports Delta Lake, an open source storage layer. The comprehensive support enables data ingestion into lakehouse environments, where data warehouse management features are combined with low-cost storage. www.talend.com and https://databricks.com

DataStax has released DataStax Enterprise 6.8, adding new capabilities for enterprises to advance bare-metal performance, support more workloads, and enhance developer and operator experiences with Kubernetes. "DataStax Enterprise 6.8 has made significant advancements in performance, Ops management, and Cassandra workloads, but most importantly it adds a Kubernetes operator. This will help enterprises succeed with mission-critical, cloud-native deployments irrespective of the scale, infrastructure, or data model requirements," said Ed Anuff, chief product officer at DataStax. www.datastax.com

Oracle has announced the availability of the NoSQL Database Cloud Service on the Oracle Cloud as a pay-as-you-go, server-less, and fully managed service, running on the latest Oracle Cloud Infrastructure (OCI Gen 2). According to Oracle, the NoSQL Database Cloud Service can seamlessly handle schema-less JSON and fixed-schema data, in addition to pure key-value data, providing users with flexible data modeling options and the advantage of being able to rapidly develop and deploy applications without a steep learning curve. www.oracle.com

Google has launched its Memorystore platform for Memcached, an open source, in-memory data store that is a caching layer for databases. Using Memcached as a front-end store not only provides an in-memory caching layer for faster query processing, but it can also help save costs by reducing the load on back-end databases, the company says. https://cloud.google.com

Koch Industries has completed the acquisition of the remaining portion of **Infor** from Golden Gate Capital. Infor, a global provider of business cloud software specialized by industry, has been a key component of Koch's technological transformation. The acquisition brings key capabilities to Koch to accelerate digital transformation, while providing Infor with resources and industry knowledge to continue growing its expertise in mission-critical software for manufacturing, retail, and distribution, among other industries. www.kochind.com and www.infor.com

MariaDB has announced the availability of MariaDB SkySQL, a database-as-a-service for transactions, analytics, or both, and optimized with cloud-native architecture. "The universal need for accessible yet robust database services has never been higher—for around-the-clock critical operations and simplified analytics for a changing world," said Michael Howard, CEO, MariaDB Corp. https://go.mariadb.com

LogDNA, a provider of multi-cloud log management solutions, has introduced performance and usability updates to enable developers to more easily query, filter, and gain insight from their log data. "The complexity of developing, deploying, and scaling applications is exponentially more complicated today than even just a few months ago, and the amount of data even small teams deal with on a daily basis is becoming untenable," said Peter Cho, vice president of product management at LogDNA. https://logdna.com

Aerospike, a provider of next-generation, real-time NoSQL data solutions, has announced the debut of Aerospike Cloud to enable customers to build, manage, and automate their own Aerospike database-as-a-service. Aerospike says that its cloud strategy is aimed at helping customers avoid public cloud vendor lock-in and minimize the complexity and cost of migrating workloads in hybrid cloud and multi-cloud environments. www.aerospike.com

Hitachi Vantara, a subsidiary of Hitachi, is acquiring assets of **Containership**, one of the earlier providers in the container ecosystem. Containership started by providing a service that helped businesses move their containerized workloads between clouds, but moved on to focus solely on Kubernetes and helping enterprises manage their Kubernetes infrastructure. www.hitachivantara.com and https://containership.io

Improving Database Change

Database release automation provider Datical recently announced the appointment of Dion Cornett as president. Datical is built on top of Liquibase and is the primary maintainer of the open source project that enables application teams to version, track, and deploy database schema changes. Cornett, who joined Datical with more than a decade of open source leadership experience from positions at Red Hat (which was acquired by IBM) and MariaDB, talked with *BDQ* about why it is critical to include the database as a key player in DevOps processes.

Dion Cornett, President
Datical

The demand for speed in application development and deployment is increasing. How is that impacting database code change and deployment processes?

Every meaningful application is using data to some extent. If you are developing applications, you are finding that database code can be a bottleneck. We see this everywhere—in hundreds of customer and prospect conversations. According to a 2019 Datical survey, 92% of app developers reported difficulty in accelerating the release of applications because of the bottlenecks they run into with database schema changes. If you are not keeping up on the database side, then you are not getting the full benefit of speedier app development.

Why is it important to introduce automation into these processes?

You have to be particularly careful with the database. If you are building an application and there is some issue with it, you can blow up the application and build a new one. That is not a luxury you have with databases. Databases are stateful. Even something simple such as adding a column can be problematic, and blowing up the database and trying to start over can mean very significant cost to the enterprise. That is why it is more important than ever to carefully automate, reduce friction, and mitigate risk around database change deployment.

How is automation being accepted in the database management area?

People have been so successful with DevOps, which was created to develop applications faster. But someone forgot to tell the database, and so I think we are sort of at the fateful point in the evolution of IT at which people are acknowledging that this is a bottleneck and looking for solutions to make database deployments more nimble, scalable, and reliable.

What are some of the key areas you will focus on at Datical?

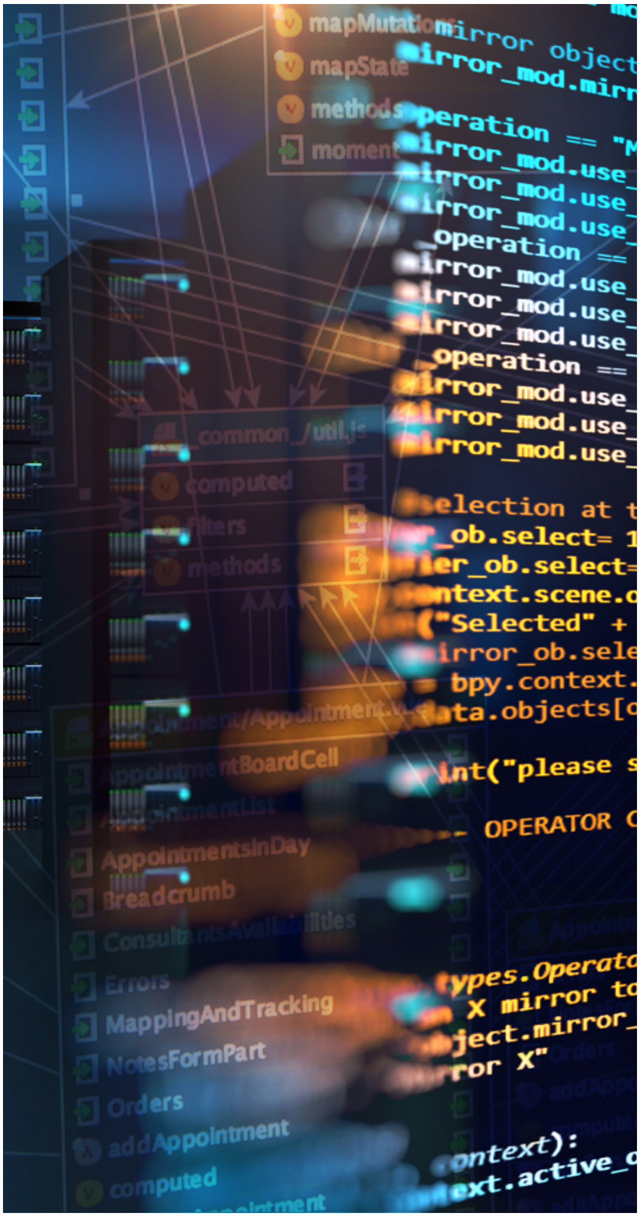
First and foremost, I want to maintain the incredible culture that the company has built here. When I was at Red Hat, I was involved in evaluating dozens of companies and had a chance to see the culture across a wide breadth of startups. People work together well, and it is a meritocracy in terms of trying to advance the mission, and then you couple that with tremendous empathy for the customer and an emphasis on creating customer value. And so maintaining those values is important.

What do you want to change?

Where I want to make an impact and drive some change is in making available a greater array of innovation to solve the challenges with the database, and the real mechanism to do this is leveraging the superior production model that represents open source—being able to engage with and be a good steward of the open source community, and it is a very robust one around Liquibase already. There have been 13 million downloads over the last 12 months, there are hundreds of contributors, and thousands and thousands of commits.

By paying attention to that community and ensuring that feedback is quickly incorporated into what we are trying to build, we are going to create better software. People that want database change automation will future-proof their software by relying on Liquibase. That is really where my expertise is coming in—working with that community to create better enterprise value.

It is more important than ever to carefully automate, reduce friction, and mitigate risk around database change deployment.



What is Datical's relationship to Liqubase?

There tends to be synergistic relationships between open source projects and strong open source companies. Linux [the Linux Foundation] and Red Hat are a great example of that synergy, as are MariaDB.org [MariaDB Foundation] and MariaDB.com [MariaDB Corp.]—and we aim to do the same thing here. Liquibase is a very powerful open source solution, and Datical has built some monetization around that, and what we want to do is have more focus on integrating those two efforts than we have in the past. The founder and the key contributors of the Liquibase community are employees of Datical. We are focusing on tightening up this integration between this strong project and the success that Datical has already had in the enterprise market to the benefit of both the project and the company.

Datical recently announced new capabilities for Liquibase, including Targeted Rollback, available for users of Liquibase Pro.

Part of the process of enhancing the relationship between business side and the community is making sure that resources can be put back into the community. One means of doing this is introducing functionality that layers on top of the Liquibase foundation.

Targeted Rollback enables developers to make a number of changes in the database, and then go after a specific change that didn't work out the way they wanted. The targeted rollback capability is an example of enhanced functionality that Datical will continue to layer onto Liquibase to provide higher value to enterprises. And, of course, with that, there is a going to be a thoroughly tested and certified version of Liquibase.

Along with a rapidly evolving, innovative community that is continually making changes to the code, it is important that enterprises have a stable version that they trust to behave for their particular requirements. You will see down the road even more capabilities around administration, ease of use, and some limited indemnification.

This interview was conducted, edited, and condensed by Joyce Wells.

Avoiding Unscrupulous Data and Business Practices Among Cloud Software Vendors

By Jon Roskill

Now, more than ever, companies are transitioning many of their critical business operations to cloud-based solutions—and ERP systems are not exempt. Because of aging infrastructure and high rates of failure, businesses will continue to switch from legacy platforms to cloud-based systems. With this change, it becomes more important for businesses to carefully evaluate their cloud solutions options to ensure they can continue to access company data and that the overall project is successful and relatively seamless.

Of the various unscrupulous practices among cloud business application vendors, procedures involving data ownership are perhaps the most worrisome.

As companies shop for a cloud-based solution, it's critical to understand that there are some major differences in business practices among cloud software vendors. Some of these practices could have deep impact on the company's overall business operations—costing more time, money, and resources in the end. For companies considering a commitment to a cloud-based business application vendor, there are a few key considerations to be aware of that will help them shop smarter and ultimately save time, avoid headaches, and get the most value for their money.

Misleading Practices

Unfortunately, it's fairly common for cloud-based business application vendors to carry out business practices and offer end-user license agreements that are misleading and border on the unscrupulous. Watch out for the following situations when evaluating vendors:

- **Hidden costs**. To draw in customers and position themselves as "affordable," a number of vendors offer large discounts or low introductory rates which end after 2 or 3 years. Price increases are not clearly detailed up-front, and the customer is made to think that they're getting a good deal for the life of their agreement. Or, the vendor may hide their inability to integrate seamlessly with a customer's business management applications or with third-party applications that are critical for companies to effectively run their business. Inevitably, it costs more to enact these integrations.
- **Implementation challenges**. The promise of a fast, easy ERP implementation may seem to be an attractive offer, but it can actually be a glaring red flag. Not because it's impossible for an implementation to be fast and easy, but because every business is unique and has different requirements; yet vendors don't always accurately represent this during the sales process. Companies should know that the implementation process is dynamic, and often difficult, and thus takes time and careful planning. Part of the chaos often comes when businesses realize that training was not part of what they paid for, and that the vendor charges extra to train users on the new system. Bottom line: Talk with the vendor to see what its approach to ERP implementation really is, and confirm with the vendor whether training is included.
- **Compliance bullying**. The SaaS end-user license agreement is often a multi-page document that vendors intentionally design to be complex and hard to understand. Vendors then contact the business (long after the sale, of course) to ensure they're adhering to the end-user license agreement. Because the end-user license agreement isn't simple and clear, users become concerned that they're out of compliance. And even though they may not be, they also may not be able to prove it. To compensate, customers often buy more licenses than they need—adding unnecessary expense to their budget. Even worse, some vendors provide license agreements as a link to a web-based document that can—and often does—change at any point of the vendor's choosing.
- **Misconceptions about data ownership.** Of the various unscrupulous practices among cloud business application vendors, procedures involving data ownership are perhaps the most worrisome. Businesses may logically assume they own their own data and have some level of control over it, but that's not always true. In many cases, companies

Jon Roskill is CEO of Acumatica (www.acumatica.com).

struggle to access their data quickly—a critical part of making informed, data-driven business decisions. Further, companies may be in for a legal tug-of-war to get control of their data in a usable format if they decide to leave their software vendor for any number of reasons, such as a merger or acquisition, rapid growth, or simply a desire to switch vendors.

- **No deployment flexibility.** Industry-specific regulations can change and may require that businesses deploy their cloud solution in specific ways to adhere to those new regulations. But, due to the end-user license agreement that is also subject to change, businesses may find out that their software deployment options are limited or inflexible.
- **Holding back on enhancements.** As companies are expected to shift, evolve, and adapt to stay relevant, it's reasonable to expect that software should too. But some ERP vendors don't even bother to invest in their own products and instead choose not to enhance or update them—sticking their customers with an outdated product that's not really a solution. As a result, the company suffers and is held back by the vendor's lack of ability or interest in delivering dynamic ERP software.

What to Look For in a Cloud Software Vendor

Despite the unfortunate existence of vendors who follow some of these bad business practices, there are many that aim to disrupt the norm and offer an approach that serves customers' interests.

For businesses that want a more customer-centric cloud vendor that helps them achieve success, they should look for business practices and end-user license agreements that offer the following elements or ones that are not materially dissimilar:

- A readily comprehensible and unchanging SaaS end-user license agreement
- A flexible, open platform for rapid integrations
- Consumption-based licensing that does not inhibit business growth
- Sustainable pricing with annual increases of no more than 3%
- ERP implementations without hidden fees
- Deployment flexibility
- Access to their data, anytime
- Consistent, 24/7 customer service
- Local business expertise
- Dual layers of support

As more companies shop for a cloud-based solution, it's critical to understand that there are some major differences in business practices among cloud software vendors.

It is imperative for businesses continuing their migration to the cloud to do their due diligence and research software vendor data practices and end-user license agreements to ensure that the transition won't cause unexpected costs, major disruptions, or loss of control. Company data continues to be among a business's most valuable assets for making informed choices, spurring advancements, and ensuring relevance in the market—and organizations must make it a priority to find a vendor that can help, not hinder, their success.

REVERSING the 80/20 RATIO in DATA ANALYTICS

By Joe McKendrick

REVERSING THE 80/20 RATIO IN DATA ANALYTICS

Even the most ambitious data analytics initiatives tend to get buried by the 80/20 rule—with data analysts or scientists only able to devote 20% of their time to actual business analysis, while the rest is spent simply finding, cleansing, and organizing data. This is unsustainable, as the pressure to deliver insights in a rapid manner is increasing. When time to answer is critical, "you can't afford to spend hours cleaning up data, nor can you waste time worrying whether your data is good enough," said Peter Bailis, Stanford University professor and CEO of Sisu.

The need to flip the 80/20 ratio is urgent. "Just 5 or 6 years ago, innovative companies were satisfied with one- or even multiple-day delays for insights from their data," said Ben Newton, director of operations analytics at Sumo Logic. "That is no longer the case. Many companies have hours, or even minutes, to respond to user behavior and market trends. The companies that are winning are basing much of their competitive muscle on the ability to leverage their data effectively and quickly."

Data teams spend "copious amounts of time finding, cleansing and organizing data," agreed Thameem Khan, general manager of data catalog and preparation at Boomi. "This creates a number of problems that hamper business progress, especially as it relates to understanding where data is, what the data says, and if the data is actually available to be used." As a result, business users may need to wait for weeks for data teams to deliver responses.

REVERSING THE 80/20 RATIO IN DATA ANALYTICS

Even the most optimistic business users are finding their enthusiasm tempered by the delays and complications that occur at the back end of their data infrastructures.

In one organization serving energy markets, its data analysts and scientists employed a platform to process data for analytics and run machine learning models that consisted of Apache Kafka, Kubernetes, and a Hadoop stack for long-term storage. "These were all complicated technologies requiring specialized and rare engineering talent," said Andrew Stevenson, CTO at Lenses.io. The analysts' screens were full of terminals running models, and if their desktops got rebooted, they had production incidents, he explained. The people charged with maintaining these systems—the platform and data engineering teams—had to keep up with trying to understand the energy markets and machine learning, as well as rewriting the models and applications, and making them deployable, noted Stevenson. Needless to say, he added, inordinate amounts of time and effort were spent "simply trying to access data, with governance, and deploy and run data applications."

TEMPERED OPTIMISM

Even the most optimistic business users are finding their enthusiasm tempered by the delays and complications that occur at the back end of their data infrastructures. "Business leaders are excited about leveraging analytics for decisioning inside of their organization, but they trigger a data landmine and report problems with accessing, preparing, cleansing, and managing data, ultimately stalling development of trustworthy and transparent analytical models," Kim Kaluba, senior manager for data management at SAS, pointed out.

In addition, business teams are slow to share their data. "They rely on data engineering teams of highly skilled engineers, who are hampered by inadequate tooling and lack of visibility into data and the complex underlying technology," said Stevenson. "They are asked to not only master complex distributed technologies but also to become an expert in every business domain. Data decays over time; if you can't discover data, process it, and quickly share your results, you are at a disadvantage."

The sheer complexity also hampers analytics efforts, especially with "many-sources operational applications from various competitors, each with their own data models and data validation rules" within enterprises, said Raghu Chakravarthi, senior vice president of R&D at Actian. For example, he related, one large auto manufacturer he worked with "had 60-plus application data sources they pulled from before performing analytics to answer a single business question. A simple operation such as 'identify customer' across operational data stores became complex when you only have a first and last name to correlate data."

In addition, variations between applications for customer hierarchical detail—some may only have one level, while others have up to 10—result in an inordinate amount of time spent cleansing, correlating, and deduping data, Chakravarthi continued. "Typically, these still result in irrelevant data, so many enterprises wrote specific business rules and cleansing logic in ETL/ELT. These rigid practices cause the 80%."

PREPPING THE ORGANIZATION

Reversing, or at least easing, the 80/20 rule requires a shift in organizational priorities—and even organizational structure. Simply opening up communication channels is a great way to start. "Teams should build processes where analysts and business stakeholders meet to discuss new questions and interpretations on a weekly basis to take advantage of new datasets and business opportunities as they arise," said Bailis.

This requires a common vision across the enterprise. "Executive mandates are not sufficient," said Horia Tipi, head of global optimization at FICO. "Stakeholders such as product, distribution, finance, and marketing teams need to see and understand the benefits of an integrated data analytics solution in order to really engage with the process." At the same time, he cautioned, this doesn't mean turning on a firehose on such diverse teams. "Marketing teams are not equipped to understand finance; while product teams often have unrealistic expectations of distribution. The solution is to have a single source of truth—a holistic data picture—that is complemented by a projection of that truth onto the screens of each individual stakeholder."

Still, in efforts to assure rapid, unimpeded flows of data-driven insights, some perspective is needed. Frequently, existing ETL processes may be enough for an ongoing data transformation flow. Organizations seek perfection too often in their drive to have the cleanest data possible, down to the transaction level, even if the data is only being used for strategic purposes such as trend analysis, observed Glen Rabie, CEO of Yellowfin. But the effort to achieve perfection may far outweigh the benefits, he noted. "Organizations need to be more efficient and prepare their data to the level that supports the detail of analysis they need to do."

Reversing, or at least easing, the 80/20 rule requires a shift in organizational priorities—and even organizational structure.

Instead of devoting too many resources to data preparation, organizations should focus on how work is allocated among data teams, Rabie continued. "Sometimes, analysts are not actually data 'analysts'; rather, they are data preparers. As a result, they feel more comfortable working with the data than analyzing it and conducting the business analysis that the organization needs. Ensure that the right data specialists are assigned to the roles in the analytic process for which they have both the skills and inclination." Rabie also urged enterprises to provide their data teams with the right resources. "Given the initial effort to prepare data, if the organization does not provide sufficient analysts, prepared data may not get the analysis it deserves." Instead, he said, organizational priorities may push the analytics team to the next dataset. "Companies should appoint more business analysts to analyze the data that has been prepared."

ONE VIEW OF THE DATA

Bringing teams together across the organization requires their involvement in a truly holistic data strategy—"understanding what data they have, understanding that all data is not equal, and then linking their data initiatives to their corporate initiatives," said Andreas Wesselmann, senior vice president of HANA and analytics, data management and platform, for SAP. Most companies have data in many systems, in ERP and other hybrid applications, and flowing in from IoT, social, and external sources, he explained. "The data is now multifaceted and in many cases not fully connected. Businesses must have a solid data management strategy, with data integration and orchestration at the top of the IT priorities list."

The name of the game is also focus, applying data exactly where it is needed across the business. For example, in operations, "much of the data is first and foremost used in process control and real-time operator insight and forensic analysis," said Richard Beeson, CTO of OSIsoft. "On the maintenance side, the data serves to schedule people, equipment, and services in the most efficient manner possible, taking into account production schedules and customer commitments which are typically in another core data source."

Before implementing any kind of technical tool, "organizations must realize the business value of the data," said Dan Wu, privacy counsel and legal engineer at Immuta. "Internal champions can identify specific use cases and form a cross-functional coalition—including governance, risk, and compliance—to rally behind them. With this coalition, a team can identify solutions to balance utility and safety."

It is important to start with a strong, intelligent management foundation that can extract data intelligence through data integration, orchestration, metadata management, connectivity, and AI and machine learning services, supporting on-premise and cloud deployments, said Wesselmann. "Focus on collaboration between various roles such as data architects, data integration experts, developers, and even data scientists. This collaboration supports DataOps initiatives for using tools within an agile framework, enabling data preparation and automation of data workflows, and provides transparency across the data roles."

TOOLS AND PLATFORMS

A cohesive organizational approach also paves the way to adoption of current architectural methodologies involving DevOps and agile protocols. "Architectural trends like containerization and microservices provide the opening for moving fast and adapting, but those architectures are only successful if the organization running them has embraced the decentralized, bottoms-up mentality that created them in the first place," said Newton. Service APIs can also supplant mechanisms such as ETL custom logic to speed up and enable teams to build and use common dimensions. This "frees up analysts and data scientists embedded within each business unit to focus on collecting detailed atomic fact data to answer things like, 'Who is my most valuable customer?' or, 'Which of my regions are underperforming?'" said Chakravarthi.

Emerging tools and platforms can help address challenges with data engineering, which is still characterized by manual tasks. "Solutions that can help accelerate this process, such as data lakes technology, can organize data from multiple sources by building relationships, removing duplicates, and automatically refreshing data," said Alex Ough, senior CTO architect at Sungard Availability Services.

"It's easier than ever to collect and aggregate this data with a wide array of flexible cloud data warehouses and cloud-native data pipeline tools," said Bailis. However, he added, "there's a premium on every analyst's time and attention. It just takes too long to properly diagnose and assess the impact of every potential change. To combat this imbalance between cheap data and expensive people, teams can look to adopt platforms

REVERSING THE 80/20 RATIO IN DATA ANALYTICS

To address the imbalance between cheap data and expensive people, organizations can use platforms that rapidly diagnose changing KPIs and suggest next steps.

that augment their ability to rapidly diagnose changing KPIs and recommend next steps collaboratively."

Even existing formats can be repurposed to achieve a simpler architecture. Stevenson advocates a DataOps approach that supports a data mesh architecture, which provides for "discoverability, visibility, and governance backed by tooling that is supported by SQL, a ubiquitous data language. This simplifies developing and managing data-intensive applications running on the data infrastructure that is now a commoditized technology. When this happens, everyone can contribute."

The challenge, of course, is working around legacy infrastructure, which may be too expensive to rip and replace. Bry Dillon, vice president, cloud, channels, and community, for OSIsoft, calls for "purpose-built operational systems that can access the data" without "disrupting critical functions of systems which could be decades old." He also recommends "systems that can normalize and contextualize the data to give as much color and depth to the data as possible. Some of this can be gathered from the source systems themselves, but many of these are quite old, and it is best to leverage the understanding by the people who operate them." AI and machine learning technologies can also play a role, especially if they "are more suited for operational datasets so they can get a faster time to value."

Business data delivery can also be accelerated through self-serve data management platforms. "These platforms come in many forms—delivered on cloud or on-premises," said Brian Sparks, product manager for data integrity at Vertex. "Some focus on simple user interfaces, while others focus on offering the most functionality. What they have in common is that they typically offer some form of ETL, some form of orchestration allowing users to set up data pipelines for repetitive data flows, and they have the ability to ensure the quality required for your business."

Data preparation tools on the market can make the difference, as can IT analytics tools. "IT teams have so far been facilitators of analytics applications for business teams, but have been slow to adopt analytics for analyzing their own processes and evaluating their performance," said Rakesh Jayaprakash, product manager for ManageEngine. "IT analytics, which involves analyzing data from IT applications, is gaining more traction as more companies realize the value of data that IT applications hold. IT teams should identify and tap into application and machine data that can be used to make business-critical decisions. For example, analyzing application logs can help identify sections of code that could potentially be causing the application to respond slowly."

ADVICE

Ultimately, the way to measure the success of efforts to reverse the 80/20 ratio is to measure the speed at which businesses can access and leverage data insights. "Always keep business benefit in mind and embrace the perspectives of all the stakeholders—get them to understand and appreciate things from their unique perspective," Tipi advised. "Find your key differentiating features as a business, and do everything you can to replicate and expand them while eliminating friction. Eliminate data friction, analytic friction, operational friction, and measurement and capital reinvestment friction. The tools exist to reduce them all; you only need to ask the right questions to find them."

Wesselmann urged the approach of taking on an executive sponsor to help bring about organizational commitment to relieving the onerous tasks faced by data teams. "Tie it to corporate initiatives and secure an executive sponsor," he advised. "For example, if the business objectives are to improve revenue, cut costs, or provide a different service, there is always key data required to execute on that vision. This gives you an idea of where the data management strategy should start and helps you select the right use cases." And don't try to boil the ocean, he added, but instead start small, show success, and then move on.

In addition, while technology is an important tool to enable faster data delivery, moving to new approaches, such as AI and machine learning, "does not relieve companies of their need for fundamental data management," said Dillon. "This goes beyond collecting data and includes ensuring it is available to the appropriate tools and appropriate experts. The success of any new technology will be highly dependent on the corresponding people and processes needed to leverage it."

Keep in mind that "the definition of a database has completely changed over time," said Khan. And as that has happened, the data and the required skill sets of DBAs have also changed, he noted. "Database admins need to think of different technologies for their applications and really understand data well if they want to make the most of it."

melissa®

The Power of Clean Data

Today, it is well-understood that data is the foundation for initiatives that provide enormous benefit to organizations. And, whether it's used for analysis, targeted marketing campaigns, fraud detection, or some other purpose, accurate and deduplicated data is critical.

However, 91% of businesses suffer from data errors, the most common of which are incorrect and inaccurate data, outdated data, missing information, and duplicate data.

What's the problem? Prepping enormous volumes of data is time-consuming and tedious. It's commonly estimated that data analysts spend roughly 80% of their time gathering, managing, and cleansing data, and only 20% of their time actually doing the data-driven activities their jobs call for. And, the fact that reams of data is flowing into organizations from more sources than ever before is only exacerbating the problem.

MELISSA DATA PROFILING AND CLEANSING TOOLS

Using cutting-edge technology and multi-sourced global reference data, Melissa provides the solutions that support Know Your Customer (KYC) initiatives, reduce costs, and improve fulfillment. More than simply ensuring that you have quality data for analysis, they improve the efficiency and expediency of data preparation.

Two complementary tools from the Melissa suite of data quality products form the bedrock of data analysis—Profiler and Generalized Cleanser.

PROFILER

The first step in improving data quality is to profile your data to gain insight into its usability. Profiler provides information on your data tables—whether it is customer information or another type of data—such as financial data or business information. The tool generates simple-to-advanced profiling information, including basic data statistics and details to identify data errors at their source. You can then monitor the performance of your source data over time to specified requirements of pre-set limits. This enables you to enhance data management and data warehousing efforts by identifying weak points in your data, optimizing data quality over time by continuously reviewing data, and enforcing business rules on incoming records to maintain data standardization.

GENERALIZED CLEANSER

After using Profiler, organizations can take the output and run it through Melissa's Generalized Cleanser tool to standardize phone numbers, email addresses, physical addresses, and other types of fields as far as case, punctuation, or other considerations. Profiler makes it possible to build data cleansing scripts for a wide range of data errors and inconsistencies. It combines six operations that allow you to cleanse data and save operations (simple or complex) for future projects. The tool is used in many different industries, giving you the flexibility to standardize and validate inventory lists, correctly format data, and much more. The Generalized Cleanser allows you to gain greater control of your data, save time and resources, and create customized rules to standardize data. It enables you to cleanse any type of data and achieve a higher standard of data quality for integration, warehousing, and analytics.

PART OF A COMPLETE PACKAGE

In addition to the Profiler and the Generalized Cleanser tools, you can advance your data prep processes even further using Melissa's MatchUp or Personator tools.

MatchUp finds all the common data elements between multiple lists and can use suppression to find data unique to each individual list. The tool uses more than 16 advanced fuzzy matching algorithms and deep domain knowledge to find even the hardest-to-detect duplicate records.

In addition, Personator taps a multi-sourced dataset containing billions of records to validate each element of a U.S. or Canadian contact record and match names to addresses to verify identity. Personator also enriches contact records by filling in missing contact information, adding current addresses for customers and prospects that have moved, and adding detailed consumer demographics.

SUPPORTING ANALYTICS, PERSONALIZATION, AND GOVERNANCE

Using these tools, you can identify data streams with particularly poor quality that need to be dealt with real-time point-of-entry data quality controls, and then monitor the streams over time to verify improvement.

Moreover, as we face a new frontier as data privacy regulations and master data management becomes a greater concern, the combination of solutions supports compliance-related initiatives. If you needed, for example, to show a customer how their data is being used, or identify their information and purge it at their request, you would be able to do that.

The power of clean data is clear. With the amount of data that is streaming into organizations, a solid data quality and governance plan is no longer a nice-to-have; it is a must. The flexibility of these tools can help any organization wrangle their data, understand its weaknesses and how it needs to be cleaned, and ultimately make it fit-for-purpose for their own unique requirements.

For more information, visit www.melissa.com or call 1-800-MELISSA.

INSIGHTS

Artificial Intelligence Grows Up in 2020

By Scott Zoldi

For several years, AI has been the *enfant terrible* of the business world, viewed as a technology full of unconventional and controversial behavior that has shocked, provoked, and enchanted audiences worldwide. That's all going to change. In 2020, AI will grow up, encountering new demands in the areas of responsibility, advocacy, and regulation.

Expect to see the rise of international standards to define a framework for safe and trusted AI in 2020 because regulation keeps companies honest.

Move Over Ethical AI, It's Time for Responsible AI

For a few years, I've been working on new data science patents, pushing AI technology to be more defensive, explainable, and ethical. Driven by the ever-rising onslaught of new AI applications—coupled with the fact that regulation around AI explainability, transparency, and ethics is still emerging—there will be higher expectations for responsible AI systems in 2020.

For example, if a medical device—such as a heart pacemaker—were rushed to market, it could be poorly or negligently designed. If people using that device were harmed, there would be liability. The company providing the device could be sued by individuals or groups if a lack of rigor and/or reasonable effort were proven.

Similarly, there will be a more punitive response for companies that consider explainable, ethical AI to be optional. "Oops! We've made a mistake with an algorithm and it's having a harmful effect" will no longer be the basis of interesting news stories about AI gone rogue, but instead a call to action.

In 2020, AI insurance will become available with companies looking to insure their AI algorithms from liability lawsuits. Using blockchain or other means for auditable model development and model governance will become essential in demonstrating the due diligence necessary in building, explaining, and testing AI models.

Get Ready for the AI Advocacy Fight

Can AI be harmful? Think about someone being denied rightful entry to a country due to inaccurate facial recognition. Or someone being misdiagnosed by disease-seeking robotic technology. Or someone being denied access to a loan because a new type of credit score on non-causal features rates them poorly. Or someone being incorrectly blamed as the cause of an auto accident by the insurance company mobile app loaded onto that driver's phone.

There are already numerous ways that people are treated unfairly in our society, with advocacy groups to match. With AI advocacy, there may be a different construct, because consumers and their advocacy groups will demand access to the information and process on which the AI system made its decision. AI advocacy will provide empowerment, but it also may drive significant debates between AI experts to interpret and triage the data, model development process, and implementation.

AI advocacy will move from being a radical idea to a commonplace function of the growth of AI adoption.

AI Regulation Matters

Industry leaders hold a negative, blanket view that government regulation is an innovation inhibitor. This couldn't be further from the truth with AI. Regulators and legislators are trying to protect consumers from the negative effects of technology (in reality, the human creators that misuse AI/machine learning) through vehicles such as the EU's GDPR, California's CCPA, and other regulations. However, often, demands are being made of technology about which there is little understanding.

Granted, at the opposite end of the scale, there are the companies that clasp their metaphorical hands and say, "We are ethical; we will do no evil with your data." But, without a standard of accountability, we'll never know for sure if this is actually the case. For both extremes (and those in between), we will see the rise of international standards to define a framework for safe and trusted AI in 2020 because regulation keeps companies honest. Hopefully, we'll also see AI experts support and drive regulation of the industry, ensuring fairness and inculcating responsibility.

Make AI Experience Count

As AI grows to be a pervasive technology, there is little trust in the morals and ethics of many companies that use it. As a data scientist, that is disheartening. However, as modern society discovers more about the damage that can be done by misuse of AI, it's clear that experience—not the mantra of "Move fast and break things"—matters. It's time for AI to grow up in 2020.

Scott Zoldi is chief analytics officer at FICO (www.fico.com).

YOUR CONNECTION TO THE INDUSTRY
BDQ
BIG DATA QUARTERLY
database
TRENDS AND APPLICATIONS

Ethical AI:

Q&A With Fractal Analytics' Suraj Amonkar

Suraj Amonkar

The goal of the bill would be to ensure facial recognition does not produce bias or inaccurate results or 'create a constant state of surveillance of individuals in the United States that does not allow for a level of reasonable anonymity.'

Suraj Amonkar is Fellow, AI @ Scale, machine vision and conversational AI, at Fractal Analytics, an AI and analytics company. Recently, he shared his views on the proposed legislation and the issues it addresses.

The Ethical Use of Artificial Intelligence Act was recently introduced by U.S. Senators Cory Booker (D-N.J.) and Jeff Merkley (D-Ore.) with the goal of establishing a 13-member congressional commission that will ensure facial recognition does not produce bias or inaccurate results. Suraj Amonkar, Fellow, AI @ Scale, machine vision and conversational AI, at Fractal Analytics, recently shared his views on the proposed legislation and the issues it addresses.

What is the goal of the Ethical Use of AI Act?

The Ethical Use of AI Act is aimed specifically at using facial recognition as an AI technology. According to the bill text, it is being enacted because facial recognition is being marketed to police departments and government agencies. The technology has a history of less accurate performance for people of color and women, and facial recognition can chill First Amendment rights if used to identify people at political speeches, protests, or rallies. The goal of the bill would be to ensure facial recognition does not produce bias or inaccurate results or "create a constant state of surveillance of individuals in the United States that does not allow for a level of reasonable anonymity."

Why is this important?

I think the Act has been introduced to regulate use of facial recognition for specific use cases such as reducing crime rates, aiding forensic investigations, finding missing people and victims of human trafficking, identifying perpetrators of crime on social media and tracking them. It is important to bifurcate the use of facial recognition so that it does not muddle with basic rights of individuals including, but not restricted to, their privacy rights.

What is the problem with facial recognition?

The problem is not with facial recognition as a technology, but there are issues in the way it is implemented and used. For instance, some facial recognition technologies have encountered higher error rates when seeking to determine the gender of women and people of color. The use of this technology causes concerns about how much people are being watched and if hackers can access this data, causing more harm than good. There is also a growing concern with the possibility of misidentifying someone and leading to wrongful convictions. It can also be very damaging to society by being abused by law enforcement for things like constant surveillance of the public.

What are the risks of facial recognition misuse?

Some of the risks include misidentifying people, wrongful convictions, misuse of private data, stalking, identity fraud, and predatory marketing. A camera in a retail store could do more than identify theft—it could use your face to link your online and offline purchasing activity, leading to intrusion of privacy and what some call predatory marketing.

How is this being received in the AI community?

It would not be wise of me to speak on behalf of the entire AI community. As a proponent of the technology, I do not believe in a complete moratorium. By limiting its use or delaying, it means we are letting go of all the benefits it brings. Of course, there is a need to regulate this technology, but that's the case with every other technology area.

The onus is equally on providers of this technology as well. We have seen some examples where the technology implementation has thrown up huge errors, leading to bias or inaccurate results. This makes it especially important that all tech companies continue the work needed to identify and reduce these errors and improve the accuracy and quality of facial recognition tools and services.

DATA SUMMIT CONNECT

ONLINE EVENT!

The Premier Data Management and Analytics Conference Is Going Online!

dbta.com/datasummit

JUNE 9–11 2020

We can't get together in Boston in May, but you can still see where the world of big data and data science is going and find out how to get there first by joining us for *Data Summit Connect*, a free series of video webinars that will run from June 9 to 11.

We're also offering two online workshops on June 8 for just $199 each. These 3-hour, in-depth workshops offer training from expert instructors that you can't get anywhere else.

RESERVE YOUR SEAT TODAY!

CONNECT

f in #DataSummit

ORGANIZED AND PRODUCED BY Information Today, Inc.   FROM THE PUBLISHERS OF BDQ BIG DATA QUARTERLY database TRENDS AND APPLICATIONS

BIG DATA BY

NEW REQUIREMENTS SPUR DATA QUALITY AND DATA INTEGRATION

A combination of factors are heightening the need for high-quality, well-governed data. These include the need for trustworthy data to support AI and machine learning initiatives, new data privacy and data management regulations, and the appreciation of good data as the fuel for better decision making.

Multiple Data Sources, Governance, and High Volume Are Top Data Quality Challenges

1. The top 3 challenges companies face when ensuring high-quality data are multiple sources of data **(70%)**, applying data governance processes **(50%)**, and volume of data **(48%)**.

2. About three-quarters (78%) of companies have challenges profiling or applying data quality to large datasets.

3. 29% say they have a partial understanding of the data that exists across their organization, while **48%** say they have a good understanding.

Source: Syncsort's 2019 "Enterprise Data Quality Survey"

The Data Prep Market Is Growing

The **global data prep market, including tools for data curation, data cataloging, data quality, data ingestion, and data governance—whether on-prem or in the cloud—is growing. The market** is expected to rise from its initial estimated value of $3.38 billion in 2018 to an estimated value of **$21.48 billion by 2026,** registering a compound annual growth rate of 26% in the forecast period to 2026.

Source: "Global Data Prep Market—Industry Trends and Forecast to 2026" from Data Bridge Market Research

There are **4 key drivers** for the increase:

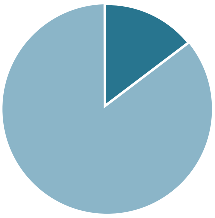
1. The rising need for adhering to regulatory and compliance requirements
2. The need for on-time qualified data
3. The benefits of streamlined business operations
4. The fact that data prep tools help companies in predictive business analytics

THE NUMBERS

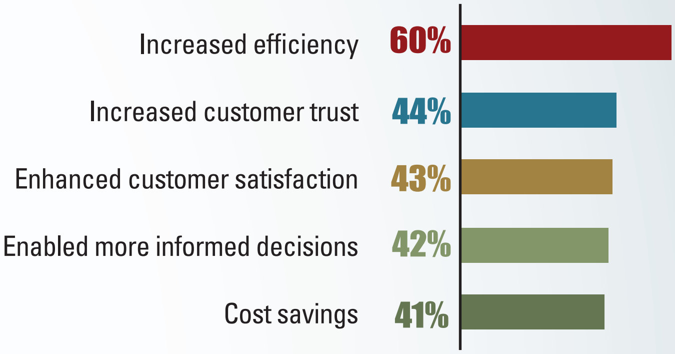
The Need for Data-Driven Insights

Maintaining a competitive business edge depends on the ability to leverage accurate and reliable data to make informed and strategic decisions.

85% of organizations see data as one of the most valuable assets to their business.



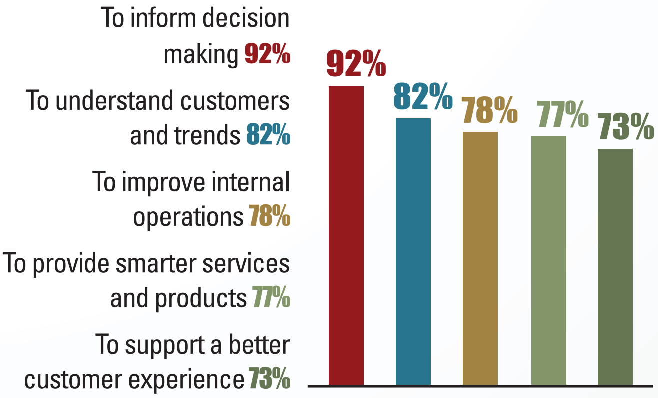
Key reasons for having a strategy to maintain high-quality data are:



Source: "2020 Global Data Management Research," produced by Insight Avenue for Experian

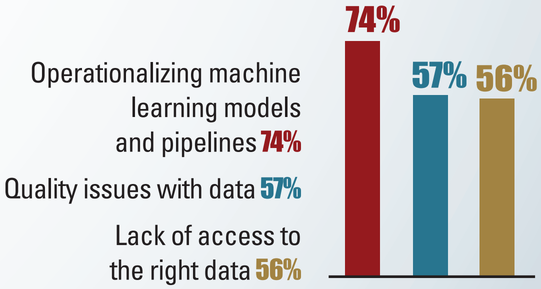
Top business uses of data driving organizations' data strategies

Although there are many business objectives driving data strategies, the most frequently mentioned are improving the decision making of end users and uncovering customer preferences and patterns.



Top challenges being encountered with your machine learning projects

The adoption of machine learning—which has been described as "getting computers to act without being explicitly programmed"—is increasing. This is due to the vast and rapidly growing volumes of data and the associated challenges of finding value from that data. However, fundamental data issues underlying machine learning models are problematic, pointing to challenges with data quality and access to the right data.



Source: "Profiling the Data-Driven Business, 2019," produced by Unisphere Research, a division of Information Today, Inc., and sponsored by Pythian.

The In-Memory Computing Landscape in 2020

By Nikita Ivanov

As companies have evolved toward digital business models and undertaken digital transformation initiatives, they have increasingly faced two challenges. First, the data they need to drive their real-time business processes is typically spread across multiple, siloed datastores. Second, their existing applications often cannot scale to address the increase in end-user demands for real-time engagement.

Thanks to the relatively low cost of RAM today and the availability of open source solutions, in-memory computing technologies have progressed dramatically over the last few years, becoming a foundation for accelerating and scaling real-time business processes in support of the range of digital transformation and big data/fast data initiatives. As we move through 2020, in-memory computing will be particularly important in enabling data centers to accelerate the use of the following new strategies for supporting real-time business processes and analytics:

- Highly performant digital integration hub (DIH) architectures for accessing data in real-time from multiple, siloed datastores and data streams. The datastores may span on-premise and cloud databases as well as SaaS datastores.
- Hybrid transactional/analytical processing (HTAP), also called hybrid operational/analytical processing (HOAP), to enable transaction processing and analytics on the same system.
- Creating end-to-end business processes based on a combination of mainframe data and other datastores that leverage DIH architectures.
- Combining in-memory computing platforms with non-volatile RAM to fine-tune the balance between optimal performance, data protection, and overall system cost.

In-memory computing technologies have become a foundation for accelerating and scaling real-time business processes.

DIH Architectures

Consider a financial services institution that offers a variety of products, including mortgages, credit cards, and core banking services. As customers interact with the company about one of its products, it may want to take advantage of cross-sell and upsell opportunities. To offer those additional products, the institution needs to be able to access a customer's data from all the disparate business units to understand their current situation and which additional products might be a fit for them. To ensure a positive experience, the required data must be processed and a decision on which new products to offer must be made in real time.

Today, most enterprises will struggle to implement such real-time business processes because the required data may reside in multiple internal and external systems as well as in separate data warehouses and data lakes.

To aggregate and process data in support of end-to-end business processes, developers must typically make API calls to multiple data sources for every action. However, waiting for data to be retrieved from disparate data silos can be slow compared to the real-time experience demanded by the web visitor. The company may also face limitations on the number or types of API calls it can make on the datastores. Further, once the data is aggregated, it must be processed in real time. The complexity of overcoming these challenges makes it extremely difficult or even impossible for many organizations to create the real-time business processes they need.

An approach to solving this challenge is what Gartner calls a DIH architecture. A DIH architecture creates a common data access layer for aggregating and processing data from multiple on-premise and cloud-based sources and streaming data feeds. The DIH architecture's API services layer automatically synchronizes changes made by the applications consuming the data to the back-end datastores.

A DIH architecture can enable multiple customer-facing business applications to access a single view of the aggregated data and process the data at in-memory speeds without movement of the data over the network. By caching the required data, calls to the siloed datastores can be reduced using change data capture to periodically update the datastore, and delays caused by waiting for data to be retrieved from some datastores can be eliminated. This architecture can also eliminate the need for business applications to make direct API calls to each datastore every time a new customer interacts with the company.

At a recent industry conference, 24 Hour Fitness described its use of a DIH architecture. The system caches data from the company's SaaS billing system, updating the cache every 15 minutes. This cached data can then be accessed simultaneously and in real

Nikita Ivanov is CTO of GridGain Systems (www.gridgain.com).

A DIH architecture creates a common data access layer for aggregating and processing data from multiple on-premise and cloud-based sources and streaming data feeds.

time by multiple business applications, eliminating the need for an API call to the SaaS billing system for each operation. Application logic can also be deployed on the server nodes of the in-memory computing platform cluster, enabling parallel processing of the cached data with no data movement over the network. This architecture provides customers and business analysts with fast access to the information they need, and the company can easily implement additional functionality.

Here are other examples of the use of DIH architectures:

- Combining current purchase data with inventory levels and webpage visit data to enable retailers to power real-time recommendation engines
- Creating a common data layer for IoT platforms, so companies can process and query current and archived sensor data to drive real-time system awareness
- Combining and processing travel reservation data from many travel providers in real time to instantly present visitors to a travel reservation website with multiple travel options

To create their DIH architectures, many companies are turning to in-memory data grids, which can cache data from a variety of underlying datastores, including databases, SaaS applications, and incoming data streams. Application code can then be deployed on the servers in the in-memory data grid cluster to perform parallel processing with no data movement across the network. With the data cached in memory for instant access and with collocated computing on the cluster nodes, line-of-business applications, including consumer-facing websites and back-office applications, can implement real-time business processes that would be impractical or impossible to achieve without a DIH.

Hybrid Transactional/Analytical Processing

For decades, companies have relied on bifurcated data infrastructure to meet their performance SLAs. Separating OLAP systems from OLTP systems ensures analytics won't impact operational systems. The downside of this approach is that it requires a time-consuming ETL process to periodically copy data from the OLTP system to the OLAP system. Today, the delays inherent in periodic ETL are an obstacle to driving real-time business processes based on OLAP real-time analytics.

In-memory computing is playing a vital role in solving this challenge, providing the speed and scale necessary for HTAP. HTAP enables a system to perform real-time analytics on a company's operational dataset without impacting performance. By running real-time analytics on the operational data in RAM with massively parallel processing (MPP), an in-memory data grid can deliver the performance at scale required for HTAP, providing both real-time transactional and analytical processing using the operational datastore. HTAP also has the long-term cost benefit of reducing the scope or eliminating the need for a separate OLAP system.

In-Memory Computing and Mainframes

The use of mainframe computers remains pervasive in the financial services industry. The IBM Z platform, for example, is used by 92 of the world's top 100 banks, all top 10 insurance organizations, and 64% of the Fortune 500. However, firms that rely on a mainframe for transaction processing may still want to implement a DIH to create real-time business processes based on a combination of operational and historical data.

In-memory computing platforms that have been optimized to run on a mainframe enable these firms to take advantage of both DIH architectures. For example, a company can use an in-memory computing platform to create an SQL-driven data access layer that runs on the mainframe. The DIH can connect to data sources, including operational databases running on or off the mainframe, as well as a portion of the data held in the data lake. Processing can then be run on the combined dataset held in the in-memory computing platform engine at the heart of the DIH to drive real-time business processes. This can enable a financial institution, for example, to obtain a 360-degree customer view based on analyzing all the data for a particular customer stored in the firm's operational database and historical data lake to drive upsell or cross-sell programs or to drive seamless customer interactions across all the firm's customer touchpoints.

To create their DIH architectures, many companies are turning to in-memory data grids, which can cache data from a variety of underlying datastores, including databases, SaaS applications, and incoming data streams.

Non-Volatile Memory

One of the most exciting emerging developments in memory technology is in the area of non-volatile RAM, or NVRAM. Nearly all computing today still relies on separating very fast volatile memory (RAM) used for running applications from slower, non-volatile memory (hard disks, SSDs, etc.) used for storage. The challenge with this approach is that applications and data need to be loaded into RAM from storage each time a computer is turned on, and all the data in RAM is lost each time the computer is turned off. This means there is always a potential for data loss in the event of a system crash, and the larger the database, the longer users must wait in the event of a restart.

Today, as data requirements scale to terabytes and petabytes, the situation has become critical. Strategies for preventing data loss have become more cumbersome and expensive, as it takes hours or days for data to load into RAM on even the most powerful systems. However, new non-volatile RAM memory technology, such as Intel Optane, has become generally available, and as the price begins dropping, NVRAM will become a vital solution for protecting data and accelerating large-scale database systems. NVRAM can also be combined with in-memory computing platforms to allow companies to unite in-memory speed with the lower cost and durability of non-volatile storage in order to fine-tune the balance between optimal performance, data protection, and overall system cost.

The In-Memory Computing Platform

The key technology powering many of these computing advances is the in-memory data grid, which is typically delivered as part of an in-memory computing platform. An in-memory computing platform pools the available RAM and compute of a server cluster, which can be easily scaled by adding nodes to the cluster. By maintaining data in RAM, the platform eliminates the delays caused by accessing data stored in disk-based databases. Further, by utilizing the MapReduce programming model to distribute processing across the cluster, the platform also provides MPP and can minimize or eliminate movement of the data in the grid across the network prior to processing.

This combination can improve application performance by up to 1000x and create a common high-performance data access layer that makes data from many datastores available in real time to many applications. In-memory computing platforms can typically be run as a standalone in-memory database or as an in-memory data grid inserted between an application and its existing data layer.

Businesses undergoing digital transformations are under tremendous pressure to achieve unprecedented levels of data processing performance and scalability to support their real-time business processes. At the same time, we are witnessing an extraordinarily rapid evolution of in-memory computing technologies designed to support these business initiatives. This means that developers and system designers must fully understand the potential of in-memory computing to impact their business models, and they must pay close attention to the latest developments. As with all new technologies, internal teams should rely on third-party experts if they don't have the expertise for such evaluations.

Security Factors to Take Into Consideration in a Multi-Cloud World

By Patrick Lastennet

The cloud is here, and as enterprise cloud strategies continue to mature, we're starting to see them take more of a multi-cloud approach, where they can capture some parts of what they need from one cloud provider and other parts of it from another, while also keeping more sensitive data on-premise.

With a multi-cloud strategy, businesses are finding that they can gain scalability, resilience, and significant economic savings. However, this approach requires businesses to transition their architecture to a much more complex and decentralized model, which makes managing the security of the entire environment extremely challenging.

Let's review the new factors that businesses must take into consideration as they develop security strategies in a multi-cloud world.

Multi-cloud environments require businesses to focus on the fundamentals of keeping all of their data safe, instead of just stopping hackers, malware, and other sophisticated attacks at the perimeter.

Consideration #1: The Multiplicity of Environments You Need to Protect

Leveraging a multi-cloud environment means a far more fragmented architecture. For example, a business could tap a number of different cloud providers such as Amazon Web Services and Google Cloud Platform, several SaaS platforms—including Salesforce and Microsoft Office—as well as their own proprietary data centers for extra sensitive data storage.

While this strategy provides businesses with an amazing capability to build on top of, it requires maintaining ongoing, high-level security for all applications used in the business, which is especially difficult due to a lack of visibility across a much wider attack surface. As such, businesses will need to identify what components their new environment consists of and make sure that they deploy a security strategy that mirrors the dynamic and distributed nature of this infrastructure.

Consideration #2: The Perimeter Is Obsolete

In the past, enterprises have prioritized strong perimeter defenses for their on-premise architectures and could expect that the applications and workloads behind them would be secure. But in a multi-cloud environment where data is distributed across a larger landscape, much of the visibility and control that they were used to is lost. There is no perimeter in the cloud. Therefore, businesses are left to assume that all of their data can be now be accessed by unwanted parties.

Multi-cloud environments require businesses to focus on the fundamentals of keeping all of their data safe, instead of just stopping hackers, malware, and other sophisticated attacks at the perimeter.

Consideration #3: Interoperability Between Tools Offered by Cloud Providers Doesn't Exist

Since a multi-cloud environment is a combination of public cloud, private cloud, and on-premise environments, the data within each of these architectures needs to be able to communicate with each other to deliver services across the entire enterprise and provide true value. While IT teams do the challenging work around shared processes, APIs, containers, and data models to enable this communication, the security piece is not quite there yet.

The major hyperscale cloud platforms all invest heavily in security, but data-in-transit over the public internet can still be intercepted. As soon as any enterprise data touches the internet, it's fundamentally at risk—whether from distributed denial of service attacks, malware infections, or other threats. With enterprise IT teams often short of time, money, and specialist skills, network vulnerabilities can creep in all too easily—whether intentionally or accidentally. Businesses need to ensure that all of their data is protected, regardless of where it is stored and what other data it is communicating with.

All Things Considered

As leveraging a multi-cloud environment becomes increasingly critical to business performance, enterprises cannot undermine the protection of their assets being stored within it. Data-centric security practices must be taken into consideration at the onset of their environment's progression.

In addition to ensuring that they have private and secure connections to the various cloud platforms that they're leveraging, fortifying encryption keys will become a necessary layer of added security. With a variety of key management options—hardware security models (HSMs) on-premise, multiple HSMs located in cloud provider data centers, or individual HSMs for each of the cloud providers—enterprises can work with a managed service that provides them with the perfect solution for key encryption management to ensure their entire multi-cloud environment is protected at the highest level of security.

Patrick Lastennet is director of enterprise, Interxion (www.interxion.com).

Advancing Data Science for Emergency Management and Public Health Response

The World Health Organization (WHO) declared a pandemic for the novel coronavirus disease (COVID-19) in March 2020. Private industry across the world has stepped up to help public agencies on the front lines of the pandemic to fast-track vaccine development and provide tools to inform the public on how best to protect themselves and their communities.

The pandemic is revealing gaps in our critical infrastructure security, supply chain fragility, and the utilization of modern technologies to mitigate and recover communities globally. Although advanced technology platforms have been used by large international corporations, the pandemic is exposing the fact that emergency management and public health agencies are behind the curve or underutilizing data science, open source software, and high-performance computing resources.

Government and industry have an opportunity to work together on educating emergency management and public health agencies on these capabilities, providing a unity of effort in identifying, selecting, and deploying the best tools to solve the most pressing challenges of our time—whether it is this pandemic, climate change, or some other major disaster.

Resolving the technology gaps in government and industry—particularly among small businesses—regarding emergency management and public health requires transformation in education and policy as well as agency culture. Coast Guard Commandant Thad Allen, who led the U.S. response to Hurricane Katrina, has stated that complex crises cannot be addressed without collaboration with the public and other private organizations. Admiral Allen uses the term "unity of effort" to summarize what is required to overcome whole-of-society disasters.

To achieve unity of effort, governments can make basic policy changes to advance data science and innovative technologies for emergency management and public health response and preparedness. There is already existing reform and legislation that require full implementation. The Open Government Data Act, signed by President Donald Trump in 2019, requires U.S. government agencies to maintain and publish a comprehensive data inventory of all data assets. This data inventory, once implemented, will transform government sources of data from documents and siloed databases into open data, using machine-readable formats, for use by any agency or private company.

In emergency preparedness, sky color is an often-referenced metaphorical state of emergency. A Blue Sky is a normal day. Gray Skies imply some type of a disaster is occurring, which then consists of a range going all the way to Black Skies.

Government open data is critical for Blue Sky preparedness and risk reduction efforts to inform investment in critical infrastructure and industries to make existing fragile supply chains more resilient to global shocks such as pandemics. For Gray and Black Skies, employing the open data in readily available and machine-readable formats, such as comma-separated values, will allow the data to be ingested into databases for research, analysis, and decision support. This will help emergency management and public health officials to make data-driven decisions at the speed of thought, rather than the speed of bureaucracy.

Government has the opportunity to educate the public and emergency management professionals about the availability of data and corresponding data science tools for maximum utility in crises. Open source software communities have existed for decades to help government agencies learn best practices on democratizing data and software to train, reskill, and equip leadership and staff. U.S. National Laboratories and the intelligence community have led the world in high-performance computing and employment of graphics processing units (GPUs) and accelerated data science for their missions. Public health and emergency management agencies have yet to employ GPUs, data science platforms, and open source software to their full extent to prepare and respond to crises.

An emergent ecosystem of companies and partners is building and deploying technologies to advance emergency management and public health preparedness and response. Disaster Tech is a public benefit company leveraging GPU-accelerated data science tools that are purpose-built for the emergency management and public health agencies to prepare and respond to disasters. Disaster Tech has partnered with NVIDIA, Microsoft, Kinetica, Indiana University, University of Delaware, and local and state emergency managers to enable the unity of effort between industry, academia, and practitioners that is required for responding to a pandemic such as COVID-19 or future disasters.

Jim Scott is head of developer relations, Data Science, at NVIDIA (www.nvidia.com). Over his career, he has held positions running operations, engineering, architecture, and QA teams in the big data, regulatory, digital advertising, retail analytics, IoT, financial services, manufacturing, healthcare, chemicals, and geographical information systems industries.

Pandemics Happen—AI and Machine Learning Can Provide the Cures

Near the end of the Middle Ages, at the beginning of the early Renaissance, an event occurred as significant as any geological boundary that signifies a great extinction event. The years between 1347 and 1353, a period known as the "Black Death" killed approximately half the population of Europe and could roughly be compared to the asteroid impact 66 million years ago known as the Cretaceous-Tertiary (K-T) boundary. While the rogue asteroid that impacted near the Yucatan Peninsula hurried the dominance of mammals while eliminating dinosaurs, the Black Death nearly wiped out human civilization. The plague came in three forms, bubonic, pneumonic, and septicemic, but it was caused by a simple bacterium, known to the modern world as *Yersinia pestis*.

Interestingly, this bacterium still exists and infects a dozen or so people yearly, even in the U.S. The infections occur in such places as southwest Utah, where hiking trails have signs warning that the bubonic plague has been found in the local area and advice to avoid dead animals that you may encounter on your route. A substantial but well-known and well-understood course of antibiotics usually proves sufficient to the task of arresting the disease when someone is unlucky enough to contract it. Yet, less than 700 years ago, it nearly caused Armageddon.

What We've Learned

The difference-maker is the scientific knowledge that humanity has acquired between then and now. Around 1352, the University of Paris (often known as La Sorbonne), which employed some of the best minds of the late Middle Ages, concluded that a primary cause of the "Great Pestilence" was infected air due mostly to astrological effects.

In the same way that the Yersinia pestis bacterium ushered in the age of science, which accompanied the Renaissance, so should SARS-CoV2, and its respective disease COVID-19, propel the world toward a change in how we approach disease in our time.

Now we know more. But how much more? Who would have conceived in December of 2019 that it would be possible for the world economy to be shut down by a virus? Or that the most prominent social norm would become "social distancing?" That the main social requirement throughout the world would be to accept the directive of social distancing somewhat blindly is frightening enough, but there are deeper systemic problems that have paved this perilous path. Will future centuries view us in the same way as we view the leaders at the University of Paris in 1352?

However, there is an answer. In the same way that the *Yersinia pestis* bacterium ushered in the Age of Science, which accompanied the Renaissance, so should SARS-CoV2, and its respective disease COVID-19, propel the world toward a change in how we approach disease in our time.

AI and Machine Learning

Twenty-first century computing can be the answer. AI and machine learning can be used for much more than recognizing a cat in the midst of a pack of dogs or helping sports gamblers gain an advantage in the next week's games. AI and machine learning can be utilized to accelerate the recognition of the new germs—be they viruses or bacterium that exist among us—that we have yet to encounter in a significant manner. Then, that same technology can be used to accelerate the development of countermeasures such as vaccines, anti-viral drugs, antibiotics, and treatment protocols.

Michael Corey, co-founder of LicenseFortress, was recognized in 2017 as one of the top 100 people who influence the cloud. He is a former 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.

It takes a minimum of 18 months to develop vaccines and anti-viral drugs. The bureaucracy and the human trials, when combined with the discovery process which involves endless hours of iterative testing, should be the perfect target for the use of the modeling capability of machine learning and the inference capability coming from new AI algorithms.

It takes a minimum of 18 months to develop vaccines and anti-viral drugs. The bureaucracy and the human trials, when combined with the discovery process, which involves endless hours of iterative testing, should be the perfect target for the use of the modeling capability of machine learning and the inference capability coming from new AI algorithms. Since we are presently learning that 18 days is more than enough time to unravel a great percentage of the world's economy, 18 months constitutes a true catastrophe regardless of the actual real effects of the disease itself.

What Is Needed Now?

We call for a modern Manhattan Project- or moon landing-scale effort to bring together all the great minds and companies of the Six Cities of Silicon Valley (www.dbta.com/Editorial/News-Flashes/The-Six-Sister-Cities-of-Silicon-Valley-125972.aspx). Their mission should be to invent new uses of AI and machine learning that will use the incredible power of modern computing with graphical processing units (GPUs) to solve this problem. That mission should be to discover the prevention of disease itself!

No one knows how bad the year will get. No one knows if historians of future centuries will deride 21st century society as we now view the figures of the mid-14th century. Those people were undoubtedly using every academic idea and scientific tool at their disposal to try and lessen the effect of the catastrophe they were immersed in. We have the ability and the tools to make sure that this type of event never again brings the world economic system to the brink of collapse. Let's create our generation's version of the "moon landing" and get it done!

Stemming Your Data Contagion

As I work from home pondering data practices amid the novel coronavirus pandemic, it is not surprising that an unfortunate analogy comes to mind. Data is an intrinsic part of business today. New sources are constantly being created and new ideas to explore are being conceived.

It is a matter of when, not if, your organization will confront a never-before-seen data source—a source that, if managed improperly, could result in catastrophic consequences to your brand and bottom line. In some cases, that data will be imported from outside your four walls. In others, the data will spring from new business processes or the fertile minds of your employees manipulating existing assets to create altogether new analytic insights.

Regardless of the source, far too many governance programs focus on reflexively and reactively imposing rigid controls on what is already known. As a result, while organizations relentlessly herald the need for everyone to become data-driven, existing governance practices encourage those venturing into the fray to do so covertly to avoid potential judgment or punitive action.

You cannot stop the flow of data into and within your enterprise. You can, however, direct that flow to ensure appropriate containment. To do so, your organization must, at a minimum, invest in the following:

- Data workspaces supporting different levels of containment, from sandboxes to warehouses
- Methods to easily check new content into a managed catalog
- "How-to" guidance and tools to categorize new content with minimal bureaucratic overhead
- The ability to ramp resources up and down quickly based on evolving demand

Note that the goal here is not to encourage people to only come forward with pristine, fully vetted, high-quality information sources, or to do so after they've gotten themselves in a quagmire that they can't get out of alone. The goal is to encourage people to "*fess up*" *to everything* and *provide an honest assessment* of what is known—and not known—about the data *as soon as possible*. In this way, sources that may be incomplete, whose value and quality are unclear, or that might present a heightened risk, can be quickly and effectively quarantined. In most cases, information can likely be made widely available in short order with few, if any, restrictions. Others may require a negative-pressure data containment area accessible only to specially certified resources until the data has been comprehensively assessed. But again, the key is not to make people afraid to come forward but to make it a no-brainer to do so.

Once a data asset goes public (so to speak), ensuring it continues to be used and shared appropriately is the name of the game. Mitigation is often seen as a reactive, catch-up strategy. And too often it is applied as such. However, organizations that are proactive in educating and engaging their workforce can increase their overall resiliency and decrease risk. Social distancing only works when a critical mass of the population engages—so too with data governance. To that end, organizations must provide the following:

- Easy-to-understand usage policies regarding how, when, and for what purposes data can be used
- Highly visible content labels and warnings at the point of consumption (if they must dig for information, it won't happen)
- Low-friction mechanisms to gain access to content in accordance with defined policies
- Efficient workflows and resources for answering questions and initiating requests
- Readily available training and guidance on data tools and environments

Nonetheless, the best policies and procedures are useless if not clearly communicated and understood. Your employees, by and large, want to do the right thing. Ensure that your data policies are simple and clear. Make sure guidelines and policies are consistent—and that the same message is given in print, in video, or in person. Resist the urge to try cover every eventuality or obfuscate hard realities; you can't do it, and a 100-point rule list complete with too many "if-but-except" clauses will discourage people from attempting to engage at all. Rather, create an environment in which people are motivated to engage, are informed and enabled to make the right decisions, and secure enough to ask when they aren't sure. Is your governance program up to the challenge?

Kimberly Nevala is a strategic advisor at SAS (www.sas.com). She provides counsel on the strategic value and real-world realities of emerging advanced analytics and information trends to companies worldwide and is currently focused on demystifying the business potential and practical implications of AI and machine learning.

IoT and Data Power the Next Generation of Clean Energy

WIND POWER IS ONE OF THE FASTEST-GROWING ENERGY sources, thanks to its near-zero carbon emissions. Not only is it cleaner than fossil fuel alternatives, it is also cheaper and more sustainable.

Over the last decade wind turbines literally have sprung up everywhere, from coast to countryside. In Europe alone, installed capacity has increased from around 13 gigawatts in 2000 to more than 180 gigawatts in 2018.

There are challenges to wind power, of course; modern wind turbines are huge and can impact a view. Wind farms are often remote, which means that maintenance involves frequent and difficult trips by service staff—who may have to climb the structures to undertake repairs.

Repairs and maintenance, therefore, require experienced staff to keep pace with booming installations and regulatory requirements. There has been a gap in the development of operations and maintenance skills in the EU, with a current shortage of 7,000 qualified personnel required by the European wind energy sector each year. This figure could increase to 15,000 by 2030.

Sensors Can Work Where People Can't

Now that wind power is no longer seen as an optional alternative but instead as a vital component in fulfilling the demand for energy, new methods of support are in high demand to keep up with the amazing growth through robust digital technology.

When you combine the skills shortage with the possibility of dangerous working conditions for maintenance personnel, you can see that there is a need for a new solution. Operators need to find a way to maintain the critical infrastructure at wind farms, while also reducing travel to their facilities. But how can they ensure the continuity and safety of operations from afar?

There are ways to do this using IoT, integration, and business transformation technology. Applied properly, they can help to ensure business continuity in the most trying of times.

The ability to monitor wind turbines remotely becomes more important when maintaining appropriate turbine availability levels. Condition monitoring overcomes the issue of accessibility while enabling wind farms to extend turbine maintenance intervals, manage resources more effectively, and avoid costly downtime.

A consistent, end-to-end architecture based on an IoT platform allows real-time analytics to manage autonomous operations, such as switching on rotor blade heaters when ice starts to form, and the automatic handling of alarms and exceptions.

This is a real-life project, covering an impressive 7,000 existing turbines, and 1,000 new turbines, which is being rolled out over the upcoming 5 years.

The benefits are in various areas. The first is that the commissioning process—time to value—is decreased by the standardization of monitoring and control. More importantly, there is a full and carefully implemented investment program in new component and maintenance technologies, giving the operators control over planned maintenance and helping to eliminate unplanned maintenance levels.

On top of that, in these difficult times, this program is paying off in an unexpected way: As some of the turbines are in restricted areas and travel for the maintenance engineers is not possible, being able to monitor and control the turbines remotely is a precious bonus.

Powering the next generation of renewable energy with IoT positively impacts running costs, while ensuring that wind power will continue to be both a viable economic solution for energy production for the future and a benefit to the environment.

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

GET THE ELEPHANT OUT OF THE ROOM

Bad address and contact data that prevents effective engagement with customers via postal mail, email and phone is the elephant in the room for many companies. Melissa's 30+ years of domain experience in address management, patented fuzzy matching and multisourced reference datasets power the global data quality tools you need to keep customer data clean, correct and current. Tell bad data to vamoose, skedaddle, and Get the El out for good!

Data Quality APIs

- Global Address Verification
- Global Email
- Identity Verification
- Global Phone
- Geocoding
- Demographics/ Firmographics
- U.S. Property
- Matching/ Deduping

Integrations

- talend
- Microsoft SQL Server
- pentaho
- salesforce
- Microsoft Dynamics

Activate a Demo Account and
Get a Free Pair of Elephant Socks!
i.Melissa.com/dbta

melissa

www.Melissa.com | 1-800-MELISSA