Volume 4 Number 3 ■ FALL 2018

BDQ

BIG DATA QUARTERLY

BIG DATA

50 Companies Driving Innovation

WWW.DBTA.COM

20

Using Data and Analytics in a Customer-Centric Way 11

Six Pitfalls to Avoid During IT Automation 18

Oracle DBAs Versus SQL Server DBAs 37

THE PUBLICATION FOR THE ERA OF

BIG DATA

Brought to you by the editors of *Database Trends and Applications* magazine, *Big Data Quarterly* is for information management and business professionals who are looking to leverage big data in organizations of all kinds. Subscribe today and stay informed on big data, data science, and the technologies and business strategies surrounding them.

This is a must-read publication for data scientists, CIOs, and other professionals involved with big data projects.

LIMITED-TIME FREE OFFER!*

SUBSCRIBE NOW.

dbta.com/BDQ/Subscribe

*Free to qualified U.S. subscribers.
Regular subscription rate is $97.95 per year.

BDQ
BIG DATA QUARTERLY

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

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

CORPORATE HEADQUARTERS 143 Old Marlton Pike, Medford, NJ 08055

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

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

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

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

Information Today, Inc.

CONTENTS

editor's note | *Joyce Wells*

2 **Taking Big Data and Analytics to the Next Level**

departments

3 BIG DATA BRIEFING

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

11 INSIGHTS | *David Judge*

Digital Transformation: Using Data and Analytics in a Customer-Centric Way

18 INSIGHTS | *André Christ*

Six Pitfalls to Avoid During IT Automation

32 TRENDING NOW

Enter the Chief Data Officer: Q&A With Accenture Applied Intelligence's Ramesh Nair

features

4 THE VOICE OF BIG DATA

Ensuring Optimal Database Performance in the New Cloud World: Q&A With Patrick O'Keeffe, Quest Software VP of Engineering

6 FEATURE ARTICLE | *Joyce Wells*

Tooling Up for Analytics

SPECIAL SECTION > BIG DATA 50

20 **Introduction**

21 **Big Data 50: Companies Driving Innovation**

30 BIG DATA BY THE NUMBERS

Data Governance: Bringing Order to Chaos

columns

36 DATAOPS PLAYBOOK | *Jim Scott*

Data Operations Problems Created by Deep Learning

37 CLOUD CURRENTS
Michael Corey & Don Sullivan

Oracle DBAs Versus SQL Server DBAs

39 GOVERNING GUIDELINES | *Anne Buff*

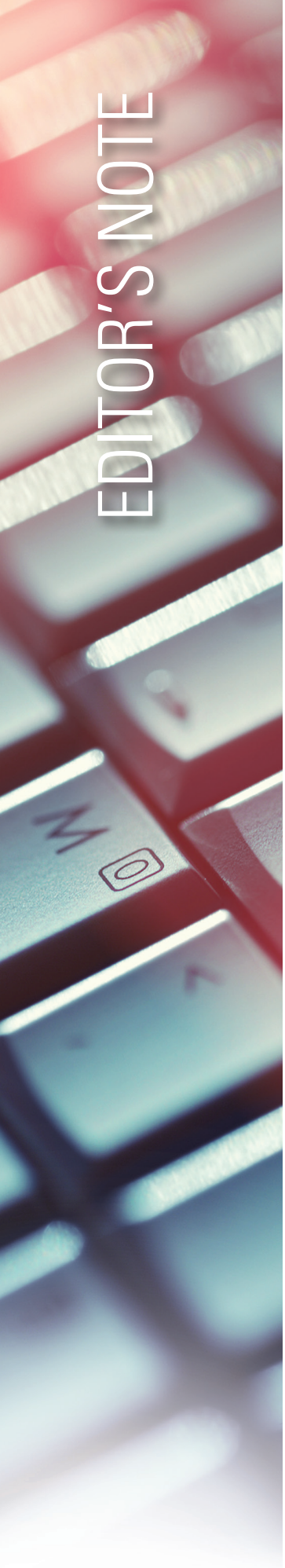
Data Governance: We Are Programmed to Receive

41 THE IoT INSIDER | *Bart Schouw*

Good Habits Light the Way to IoT Innovation

43 DATA SCIENCE DEEP DIVE | *Bart Baesens*

Data Warehouses, Data Marts, Operational Data Stores, and Data Lakes: What's in a Name?



EDITOR'S NOTE

Taking Big Data and Analytics to the Next Level

By Joyce Wells

While the concept of "unstructured" data flowing into organizations from web interactions, social media, smartphones, gadgets, and machinery with sensors has now become familiar, the technologies that help organizations derive actionable insights from these sources are still evolving. The shifts are often subtle.

Yes, artificial intelligence is coming to the fore, but experts note that, at least for the near-term, it will be used to enhance, not replace, human intelligence. Hadoop MapReduce is still a foundational technology, but many industry leaders note that Spark is increasingly being used as well for processing big data. Cloud is frequently the first choice for new tech deployments, but a hybrid or multi-cloud approach is often desirable to decrease risk. Meanwhile, data governance, long thought of as necessary but rather unexciting, is getting new attention in the era of heightened regulatory mandates and spurring conversations about the additional dividends it pays in terms of reliable data for a variety of constituents.

Underlying many of today's technology changes is the need for faster access to trustworthy information on which to base decision making. In this issue of *Big Data Quarterly*, these and other big data topics are explored from a variety of viewpoints.

In our feature article on "Tooling Up for Analytics," VMware's Avanti Sane points out that with the pressure for fast decision making, the time it takes to send up IoT data to the cloud creates too much latency, making edge computing critical to real-time analytics. In addition, notes Radiant Advisors' John O'Brien, graph databases are "finally getting their day in the sun," due to their ability to enhance data lakes by shedding light on relationships and storing context. And, while blockchain offers the promise to improve the security and transparency of business-to-business transactions, it is important to carefully evaluate tool sets for building new applications and interoperating with other blockchain networks, notes Oracle's Frank Xiong.

These changes are reflected in the roles of data professionals. According to Quest's Patrick O'Keefe, the increasing heterogeneity of systems and rise of the DevOps culture are having major impacts on DBAs. Accenture's Ramesh Nair also explains the need for a comprehensive approach that ties a variety of elements together is prompting many companies, particularly in financial services, to appoint chief data officers, or CDOs. And, there are many other articles in this issue that highlight the evolving trends in big data and analytics.

As an additional resource to help data-driven organizations take their big data and analytics strategies to the next level, this issue of *Big Data Quarterly* includes the fourth annual "Big Data 50—Companies Driving Innovation." Each of the companies on this list is helping to drive the big data ecosystem forward with innovative technologies, products, and services.

To continue to stay on top of the latest big data trends and research, visit www.dbta.com/bigdataquarterly. And don't forget to mark your calendar for the next Data Summit conference coming to the Hyatt Regency Boston from May 21 to 22, 2019, with preconference workshops on Monday, May 20.

BIG DATA BRIEFING

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

Oracle has announced the availability of the company's latest Autonomous Database Cloud Service, **Oracle Autonomous Transaction Processing**. Leveraging machine learning and automation capabilities, the company says the new transaction processing service delivers cost savings, security, availability, and productivity and can support a complex mix of high-performance transactions, reporting, batch, IoT, and machine learning in a single database. www.oracle.com

OpenText, a provider of enterprise information management, is releasing a next-generation hybrid-cloud platform that brings together intelligent automation, security, and EIM applications. **OpenText OT2** enables developers to rapidly build SaaS applications in the OpenText Cloud. www.opentext.com

The dbKoda team has released **dbKoda 1.0**, the first production release of its open source integrated development environment for MongoDB. dbKoda offers features for developing MongoDB applications and for managing and tuning MongoDB databases. www.dbkoda.com

Amazon Web Services, Inc., an Amazon.com company, unveiled **Amazon Aurora Serverless**, a new deployment option for Amazon Aurora that automatically starts, scales, and shuts down database capacity with per-second billing for applications with less predictable usage patterns. Amazon Aurora Serverless offers database capacity without the need to provision, scale, and manage any servers. https://aws.amazon.com

SAP is making improvements to SAP **S/4HANA Cloud**, delivering dozens of new AI-powered scenarios that further power the intelligent enterprise. Building on SAP's industry domain expertise to help every customer become a best-run business, these updates will increase agility and flexibility as customers adapt to rapidly changing business conditions. www.sap.com

Instana, a provider of application monitoring for containerized microservice applications, is releasing a set of customization features across its solution, creating a personalized APM experience for users. The new capabilities, called Application Perspectives, go beyond traditional custom dashboarding to deliver an automated APM experience. www.instana.com

Yellowbrick Data has emerged from stealth to announce the debut of its analytic solution for hybrid cloud, the cornerstone of which is the **Yellowbrick Data Warehouse**. According to Yellowbrick, the data warehousing segment has been in need of a refresh. The company contends that a fragmented ecosystem of solutions based on legacy architectures has hindered efficiency and forced businesses to compromise on how quickly and easily they can derive insights from data. https://yellowbrickdata.com

Databricks is partnering with **RStudio**, providers of a free and open-source integrated development environment for R, to increase the productivity of data science teams and allow both companies to integrate Databricks' Unified Analytics Platform with the RStudio Server. The RStudio and Databricks integration removes the barriers that stop most R-based machine learning and artificial intelligence projects. https://databricks.com and www.rstudio.com

Looker is integrating its platform with **Google Cloud BigQuery ML**, accelerating the time-to-value of data science workflows. With Looker and BigQuery ML, data teams can now save time and eliminate unnecessary processes by creating machine learning models directly in BigQuery via Looker—without the need to transfer data into additional machine learning tools. https://looker.com

MarkLogic, a provider of an enterprise NoSQL database platform, has introduced the **MarkLogic Query Service**, a new way to give customers elasticity in the cloud for their mission-critical, enterprise-grade workloads. www.marklogic.com

Pure Storage, provider of an all-flash storage platform, has introduced FlashStack with FlashBlade to accelerate data warehouses. The company also announced that AIRI and AIRI Mini, AI-ready infrastructure solutions from Pure Storage and NVIDIA, are now available with Cisco Nexus ethernet switches. www.purestorage.com

Microsoft is releasing **Azure IoT Edge** and introducing robust capabilities to enable enterprise customers to bring their edge solutions to production. The new updates are open and flexible to provide users with greater choice. https://azure.microsoft.com

Ensuring Optimal Database Performance in the New Cloud World

Patrick O' Keeffe, Quest Software
VP of Engineering

Today, data management environments are highly complex and often span multiple vendors with deployments across on-premise data centers, clouds, and hybrid installations. In addition to the heterogeneity of systems, the processes surrounding database development and management have also changed. DevOps, a methodology for data scientists, DBAs, and others to participate in an agile workflow, puts a premium on speed and also means that DBAs do not wield the firm control they did in the past.

Amidst those shifts, there is also greater need for stellar database performance due to the greater emphasis on real-time responses. As database environments are evolving, so is the job of DBAs. The movement to the cloud—along with more automation of tasks typically associated with the DBA role—has led to some speculation that organizations may need fewer DBAs, but according to a recent Unisphere research study, so far, that does not appear to be true. According to more than 60% of the respondents, the number of people with the title "DBA" is holding steady, while around 20% say the number of people within that role is actually increasing. What is happening, however, with the greater diversity of platforms and methodologies is that data professionals' lives are becoming more difficult.

Recently, Patrick O' Keeffe, Quest Software VP of engineering, discussed how DBAs' jobs are changing and what is needed to ensure optimal database performance.

—*J.W.*

How has cloud changed the role of database administrators (DBAs)?

The cloud has changed the role of the DBA in a number of ways. It has added a level of complexity. DBAs used to be responsible for only a small number of servers on hardware. Then along came virtualized environments that took away hardware concerns but added others. At the same time, the number of databases a single DBA needed to manage was growing. Cloud has just added to the responsibility mix—it is yet another environment in which DBAs have to manage risk as data stewards.

The cloud has also brought along new technologies for the DBA to master. Relational databases have been with us for some time, but the cloud has added database-as-a-service approaches for traditional RDBMSs as well as NoSQL offerings both in the transactional and analytics spaces.

Lastly, the cloud has also ushered in new approaches to software development. Businesses want higher velocity at lower risk—this is the fundamental driver of DevOps cultures built on top of continued integration/continuous delivery [CI/CD] and Agile. The cloud both enables and tacitly mandates this culture—and configuration-as-code is de rigueur here. DBAs as data stewards have been slower to climb on board this train.

What are the issues with cloud that become more complex or easier?

The cloud takes away some issues and adds others. Take, for example, an organization adopting a platform-as-a-service offering. Things like OS patching and backups, to some degree, cease to be concerns, while data governance, security, and privacy start to become even more critical.

How is DevOps changing the DBA role?

DevOps is about putting creators and risk managers on the same team, and this itself is new for DBAs from an organizational perspective. Businesses undergoing digital transformation are doing more software development and at the same time demanding higher velocity from these teams. This is demanding that DBAs work more closely with development teams.

Is the fail fast mantra of the methodology a double-edged sword?

All innovation is change, which means that if you want to be innovative, you need to be prepared to change. You also need to be prepared to fail. Failure, always being an option, can be difficult for some organizations to confront culturally because often there are organizational or process constraints built specifically to try and avoid failure. So, for innovation to take place, the environment needs to be psychologically safe, and for that to happen these barriers need to be removed. This means that if you're going to try to build a culture where failure is normal—and as it happens, necessary—you need to make failure as cheap as possible. Hence, the desire to fail fast. I'd say it's absolutely positive.

What kinds of challenges does DevOps present for DBAs? Can you identify some of the dangers?

Tooling is an issue as databases are many, varied, and complex. The tooling to support CI/CD for traditional databases is still not as mature as we'd like. This is potentially driving developers to use "simpler" databases for easier enablement of these workflows. This dynamic of development teams making technology adoption decisions is ascendant and represents business risk. The challenge is that development teams often do not consult experienced DBAs on these data storage adoption choices, and this is a missed opportunity because the initial choice can sometimes end up as technical debt that needs to be paid down at a future time.

What are some approaches that can be put in place to alleviate the risk associated with cloud and DevOps?

It is important to encourage DBAs as data stewards to become involved in DevOps initiatives. They have much to contribute due to their experience. Having them involved early will mitigate risks around technology choice, and will ensure that requirements arising from needing to manage risk are at least on the table.

'DevOps is about putting creators and risk managers on the same team, and this itself is new for DBAs from an organizational perspective.'

What are the capabilities that DBAs need to succeed in their roles today?

DBAs need to have an open mind, eagerness to learn, and the ability to adapt. Today's DBA should be ready to raise their hand and take a seat at the DevOps table. They should also be well-versed in all different types of databases—relational, non-relational, and cloud-based. As providers each have completely different "SQL dialects" for programming within databases, DBAs also need to be fluent in each of them. Critical to the success of DBAs is staying on top of the latest tools available to make their jobs easier and create a positive impact in today's data-driven business environment.

TOOLING UP for ANALYTICS

By Joyce Wells

NEW TOOLS AND technologies are becoming established in enterprises, helping organizations extract more value from data. To explore the ways that data-centric companies are managing their data more effectively, *BDQ* asked industry experts about the trends that loom large in the future and those that appear to be waning.

Data Lakes and the Rise of Spark

Data lakes, which can handle large, diverse sets of data in its original state, continue to expand, and organizations are increasingly making use of Spark for processing and analyzing the data. While Hadoop MapReduce is still leveraged for big data processing, many industry experts say that its use is leveling off.

"Most Global 2000 companies have data lake initiatives underway, and many of these deployments have moved beyond exclusive use of MapReduce to use a mix of MapReduce and Apache Spark as their primary means of processing and analyzing data," said Kelly Stirman, CMO and VP, strategy, at Dremio. "When these companies have moved to cloud deployments on AWS and Azure, many are using object stores like Amazon S3 and Azure Data Lake Store instead of HDFS [Hadoop Distributed File System] to store their data, while processing of the data is handled by Spark and some cloud-native processing services. Because the data lake does not provide sufficient performance and concurrency for BI workloads, many companies move data from their data lake into a data warehouse such as Teradata on-prem, and cloud services such as Redshift and Azure SQL Data Warehouse." In addition, Stirman noted, data-as-a-service platforms are being used with the data lake and the data warehouse to provide a uniform access layer that is capable of joining and accelerating data between the two.

Spark is being adopted because its power and capabilities are "significantly more advanced" than most analytics tool sets available within the Hadoop ecosystem, agreed Brian Schwarz, VP of product management at Pure Storage, who also pointed out that Spark's core strengths are its flexibility and the power to simplify company's environments. "Spark is generally perceived as the most dexterous and agile technology for big data processing given that it can handle large amounts and different types of analytics." Instead of deploying three or four smaller, more specialized tools, the libraries within Spark give companies wide flexibility to handle tasks such as SQL queries or other types of unstructured analysis, pattern matching, machine learning (ML) library searches, and regression analysis, said Schwarz.

Hadoop is still an important foundational technology that many customers are using within their data lake with Apache Spark, stated Mike Lehmann, VP of product management, Oracle. However, he added, "Oracle is seeing a transition happening where there is less of a dependence on the core Hadoop ecosystem and more focus on putting data directly into cloud object storage, doing data processing and manipulation using Apache Spark, real-time interactive queries directly against the data lake, and the overall infrastructure running on cloud-native foundations such as Kubernetes."

A cacophony of streaming data from devices and sensors is contributing to the demand for real-time analytics, which necessitates edge processing.

Analytics Gets Closer to the Edge

A cacophony of streaming data from devices and sensors is contributing to the demand for real-time analytics, which necessitates edge processing, executives noted.

The need for real-time analytics is motived by IoT, said Avanti Sané, product marketing, Internet of Things, VMware. "There are now millions of devices across the globe transmitting billions of bits of data. The majority of data being generated by these devices is time-sensitive, but by the time a piece of data is sent up to the cloud and brought back down, it has already become obsolete," Sané said. As a result, edge computing, where processing occurs at the extremes of a network near where the data is generated, "greatly accelerates" the steps, making it critical to real-time analytics. "In addition, the advent of 5G will bring faster networks that are capable of accommodating mass volumes of data and will thus bring IoT adoption further into the mainstream," she noted.

Real-time analytics is increasingly being utilized in many industries, said Oracle's Lehmann. Examples range from manufacturing and factory environments to online commerce clickstream analysis to real-time fraud detection in financial services, among many others. "To start, organizations need to develop an expertise in streaming data in real time to bring the data for analysis first," said Lehmann. "Next, the data needs to be analyzed, in motion, in real-time. And the requirements for that analysis are growing. Organizations need to be able to deploy trained and tested machine learning models against real-time data streams to derive more sophisticated insight in real time—it's not enough to do simple rule-based analytics on real-time data anymore." And, finally, he said, real-time analysis needs access to the real-time data as well as historical or contextual data in databases, in data lakes, and in cloud-based object storage.

AI and ML Enhance Human Decisions

The combination of artificial intelligence (AI) and ML is seen as a way to incorporate information about opportunities and risk into the decision making process with a level of speed that surpasses human capabilities. In areas as diverse as banking and finance, marketing, and healthcare, AI and ML can inject intelligence to improve the efficiency of processes.

"There's a vast movement toward machine learning, and ultimately autonomy, happening in many industries, including manufacturing, automotive, healthcare, information technology, and others," said Tim Hall, InfluxData VP of products. But, he noted, it is important to remember that this is a journey that originally began with business intelligence [BI], data warehousing, and reporting—"essentially, the desire to understand what happened based on information that had been gathered."

Now, the physical world is being instrumented via sensors embedded in an increasing number of consumer products and a very rapidly growing variety of industrial products, Hall continued.

In manufacturing, he said, "Industry 4.0" represents a new industrial revolution that utilizes robotics, automation, and data exchange of cyber-physical systems to create a smart factory that includes machines that learn with the objective of being able to make decisions for themselves based on the information, scenarios, and goals they are directed to achieve.

"In today's highly digital and fast-moving world, there's simply too much data, in too much complexity, for people to extract what they need through all of the noise," said John O'Brien, principal analyst at the research and advisory firm Radiant Advisors. "The promising use of AI/ML exists for companies to tackle their existing operational challenges where the problems are too complex to solve with traditional rules-based analytics." However, he noted, "One obstacle for training AI/ML models has been a lack of quality training data." The data has often been too scarce and difficult to generate objective real-world training datasets; the existing data has had too much historical bias; or the data has been the product of bad processes. A trend for AI and ML for the foreseeable future, said O'Brien, will be in "human assistance"—using AI and ML where possible to assist humans to be faster and more accurate.

AI and ML are becoming prevalent in the enterprise, but the area of most opportunity is the integration of those technologies with BI-style analytics, agreed Priyank Patel, co-founder and chief product officer, Arcadia Data. For example, he said, "You can use BI and visual analytics to present AI/ML outputs in a human-readable form. Visualizations like heat maps and flow charts can help non-technical users quickly navigate to insights that were uncovered by the complex algorithms built by your data scientists. And, since BI tools are designed for non-engineers to build visualizations, you create a self-service model for analyzing the outputs of your advanced analytics."

AI and ML processes can also be integrated into BI tools to help business analysts be more productive with their analytical work. You can think of AI for analytics as being similar to GPS navigation for business users "who have to deal with tons of information and big data, and need some help navigating to find useful insights," said Patel. This leaves the power of decision making in the hands of humans, but it accelerates the process.

Microservices and Containers Support Hybrid Clouds

Cloud implementations are seen as an important element for providing easy access to advanced technology to companies of all sizes. With the adoption of cloud there is also concern about vendor lock-in, however. This makes hybrid and multi-cloud scenarios that incorporate a mix of deployments more appealing, despite the risk of greater complexity.

"The primary purpose of cloud is to get on-demand IT resources and elastic scalability, which enables companies to avoid building their own and focus on their core competencies while allowing them to easily meet demand without overprovisioning," said Eric Holzhauer, principal manager, strategy and product marketing, MongoDB. "Cloud also levels the playing field, so to speak. Whether you're an individual developer, a small startup, or a Fortune 500 company, you have access to first-class, elastic, globally available, and high-performance resources." The majority of MongoDB users are deploying in the cloud, either by installing it on cloud infrastructure themselves or using MongoDB's managed cloud database, he added.

"We're seeing a higher need—now more than ever—for hybrid cloud strategies among customers wary of lock-in with a single cloud vendor," said Scott Clinton, VP of product marketing at Hortonworks. And, he said, for developers building the next generation of cloud-native data applications, the use of containers and microservices lets them move fast, deploy more software efficiently, and operate with increased velocity in the DevOps environments in which more data applications are being built.

These technologies, and Kubernetes and Docker, in particular, allow applications to be packaged in ways that make them very transportable across clouds and easily able to scale elastically on a cloud platform, noted Guy Harrison, CTO of Southbank Software. "Microservices are the modern equivalent of a modular code architecture since they allow loose coupling between cloud-based application components."

Containers provide a strong model of isolation, separating the software operating environment from the environment it is physically deployed within, added Jim Scott, director, enterprise strategy and architecture, MapR. "This creates a substantial value proposition for those who want to run software in more than one location." In the cloud, where companies don't know what hardware is under the covers, containers simplify the deployment and movement of software applications from on premise to the cloud. "This is a critical requirement when running cloud, multi-cloud, hybrid cloud/on-premise, or even multiple on-premise environments." Containers, Scott added, "are great to use with microservices" because they are lightweight and can be physically isolated from one another.

Data Governance in the GDPR Era

Data governance and data security initiatives are getting heightened attention with the recent implementation of the General Data Protection Regulation (GDPR) in the European Union; the California Consumer Privacy Act of 2018, which takes effect in 2020; and an assortment of other regulations surrounding handling of data.

While most organizations still view governance as a cost of doing business that slows them down, some have realized that governance requirements are forcing them to gather data and create views that no individual part of ▶

Since GDPR was put into effect in May 2018, there has been an increased awareness of data governance and the risks to data security.

the business would do on their own, noted Joe Pasqua, EVP for MarkLogic. "Doing it right means collecting the data as-is—not transforming away value or dropping things on the floor—harmonizing it incrementally, and building business-level microservices to access it. This can result in a transformative tool for the business and deals with governance requirements along the way."

Since GDPR was put into effect in May 2018, and even in the months leading up to the compliance deadline, there has been an increased awareness of data governance and the risks to data security, observed Hortonworks' Clinton. "Putting data into the hands of the customers who request it allows for a never-before-seen sense of control and efficiency in which businesses are forced to improve their security protocol—reducing data loss and improving peace of mind. In addition, data governance enables IT organizations to provide trusted data to business consumers, including analysts, data scientists, and others."

The Rise of Graph DBs and Blockchain

Looking to the future, several data management technologies are poised to make an impact. "Graph databases are finally getting their day in the sun, in large part due to the challenges of data lake management," said Radiant Advisors' O'Brien. "These graph database implementations are well-suited to highlight relationships and store context for objects captured in data that are related to each other in a variety of ways." Because graph database engines are easy to use, highly scalable, and fast-performing, they enhance and augment data lake management, semantic layers, and governance, while automated metadata acquisition and self-service metadata collection allow for anything to be connected to anything easily, he said.

In addition, blockchain, the distributed ledger technology, offers "vast opportunities" in enterprise environments, with the promise to fundamentally transform how business is being done by making business-to-business interactions more secure, transparent, and efficient, said Frank Xiong, group VP of product development at Oracle. "It allows enterprises to extend their boundary to reach out to their suppliers, business partners, distributors, and end customers and to carry out operations transactions in automated way." Internally, a large enterprise can use blockchain technology to further integrate and automate processes to streamline operations and improve productivity, he explained.

However, although blockchain may ultimately transform society, noted Southbank's Harrison, "In the short term, there are not that many applications crying out for blockchain enablement. Private blockchains are commonly being deployed in B2B scenarios such as supply chains where multiple businesses need to coordinate. Public blockchains—Ethereum, for instance—are currently almost unused in an enterprise scenario."

The industry has yet to see blockchain-enabled frameworks to emerge that will provide "killer use cases," said Harrison. "For instance, I think that eventually companies that want to assert any sort of compliance, proof of payments—or anything else that might have legal implications—will do so on the blockchain. But for now, there isn't an easy way to do that—we need application frameworks to emerge that bake that capability in."

As enterprises evaluate potential blockchain solutions, it's important that they consider the tool sets offered that can assist in building new blockchain applications and interoperate with other blockchain networks and existing applications, said Oracle's Xiong.

Key Open Source Projects

Three Apache projects are seen as being on a strong upward trajectory for a variety of reasons.

Apache Arrow has expanded in popularity for in-memory columnar data processing, said Stirman, whose Dremio platform is based on Arrow. In addition, he said, the Gandiva Initiative, a new execution kernel for Arrow that will "dramatically accelerate" processing of Arrow data, was announced in June. In Apache Arrow Flight, an RPC [remote procedure call] for Arrow, was also recently announced to provide a modern alternative to ODBC/JDBC and allow systems to exchange data more efficiently.

"Kafka seems to be the default open source choice for developing integration pipelines," noted Harrison, while MarkLogic's Pasqua pointed out that—with the proliferation of data silos in the enterprise and the ascent of the data hub model—Apache NiFi has emerged as a way to easily route data from the silos into the hub.

When data is coming from many different systems into a central data hub, organizations need to know where it came from, when, and using what processes, and NiFi helps to capture that flow of information, said Pasqua. In the new more stringent regulatory environment, "the Wild West days" of data lakes just don't cut it, Pasqua concluded.

INSIGHTS

Digital Transformation: Using Data and Analytics in a Customer-Centric Way

By David Judge

PREVIOUS TECHNOLOGY SHIFTS, such as the introduction of enterprise resource planning software, tended to be inward-facing, involving little contact with the customer. Things are different this time around. The companies that are embracing today's digital transformation trend are focusing squarely on the customer—connecting customer-facing initiatives with business processes to stay ahead of the competition.

How pivotal a role is the customer playing? In a recent study by the SAP Center for Business Insight and Oxford Economics, the top 100 "transformative" companies were 2.5 to 4 times more likely to report value from next-generation technologies. More than 90% of these digital transformation leaders have adopted mature digital transformation strategies and processes to improve the customer experience. Nearly three-quarters said digital transformation has generated significant or transformational value customer experience and engagement.

Clearly, the customer experience matters. Here are some ways today's most transformational companies are infusing data and analytics into their business processes to drive customer value.

Today's transformational companies can tap insights from customer interactions—from individuals or from the collective audience—to reach out over different channels with relevant communications.

They are generating information—and consolidating it.

Shoppers are demanding more choice and richer content at all times, on any device. To meet demand, retailers need to deliver more personalized content, faster and across channels. To do this, they need access to data—from every customer interaction, across multiple channels.

The highest achievers are consolidating customer interactions from stores, online, and mobile channels with insight from social media feedback via Twitter, Pinterest, and other platforms. The information serves as the basis for a "single source of truth" to help retailers optimize branding, marketing, promotions, pricing, merchandising, and inventory management processes.

They are using data to reach—*and interact with*—customers.

Today's transformational companies use data to interact with customers, not just market to them better. These businesses can tap insights from customer interactions—from individuals or from the collective audience—to reach out over different channels with relevant communications. Use of "chatbots" is becoming a popular mode of interaction, automatically improving user experiences based on data.

Data is critical—make no mistake. But it can't be positioned as a replacement for creativity. In other words, it's not about creativity *versus* data—it's about creativity *enhanced by* data.

Companies gather data to analyze and improve the customer experience and then to create targeted messages emphasizing the brand promise. But emotion, storytelling, and human connections remain as essential as ever.

They are engaging with the customer of one.

The interactions described above are part of a strategic shift where marketers devote less resources broadcasting at customer *groups* and more at interacting with customers individually. This is sometimes referred to as serving the "customer of one."

The shift has been driven by improvements in technology. In the past, marketers gathered information about broad demographic groups—men 25 to 49 years of age, women of a particular ethnic group, children living on the East Coast—and pitched broadly to those groups. They made assumptions about their targets' preferences based on a few demographic points and essentially told them what they should like. ▶

David Judge is VP of Leonardo| Analytics at SAP (www.sap.com)

Now, using analytics, machine learning (ML), and contextual information about a customer's current state, companies can enter into a dialogue with that particular customer. Everything the customer has done on the site and in their public information is similar to a fingerprint, revealing preferences. Segment data is still useful, but it's less of a focus these days.

They are developing products and experiences that self-improve.

The goal of every business is to continuously improve—to learn from its mistakes, make better products, and deliver better service. In the past, businesses would accomplish this by doing customer surveys and studying their own internal processes. These latter tactics are still important today, but advanced technologies are making the improvement cycle more efficient.

Tesla, for one, has equipped its cars with sensors and Internet of Things connections to gather vast amounts of data. The company collects this data and applies analytics and ML to create a better driving experience. It feeds improvements based on this data back into the cars, creating a better driving experience, which increases the attractiveness of the cars to consumers. Presumably, more consumers then purchase Teslas, which continues the virtuous cycle. Additionally, Tesla might use the collected data for other future business opportunities that are not clear yet.

Expect more of this in the not-too-distant future. More interactions will feed more data back to businesses, which will improve products and sales processes, making customer interactions more seamless and valuable.

They are capitalizing on micro-moments.

Today's customers have dispensed with the old practice of following a linear path to a sale—viewing an ad, clicking through to a website, and then buying what they need. They're embarking on fragmented journeys that can shift at any time. Retailers that embrace digital transformation are inserting themselves into those journeys to deliver the right information at the right time. They're hitting consumers with just what they need at what Google calls the "micro-moments" that matter.

Consumers that are in the "I Want to Know" moment are looking for information, not a hard sell. Brands can leverage predictive models to use this moment as a time to deliver instructive content that answers questions rather than pushes a sale. Consumers that have moved to an "I Want to Go" moment will be responsive to maps, directions, and in-stock inventories more than specific product information. To target consumers in an "I Want to Do" moment, create how-to videos to help people use your product or service, whether that means baking cookies or buying a home. Make sure the content is useful when people come to you on mobile mid-task.

Then there's the holy grail—the "I Want to Buy" moment. Consumers decide to buy from anywhere—the home, the car, the store aisle, or while out on a walk. Use location and device clues to help them seal the deal however they prefer: on your site or app, in a store, or on the phone.

They are treating customer data with care.

It's important to remember that the companies truly leading in customer experience are those that treat customer privacy with respect and are mindful not to cross boundaries. While opportunities for companies to deliver enhanced customer experiences are greater than ever with the advancement of technology, governance is also at an all-time high. With compliance standards such as the European Union's General Data Protection Regulation being enforced, companies should make data privacy a top priority and ensure all customer information is handled properly. Data transparency not only allows companies to adhere to regulations, but also helps strengthen customer relationships. This means that at any time, customers can ask a company to show what personal information they have in their databases and companies can easily comply with the request, giving customers peace of mind.

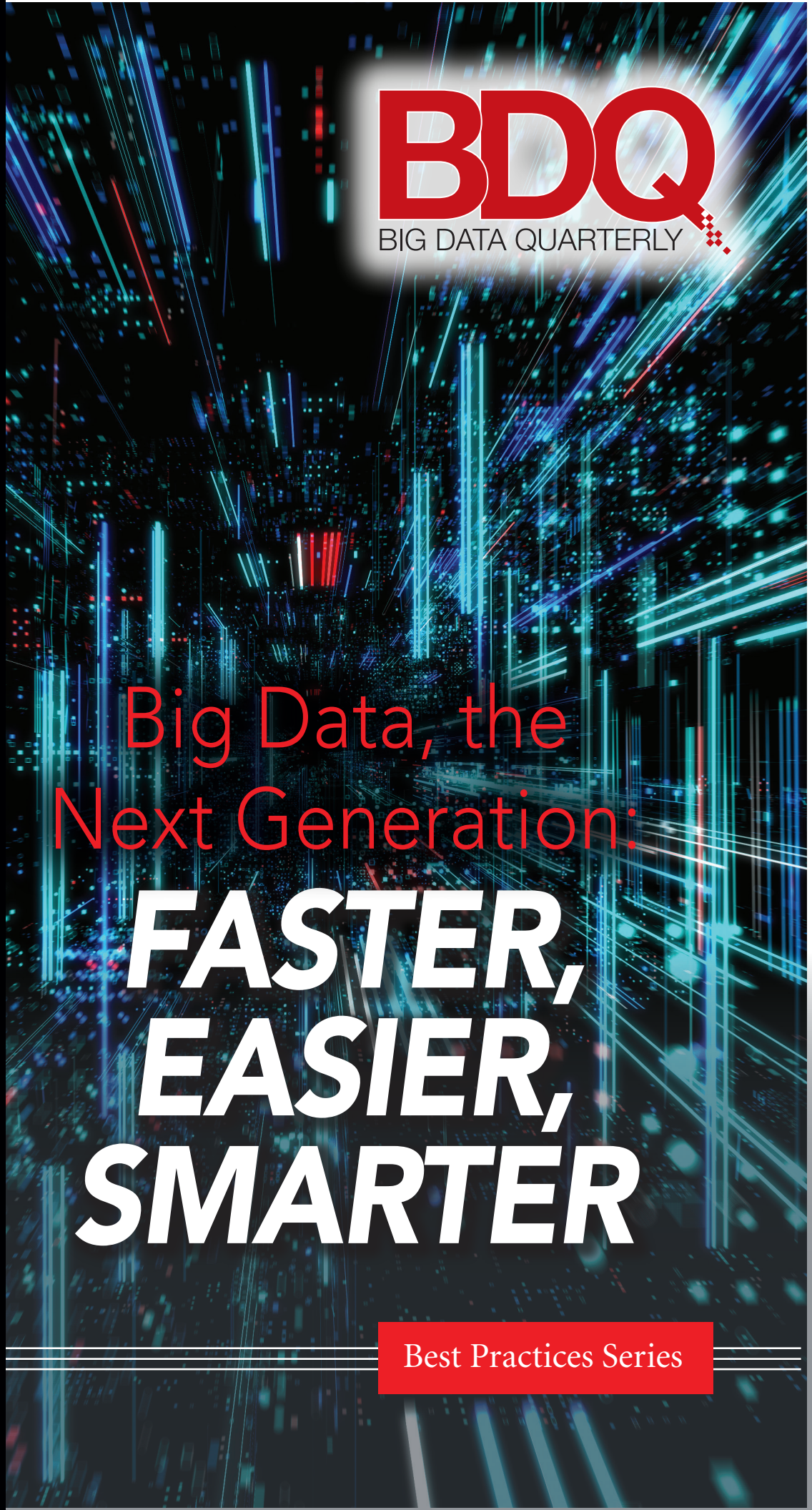
Be positioned for the future.

Today's most transformational companies are getting ahead by using data and analytics in an effective, customer-centric way. Whether it's ensuring each and every message a customer receives is specifically tailored to them, or diligently protecting their personal information, companies need to prioritize the customer in everything they do.

By taking advantage of customer insights, organizations can better serve individuals and interact with them on a personalized level. It's a practice that will position them well to take advantage of not only the current technology shift but to brace for the next one that comes along.

Aerospike
PAGE 16
MAXIMIZE THE VALUE OF YOUR OPERATIONAL DATA

MariaDB
PAGE 17
THE END OF SINGLE-PURPOSE BIG DATA PLATFORMS



Best Practices Series

"Big data" in the form we know it, pertaining to volume, variety, and velocity, has been top of mind at enterprises for close to a decade now. Capturing, deploying, and extracting value from it typically required a cadre of data specialists, scientists, analysts, and professionals.

These people will all be essential for moving forward into the next phase. However, managing and leveraging data is getting a whole lot easier. And this is not coming a moment too soon: Organizations are being inundated with data from the Internet of Things on the outside and artificial intelligence (AI) and machine learning on the inside.

Until recently, data analytics has been delivered through data warehouse and Hadoop-based stores. Typically, both environments have required investments in skills and hardware to support sizable datasets. In recent times, new technology platforms and architectures have emerged that are reshaping—and indeed dismantling—the very concept of big data.

Here are some of the factors leading to the next generation of big data:

> The rise of open source platforms for data analytics—Apache Spark, Kudu, and Impala—provides cost-effective and powerful ways to deliver data and insights in real time to decision makers and systems.

NEXT-GENERATION PLATFORMS HAVE OPEN SOURCE FOUNDATIONS.

The rise of open source platforms for data analytics—Apache Spark, Kudu, and Impala—provides cost-effective and powerful ways to deliver data and insights in real time to decision makers and systems. These platforms aren't necessarily replacing enterprise data warehouses and business intelligence environments—but are modernizing these solutions.

As AI becomes a more pervasive part of enterprise processes and systems, machine learning means algorithms and programs adjust and are constantly refreshed with an influx of data.

NEXT-GENERATION DATA COMES FROM AND RESIDES IN MANY PLACES ALONG THE IoT CONTINUUM.

"Fog computing," for example, is coming to the fore as a solution of choice for many enterprises seeking more distributed options. Fog computing—as defined by the National Institute of Standards and Technology—"facilitates the deployment of distributed, latency-aware applications and services, and consists of fog nodes (physical or virtual), residing between smart end-devices and centralized cloud services."

NEXT-GENERATION DATA RESIDES IN CLOUDS, DATA CENTERS, BEHIND APIs—AND ALL OF THE ABOVE.

The proliferation of cloud services and APIs has opened up new vistas for big data with almost unlimited capacity. Cloud-based services take much of the burden off enterprise shops in terms of managing data. Data may be provided through data as a service environments, in which any and all data, regardless of where it resides, can be surfaced or virtualized to be available to standardized service layers of applications and systems.

NEXT-GENERATION DATA TAKES ADVANTAGE OF STORAGE ANYWHERE, ANYTIME—WITH ASSURED BACKUP AND RECOVERY.

Storage is no longer confined to disk arrays stacked within data centers; it is available on demand, anytime, anywhere, thanks to cloud computing. There is no longer a ceiling on data capacity within on-premise data centers, no longer a need to employ space-saving techniques such as compression or lifecycle management progresses from disk to tape. In addition, data is resilient and recoverable on-demand in less than a second.

NEXT-GENERATION DATA FOSTERS SELF-LEARNING MACHINES.

As AI becomes a more pervasive part of enterprise processes and systems, machine learning means algorithms and programs adjust and are constantly refreshed with an influx of data. To some degree, humans no longer need to be constantly rewriting or reprogramming systems to meet new demands; the data is steering things in new directions.

NEXT-GENERATION DATA TURNS BIG DATA INTO REALLY BIG DATA.

Big data—based on internal corporate data such as transactions and customer records—has gotten even bigger than previously imaginable, expanding wildly in a geometric sense. Data is now flowing in from sources large and small (as small as miniature cameras) and flowing through enterprises.

NEXT-GENERATION DATA FLOWS THROUGH ENTERPRISES IN REAL TIME.

An emerging generation of tools, platforms, and methodologies has driven down the costs of real-time data analytics. Such data may be locked away in ERP systems, as well as externally within partner systems or the Internet of Things. Cognitive tools and technologies such as AI and machine learning are opening up avenues of discovery that are available across real-time data.

NEXT-GENERATION DATA TAKES MANY FORMS.

When the big data revolution kicked into high gear about a decade ago, it was limited to relational data and NoSQL data. Now, data architectures have opened up to imagery and a variety of content types.

NEXT-GENERATION DATA IS NOT TIED TO ANY VENDOR, NOR DATABASE, FOR THAT MATTER.

In recent years, NoSQL databases have grown in popularity, providing cost-effective and cloud-friendly environments for a variety of data types. In addition, the rise of data lakes also detaches data from particular databases, enabling data to be maintained and available for future applications.

—Joe McKendrick

AEROSPIKE

Maximize the Value of Your Operational Data

Modernize and Transform Your Enterprise Via Real-time Transaction and Analysis Processing

OVERVIEW

It might sound too good to be true: a database system that processes large volumes of operational data in real time while delivering exceptional runtime performance, high availability, and cost efficiency while still keeping your data safe. What if early adopters in banking, telecommunications, and other industries are already harnessing such a database for achieving results that are transforming their businesses in myriad ways? What if published benchmarks demonstrate sub-millisecond response times for high throughput read/write workloads over high data volumes with substantial cost savings compared with traditional alternatives?

This paper introduces key technologies that Aerospike clients are using to modernize their data management infrastructures and realize such impressive (and seemingly impossible) results as:

- Rapid read/write speeds without extensive tuning or a separate data cache
- Substantially smaller footprints than popular alternatives, often leading to 3-year total cost of ownership (TCO) savings of $3-5 million per application
- 24x7 availability, including cross-datacenter replication
- Operational ease during scale-out and maintenance
- Interoperation with popular software offerings, including Apache Hadoop, Spark, and Kafka

Sounds unbelievable, right?

THE TECHNOLOGY IN BRIEF

Aerospike provides a distributed, highly scalable database management system for demanding read/write workloads involving operational data. It was designed to deliver extremely fast—and predictable—response times for accessing data sets that span billions of records in databases of 10s – 100s TB. Other design features address fault tolerance and near 100% uptime even during upgrades and maintenance.

How? By capitalizing on proven architectural approaches—such as distributed computing and parallelism—and developing new technologies to meet business demands that hadn't even surfaced when older systems were originally built. Indeed, Aerospike's patented Hybrid Memory Architecture™ (HMA) drastically reduces traditional I/O and network communication compared with other approaches; it also uses CPU resources considerably more efficiently. The cumulative impact of these features (and others) enables Aerospike to deliver remarkable speed at scale.

APPLICATIONS AND USE CASES

Applications that benefit from Aerospike typically share some or all of these characteristics:

- Service-level agreements (SLAs) that require sub-millisecond database response times
- High throughput for mixed workloads (e.g., 3–5 million operations per second)
- Support for managing billions of business records in databases of 10s–100s TB
- High availability and fault tolerance for mission-critical applications
- High scalability for handling unpredictable increases in data volumes and transactions
- Adaptable infrastructure for managing varying types of data with minimal effort
- Low total cost of ownership (TCO)

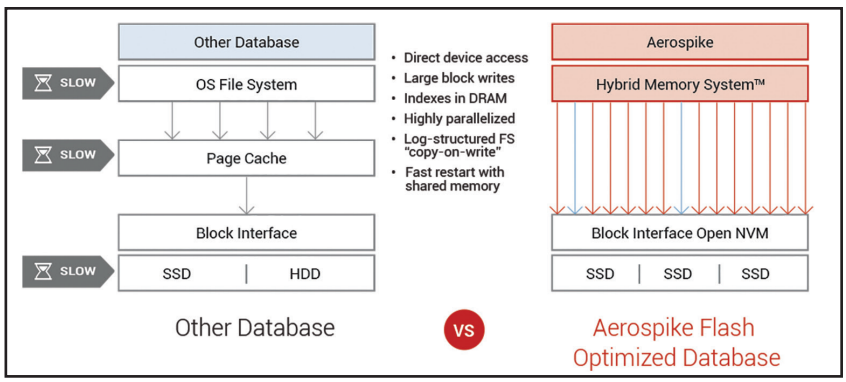
KEY FEATURES AND TECHNOLOGIES

Aerospike is a shared-nothing database system that operates on a cluster of commodity server nodes:

- It's a schema-free, key-value data store.
- Aerospike exploits volatile and non-volatile memory in a distinctive way, providing rapid access to index and user data.
- An intelligent client layer minimizes costly network "hops" needed to access data.
- Immediate record-level consistency and high availability are guiding principles.
- Access management controls and transport encryption protect sensitive data.
- Asynchronous replication across data centers provides disaster recovery.
- Ready-made connectors, a publish/subscribe messaging system, and partner offerings help firms integrate Aerospike into their existing IT infrastructures.

FULL REPORT

To get a copy of the full report, please go to: www.aerospike.com/maximize-operational-data



The End of Single-Purpose Big Data Platforms

The rise of big data led to a generation of platforms engineered for a single purpose, scaling on commodity hardware.

Apache Hadoop can scale to petabytes of data, but it was engineered for offline analytics. It can't meet the performance requirements of data-driven organizations leveraging ad hoc, interactive queries on near real-time data at scale, and with near real-time latency, to drive faster time to insight. Further, big data is no longer limited to analytical workloads, and because it was engineered for offline analytics, Hadoop can't meet the performance requirements of transactional workloads (regardless of the scale). NoSQL databases can, but they're optimized for point and range queries on small to medium working sets cached in memory, not aggregate queries on most, if not all, data stored on disk. NoSQL databases can't meet the scalability requirements of near real-time analytical workloads in a cost-effective manner—and without schemas and transactions, can't be the system of record for business-critical, mission-critical applications.

The only solution was to deploy a combination of Hadoop, NoSQL databases and relational databases in order to meet *both* performance and scalability requirements, and support *both* transactional and near real-time analytical workloads. It was a complex solution whereby data had to be synchronized between different environments and teams via batch imports, messaging systems or both: Hadoop for offline analytics, NoSQL databases for caching and relational databases for transactions. Further, the lack of standard SQL in Hadoop and NoSQL databases resulted in performance issues with traditional BI and reporting tools.

What if a relational database could scale out *and* support both transactional and near real-time analytical workloads? With MariaDB leading the way, the next generation of big data will be handled by relational databases with scalable, workload-optimized storage engines (row-based for transactional, columnar for analytical), replacing single-purpose big data platforms with hybrid databases capable of scaling both transactional and analytical workloads from a single gigabyte of data to hundreds of terabytes of data—all without sacrificing schemas, transactions or SQL.

MariaDB leverages multiple, purpose-built storage engines to support both transactional and analytical workloads at scale. InnoDB, MariaDB's default general-purpose storage engine, supports transactional workloads up to several terabytes of data. The Spider storage engine extends MariaDB with built-in, transparent sharding to support transactional workloads requiring read, write and storage scalability. And MyRocks, a write- and space-optimized storage engine developed by Facebook and supported by MariaDB, can be used with Spider for unrivaled write scalability and storage efficiency. The final storage engine, MariaDB ColumnStore, extends MariaDB with distributed, columnar data and parallel query processing to support near real-time analytical workloads on hundreds of terabytes of data.

MariaDB is available in two configurations, MariaDB TX and MariaDB AX, both with the world's most advanced database proxy, MariaDB MaxScale. MariaDB TX includes InnoDB, MyRocks and Spider, and is optimized for transactional workloads at any scale. MariaDB AX includes InnoDB and ColumnStore, and it is optimized for scalable, high-performance analytical workloads. It's the same database with the same clients, SQL parser and optimizer, but under the covers, different storage engines support different workloads—and with streaming change-data-capture enabled, the same data too.

In a hybrid transactional/analytical topology, MariaDB TX nodes automatically and continuously stream data to MariaDB AX nodes, enabling analytics on near real-time transactional data without the need for separate databases (often from different vendors), import delays or complex ETL processes. The data is stored in a row-based format in MariaDB TX for high-performance transactional queries, with the same data (or a subset of it) stored in a columnar format in MariaDB AX for high-performance analytical queries.

It's the beginning of the end for single-purpose big data platforms for anything less than a petabyte of structured/semi-structured data. MariaDB is the leading enterprise open source database solution, and the only one delivering the best of both worlds in both dimensions: performance and scalability, transactional and analytical.

MariaDB TX
https://mariadb.com/products/solutions/oltp-database-tx

MariaDB AX
https://mariadb.com/products/solutions/olap-database-ax

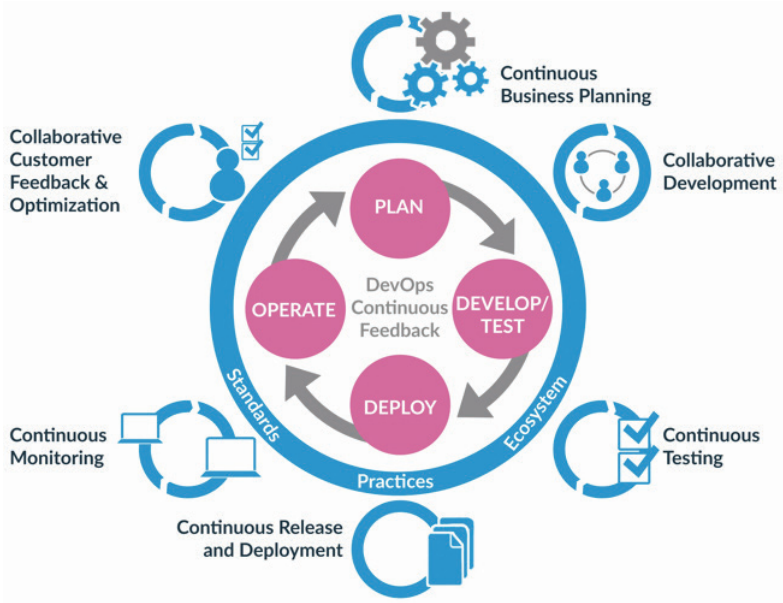
INSIGHTS

Six Pitfalls to Avoid During IT Automation

By André Christ

From planning to maintenance, the wrong tools can wreak havoc on the automation process.

IT DEPARTMENTS ARE under increasing pressure to keep up with the pace of business innovation. As data volumes rise, automation looms as a light at the end of the complex-process tunnel. The ability to leverage software, framework, and application tools to generate automated sequences based on various trigger events has become critical to the IT industry. But, it doesn't come without its challenges. When well-executed, IT automation can increase efficiency and productivity, reduce cost, and speed time to value. However, poorly crafted automated sequences and lack of integration between applications can make a lasting impact on application deployment and management, causing a ripple effect across the entire IT landscape. Deploying faulty automated sequences quickly builds a backlog of errors and makes it extremely difficult to effectively manage architecture down the road. Challenges throughout the IT automation process—from planning and implementation to maintenance and repairs—range from technical to organizational. It's crucial to pinpoint these obstacles before launching new sequences.



DevOps focuses on breaking down walls between business development and IT departments in order to increase speed in software development, enabling teams to effectively collaborate and communicate through the use of collaboration tools.

Rushing the Automation Process

Implementing DevOps and Agile methodologies is critical for future-proofing an IT infrastructure. It enables the organization to be reliable and scalable in the long run. Before kicking off a software automation project, managers need to get a bird's-eye view of how the end result will impact every line of business to ensure that the project is broad enough to be an effective, long-term solution for the organization as a whole. Once overall objectives are established, process methodology should be selected and optimized. Developers, testers, and managers should take multiple automation tools into consideration and become highly efficient in navigating various software applications through leveraging multiple technologies and techniques on scalable platforms—from analysis to testing—prior to launching an automation project.

Failure to Identify Areas of High Risk

Risk mitigation is a cornerstone of enterprise IT automation. Rolling out new automated systems within a large enterprise presents multiple potential points of failure. A lack of consideration for potential risk at the outset almost always leads to pain down the road.

Software automation testers need to create a thorough execution plan that includes a list of all potential risks. Execution plans should include efforts for risk management and mitigation, ranking the threat of the risk at stake, a plan of action for unidentified risk, and risk treatment upon notification. Firewall restrictions, antivirus sequences, and other security measures should also be taken into account throughout the development of the risk execution plan. Testers should become fluent in the security of their products and highly efficient in testing the framework from one step to the next throughout the entire automation process.

André Christ is co-founder and co-CEO of LeanIX (www.leanix.net).

Automating Inaccurate Sequences

Testing and vetting of processes are crucial to the success and efficiency of IT automation. It is vital that developers be attentive to critically examining details throughout the process of automating a complex sequence of events to verify that there are no overlooked morsels of information. Missing one key element or error within the sequence ensures it is endlessly duplicated until it is identified and repaired. Flawed algorithms result in repeated error, ultimately having the potential to harm the business as a whole and, in some cases, even harming customers' reputations. Automating complex processes is incredibly time-consuming, but failing to check and double-check sequences leaves room for potentially fatal undetected errors in the future.

Failure to Actively Maintain Your Automation Process

Once an IT automation initiative is executed, it's not complete. Rolling out an automated sequence isn't a one-and-done business effort. It is necessary to actively tweak the system, develop predictive maintenance techniques that capture and combine data, and use machine learning to identify failure before it occurs. Systems should be optimally programmed to automatically trigger a maintenance request, send a notification to a technician, place an order for a replacement part, and more. It's imperative that organizations identify how to most effectively execute system repairs, conduct proper analyses, and identify additional repairs that should be made simultaneously in order to reduce spending. Before automation deployment begins, ensure project managers compare maintenance strategies to determine optimal cost savings and identify the most efficient processes.

Investing in the Wrong Tools

From planning to maintenance, the wrong tools can wreak havoc on the automation process. Automation tools are used to identify and deploy a series of intricate actions that are prompted by a manual action or by trigger events previously specified in the development of the automation sequence. Failing to invest in optimized, future-forward platforms that support overall automation processes—and aren't limited to one developer group or solving one problem—introduces the potential for easy error. Selected tools should be broad enough to impact daily activities, aid Agile initiatives, and enable teams to manage all software and application tools effectively and efficiently. Before an IT automation project is executed, teams should consider how selected tools will amplify the right processes for the organization, what benefits the organization will gain as a result of leveraging specific tools during the automation process, and how the automated software will affect the organization long term.

Poor Organizational Vision

In addition to technical considerations, there are also challenges in how IT automation will fit organizational processes. IT is no longer seen as a mostly independent technology business unit—now, sales, marketing, and even HR are involved. When deploying organization-wide changes such as automation, executives must take steps toward company-wide transparency and collaboration—as well as ensure alignment throughout the entire organization's business model. Consider adopting a comprehensive, data-first culture that increases communication. The use of multiple training methods that support various learning preferences ensures all employees across all units are aligned with the organization's overall IT architecture.

Most importantly, build your projects around the needs of your team, instead of around the tools themselves. The goal of IT automation is to streamline processes and enable the entire organization to be more productive and efficient. Automation is meant to assist employees, not replace them.

Despite the many challenges of deploying complex automation processes to various systems and software tools within large organizations, if done correctly, automation has the potential to alleviate cross-departmental pain points and time spent on routine tasks. Effectively deployed automated systems will increase productivity, cost effectiveness, collaboration, and transparency in the long run.

The wealth of data now available to organizations, from internet-scale applications to the growth of the Internet of Things, is fueling the use of data lakes, artificial intelligence (AI), machine learning (ML), and predictive analytics solutions.

These and other technology initiatives for managing and analyzing data were explored in a recent Unisphere Research survey. In the report, analyst Joe McKendrick identified key trends that are shaping the way enterprises leverage their data, as well as the evolving priorities of data managers.

Data lakes, places to store diverse datasets without having to build a model first, are perhaps the most mature technology initiatives seen among enterprises in the survey. According to the report, sponsored by Oracle Corp., data lake adoption continues to rise as data managers seek to capture and store data from a wide variety of sources in various formats. Overall, 38% of organizations are employing data lakes as part of their data architecture, up from 20% documented in a 2016 survey.

Another salient point raised by the survey is the trend of Spark moving to the mainstream. "While it remains to be seen what role cloud computing will play in propelling Hadoop engagements, the cloud appears to be a natural setting for its newer sibling, Apache Spark," observed McKendrick, who also noted that as "Hadoop implementations have decreased, it appears some data managers are opting to favor object storage as an alternative to the Hadoop File System." According to the report, object storage, which is designed to contain identifiers, metadata, and files within single units and be deployable across devices and systems, is also relatively new, "but already has a sizable base of adherents."

ML is making its presence felt as well. About one-quarter of respondents' organizations are using ML as part of their data architecture, and another 20% are considering adoption.

Highlighting its newness, the report reveals that a majority of companies with ML, 51%, have been using it less than a year. Moreover, most respondents are either facilitating or are likely to facilitate ML applications in the cloud—33% of respondents with ML in place have implemented it in the cloud, and 21% intend to do so.

Identifying new and disruptive technologies, as well as evaluating when and where they may prove useful, is a challenge in the fast-changing big data market. To contribute to the discussion each year, *Big Data Quarterly* presents the "Big Data 50," a list of forward-thinking companies that are working to expand what's possible in terms of collecting, storing, protecting, and deriving value from data.

We encourage you to explore these solution providers by visiting their websites. You can also gain insight into trends in how data is being managed and consumed by accessing Unisphere Research's survey reports at www.unisphereresearch.com as well as an extensive library of best practices reports and white papers at www.dbta.com/DBTA-Downloads/WhitePapers.

Accenture
www.accenture.com
A global professional services company providing a broad range of services and solutions in strategy, consulting, digital, technology, and operations, Accenture develops and implements solutions to help organizations maximize their performance and achieve their vision.

Actian
www.actian.com
Through the deployment of innovative hybrid data technologies and solutions, Actian, the hybrid data management, analytics, and integration company, helps ensure that business-critical systems can transact and integrate at their very best—on premise, in the cloud, or both.

Aerospike
www.aerospike.com
Providing an enterprise-grade, internet scale, key-value store database with a patented Hybrid Memory Architecture, Aerospike helps enable digital transformation by supporting real-time, mission-critical applications and analysis for companies in the financial services, banking, telecommunications, technology, retail/ecommerce, adtech/martech, and gaming industries.

Amazon Web Services
https://aws.amazon.com
A subsidiary of Amazon.com, Amazon Web Services offers compute power, database storage, content delivery, and other functionality to provide the services organizations need to build sophisticated applications with increased flexibility, scalability, and reliability.

Arcadia Data
www.arcadiadata.com
Founded with the goal of connecting business users to big data, Arcadia Data provides visual analytics and BI software that run natively within modern data platforms such as Apache Hadoop and the cloud to analyze large volumes of data without moving it.

Attunity
www.attunity.com
Supporting availability, delivery, and management of data across heterogeneous platforms, organizations, and the cloud, Attunity solutions span data replication and distribution, test data management, CDC, data connectivity, enterprise file replication, managed file transfer, DW automation, data usage analytics, and cloud data delivery.

BackOffice Associates
www.boaweb.com
A provider of information governance and data stewardship solutions focused on helping customers manage one of their most critical assets—data—BackOffice Associates enables organizations to accelerate growth, gain actionable visibility, and reduce risks.

Bedrock Data
www.bedrockdata.com
Helping teams work better and forge deeper connections with their customers, Bedrock Data provides Fusion, which automates the process of unifying customer data across cloud applications, to create an on-demand SQL warehouse that enables Customer 360 analytics, BI, and real-time dashboards.

BlueData
www.bluedata.com
Improving how enterprises deploy big data analytics and machine learning, BlueData provides a big-data-as-a-service platform that uses Docker container technology to make it easier for enterprises to innovate with big data and AI technologies.

Cambridge Semantics
www.cambridgesemantics.com
A big data management and enterprise analytics software company, Cambridge Semantics offers the Anzo Smart Data Lake to allow IT groups and their business users to semantically link, analyze, and manage diverse data—whether internal or external, structured or unstructured.

Cloudera
www.cloudera.com

Helping to transform complex data into actionable insights, Cloudera recently launched a new version of the Cloudera Data Science Workbench, a platform that lets data scientists manage their own analytics pipelines, thus accelerating machine learning projects from exploration to production.

CloverETL
www.cloveretl.com

Developed by Javlin, a provider of solutions and services for mass data processing, CloverETL provides a data integration software suite that makes rapid development and enterprise capabilities available in a light footprint package.

Collibra
www.collibra.com

A provider of data governance and catalog software, Collibra helps organizations gain competitive advantage by maximizing the value of their data across the enterprise and addressing the gamut of data stewardship, governance, and management needs of the most complex, data-intensive industries.

Couchbase
www.couchbase.com

Built with NoSQL technology, the Couchbase Data Platform was architected for "the massively interactive enterprise," and offers its geo-distributed Engagement Database to provide developer agility and manageability, as well as high performance at any scale—from cloud to the edge.

DataStax
www.datastax.com

Powering the "Right-Now Enterprise" with an always-on, distributed cloud database, built on Apache Cassandra and designed for hybrid cloud, DataStax Enterprise gives businesses full data autonomy, allowing them to retain control and strategic ownership of their data in a hybrid cloud world.

Denodo Technologies
www.denodo.com

A provider of data virtualization software, Denodo helps organizations to achieve faster and easier access to unified business information for agile BI, big data analytics, web and cloud integration, single-view applications, and enterprise data services.

erwin
https://erwin.com

Now a standalone company, erwin, Inc. announced its evolution to "the data governance company" in January 2018 and, as part of this focus, acquired A&P Consulting, a technology and consulting services firm based in Italy.

Franz
https://franz.com

An early innovator in AI and supplier of semantic graph database technology with expert knowledge in developing and deploying cognitive computing solutions, Franz counts dozens of Fortune 500 companies among its customers—spanning the healthcare, intelligence, life sciences, telecommunications, and research sectors.

HVR
www.hvr-software.com

Providing a real-time data integration solution that supports enterprise digitization efforts, the HVR platform offers a reliable, secure, and scalable way to move large data volumes in complex environments, enabling real-time data updates, access, and analysis.

IBM
www.ibm.com

One of the largest IT companies in the world, IBM is "putting smart to work" with innovative big data solutions and services spanning AI and machine learning, cloud, blockchain, Hadoop, and Spark.

IDERA
www.idera.com

Providing database lifecycle management solutions that allow database and IT professionals to design, monitor, and manage data systems, whether in the cloud or on-premise, IDERA also offers a portfolio of free tools and educational resources to help users gain the knowledge they require.

InfluxData
www.influxdata.com

Delivering a complete time series platform built specifically for metrics, events, and other time-based data—including data from humans, sensors, and machines—InfluxData empowers developers to build next-generation monitoring, analytics, and IoT applications that deliver real business value quickly.

Informatica
www.informatica.com

Helping organizations to "unleash the disruptive power of data," Informatica provides the Informatica Intelligent Data Platform, powered by CLAIRE metadata-driven AI, to ensure automated delivery of data for self-service access by people, applications, and machines.

Looker Data Sciences
https://looker.com

Used at more than 1,300 companies such as Sony, Amazon, The Economist, IBM, Spotify, Etsy, Lyft, and Kickstarter, Looker offers a complete data platform that provides data analytics and business insights to all departments and easily integrates into applications to deliver data directly into the decision-making process.

MapR Technologies
https://mapr.com

With its data platform for AI and analytics, MapR addresses the complexities of high-scale and mission-critical distributed processing and helps enterprises inject analytics into their business processes to increase revenue, reduce costs, and mitigate risks.

MariaDB
www.mariadb.com

MariaDB offers a relational database to support enterprise needs from online transaction processing to analytics from a single SQL-compliant interface with open source MariaDB Server, a relational database, and complementary products including MariaDB, MaxScale, and MariaDB ColumnStore.

MarkLogic
www.marklogic.com

Integrating data from silos, MarkLogic's operational and transactional enterprise NoSQL database platform empowers customers to build modern applications on a unified, 360-degree view of their data.

Melissa
www.melissa.com

Providing global address, phone, email, and name identity verification solutions and data enrichments, Melissa's goal is to maximize the effectiveness of business intelligence, big data analytics, and omnichannel marketing initiatives.

Microsoft
www.microsoft.com

Microsoft offers an array of technologies and solutions for businesses of all sizes, spanning desktop applications, relational database management technology, operating systems, search, and mobile devices, in the cloud and on-premise.

MicroStrategy
www.microstrategy.com

Built to help organizations quickly deploy sophisticated analytical and security applications at scale, MicroStrategy delivers innovative software that empowers people to make better decisions and transform the way they do business.

MongoDB
www.mongodb.com

MongoDB uses a document data model that is similar to JSON and recently added support for multi-document ACID transactions to provide a globally consistent view of data across replica sets and enforce all-or-nothing execution to maintain data integrity.

Oracle
www.oracle.com

Helping organizations to devote more time and resources to adding value for their users and customers, Oracle provides capabilities in SaaS, platform as a service, infrastructure as a service, and data as a service from data centers throughout the world.

Progress
www.progress.com

With a platform for building and deploying applications, Progress offers flexible front-end tooling for delivering a multi-channel UX; a scalable, secure back end to build and run microservices; and data connectivity capabilities from any source.

Pure Storage
www.purestorage.com

Pure's data solutions enable SaaS companies, cloud service providers, and enterprise and public sector customers to deliver real-time, secure data to power their mission-critical production, DevOps, and modern analytics environments in a multi-cloud environment.

Quest Software
www.quest.com

Quest Software offers solutions that include information management, data protection, endpoint systems management, identity and access management, and Microsoft platform management that can reduce the time and money spent on IT administration and security.

RedPoint Global
www.redpointglobal.com

RedPoint Global's software solutions are aimed at empowering brands to transform how customer experience is delivered, providing a single point of control to connect customer data, determine actions in real time, and orchestrate interactions across all enterprise touchpoints.

Reltio
www.reltio.com

Providing the Reltio Self-Learning Data Platform, developed natively in the cloud and enhanced with machine learning, Reltio organizes data from all sources and formats, creating a unified dataset with personalized views for users across sales, marketing, and compliance.

Robin Systems
https://robinsystems.com

Seeking to disrupt the virtualization market, Robin Systems brings together purpose-built, container-aware block storage with application-aware management in the cloud (private and/or public), demonstrating benefits to distributed, clustered, and stateful applications including big data and databases.

SAP
www.sap.com

Known for HANA, its platform for next-generation applications and analytics, SAP is a global provider of enterprise application software that empowers people and organizations to work together more efficiently and use business insight more effectively.

SAS Institute
www.sas.com

SAS is a provider of business analytics software and services across areas such as advanced analytics, business intelligence, customer intelligence, and data management that empowers and inspires customers around the world to transform data into intelligence.

SnapLogic

www.snaplogic.com

Delivering an elastic integration platform as a service to connect cloud applications and disparate data sources, SnapLogic offers its Enterprise Integration Cloud to make it fast and easy to connect applications, data, APIs, and things.

Software AG

www.softwareag.com

With Software AG's Digital Business Platform, companies can better interact with their customers and bring them on new "digital" journeys, promote unique value propositions, and create new business opportunities.

SQream

https://sqream.com

Developer of SQream DB, a GPU database designed to enable business insight from massive data stores, SQream allows enterprises to analyze more data than ever before, while achieving improved performance, reduced footprint, and cost savings.

Striim

www.striim.com

The Striim (pronounced "stream") platform is an enterprise-grade, real-time data integration and intelligence solution, making it easier to ingest and process high volumes of streaming data—including change data capture—for real-time log correlation, cloud integration, edge processing, and streaming analytics.

Syncsort

www.syncsort.com

A trusted enterprise software provider and the global leader in "Big Iron to Big Data" solutions, Syncsort helps customers optimize traditional data systems and deliver mission-critical data from these systems to next-generation analytic environments.

Tableau

www.tableau.com

Tableau helps customers see and understand data through visual analytics, allowing them to build dashboards and perform ad hoc analyses in just a few clicks to share work with anyone, anywhere.

Teradata

www.teradata.com

With a portfolio of cloud-based business analytics solutions, architecture consulting, and industry leading big data and analytics technology, Teradata unleashes the potential of companies to achieve high-impact business outcomes.

Vertica

www.vertica.com

The Vertica Analytics Platform is purpose-built for big data analytics, designed for use in data warehouses and other big data workloads where speed, scalability, simplicity, and openness are crucial to the success of analytics.

VMware

www.vmware.com

VMware's compute, cloud, mobility, networking, and security offerings provide a dynamic and efficient digital foundation to more than 500,000 customers globally, aided by an ecosystem of 75,000 partners and powering the world's complex digital infrastructure.

VoltDB

www.voltdb.com

VoltDB provides an in-memory translytical database for applications that require a combination of data scale, real-time analytics, volume, and accuracy for use in telco, financial services, ad tech, gaming, and other industries.

BackOffice Associates

Rex Ahlstrom, Chief Strategy and Technology Officer

As today's enterprises are faced with managing enormous amounts of data, ranging from financial and HR to customer, sales, manufacturing and others, they are largely shifting away from legacy systems and onto cloud-based business suites that allow them to leverage data more strategically and collaboratively.

As part of this trend, organizations are embarking on digital transformation initiatives—which also require a data transformation—in order to foster innovation, drive key process improvements and achieve compliance with industry and legal regulations. And, in doing so, business teams are revamping the way they manage both structured and unstructured (big) data, including applying AI and machine learning, as well as implementing a crowd-sourced information governance approach to capture the cross-departmental, global knowledge of their teams.

At BackOffice Associates, we focus on helping organizations solve their most complex enterprise data transformation challenges through our unique combination of data expertise, intelligent software and solution accelerators that power digital transformation efforts. We've helped thousands of enterprises ensure their data is continuously up-to-date, accurate and immediately actionable for driving intended business results that positively impact their bottom line. Our solutions incorporate AI and machine learning to deliver refined expert guidance across the full data journey, including data migration, data quality management, metadata management, information governance, analytics and archival.

As enterprise digital transformations continue to unfold, BackOffice Associates remains committed to delivering industry-leading expertise and solutions to help our customers achieve data excellence—ultimately yielding powerful and sustainable business results.

BackOffice Associates
www.boaweb.com

Denodo

Ravi Shankar, CMO

Big data has given rise to data lakes and enabled many use cases that were not possible earlier—multi-dimensional analysis, IoT integration, and offloading data storage among other things. However, the earlier presumption that data lake will become the single repository for the entire enterprise has not materialized. The reason being that many enterprises have still kept the legacy technologies while adding new data lakes for specific lines of businesses like marketing, sales, and customer service. As a result, the data has become further siloed across the new big data systems and legacy infrastructure.

Fortunately, there is a way to unify the data across the enterprise irrespective of its location, format, or latency, and all without having to replicate the data into another central repository. That solution is data virtualization. It is part of a big data fabric that integrates the entire enterprise information and delivers it to business users in real time. As a data abstraction layer, it hides the complexities of accessing data from the underlying data systems, their formats and structures. Unlike ETL (Extract-Transform-Load), it does not replicate the information into another repository.

Denodo is a leader in data virtualization. Many Fortune 1000 organizations like Autodesk, Logitech, and Vizient rely on Denodo for enabling critical big data solutions within their organizations. Analysts such as Forrester Research have acknowledged Denodo as "Today, several large Fortune 1000 companies leverage Denodo to support their mission-critical data virtualization strategies." Denodo continues to innovate on data virtualization with support for in-memory fabric, data cataloging, and enterprise-wide deployments.

It is easy to take Denodo on a self-test-drive.
Visit www.denodo.com/test-drive.

Denodo
www.denodo.com

erwin, Inc.

Adam Famularo, CEO

erwin Helps Solve the Enterprise Data Dilemma

Most organizations don't use all the data they're flooded with to reach deeper conclusions about how to grow revenue, achieve regulatory compliance or make strategic decisions. They face a data dilemma: not knowing what data they have or where some of it is—plus integrating known data in various formats from numerous systems without a way to automate that process.

To accelerate the transformation of business-critical information into accurate and actionable insights, organizations need an automated, real-time, high-quality data pipeline. Then every stakeholder—data scientist, ETL developer, enterprise architect, business analyst, compliance officer, CDO and CEO—can fuel the desired outcomes based on reliable information.

erwin delivers the integration and automation that simplifies the total enterprise data management and governance lifecycle. Its persona-based erwin EDGE Platform creates an "enterprise data governance experience" that facilitates collaboration between both IT and the business to discover, understand and unlock the value of data both at rest and in motion.

By bringing together enterprise architecture, business process modeling, data mapping and data modeling around a data governance hub, the erwin EDGE helps organizations get a handle on how they manage their data. The broadest set of metadata connectors and automated code generation, data mapping and cataloging tools gives users the most agile, efficient and cost-effective means of launching and sustaining comprehensive data governance.

Ultimately, the erwin EDGE provides greater efficiencies to technical users and better, more dependable analytics to business users.

erwin, Inc.
www.erwin.com

Franz Inc.

Jan Aasman, CEO

AllegroGraph—Semantic Graph Database for Knowledge Graphs

Artificial Intelligence (AI) is one of the top investment areas for companies looking to improve ROI on operations and products, and to create customer 360 views. Using AI to create "Enterprise Knowledge" and link it across the Enterprise to create a "Knowledge Graph" is a key differentiator for companies in an ever-increasing competitive landscape. The foundation for Knowledge Graphs and Artificial Intelligence lies in the facets of semantic technology provided by Franz's AllegroGraph database. Semantic Graph databases, such as AllegroGraph, provide the core technology environment to enrich and contextualize the understanding of data. The ability to rapidly integrate new knowledge is the crux of the Knowledge Graph and depends entirely on semantic technologies.

An early innovator in Artificial Intelligence, Franz Inc. is a leading supplier of Knowledge Graph solutions with Semantic Graph Database technology as the foundation. If you really want to develop your corporate Knowledge Graph and address complex Artificial Intelligence problems, you need a data system that goes beyond just data. You have to create a system that can link to anything outside your own predefined parameters—and that can learn from previous experiences. That is where a Semantic Graph Database, like AllegroGraph, comes into the picture.

Franz Inc. provides a variety of services as part of its Knowledge Graph platform solution: from architectural consulting and technical seminars to training. Franz's flagship product, AllegroGraph, provides the necessary power and flexibility to address high-security data environments such as HIPAA access controls, privacy rules for banks, and security models for policing, intelligence, and government.

Contact Franz Inc. to unleash the potential of your Company's Knowledge Graph.

Franz Inc.
https://franz.com/

IDERA Software

Heidi Farris, VP & GM of Database Tools

While relational platforms continue to comprise the majority of database deployments around the world, NoSQL and Big Data platforms have increased in popularity for data warehouses and other data-intensive configurations that require a more flexible approach to managing data. IDERA Software enables database professionals with enterprise architecture and database development tools that help them work with diverse data sources in their organizations.

ER/Studio is the company's flagship data modeling and architecture suite used by many companies to build the enterprise model foundation for their data governance and compliance programs. In 2015, ER/Studio added round-trip data modeling support for Big Data platforms such as Hadoop Hive and MongoDB, including the ability to show nested object relationships. The addition of these popular NoSQL databases along with cloud, data warehouse, and relational platforms provides a comprehensive view of the enterprise data landscape.

Aqua Data Studio by AquaFold is a recent addition to the IDERA Software database tools portfolio. This impressive database administration and development tool includes support for over 28 diverse data sources. These include relational, cloud, NoSQL, and data warehouse platforms, among others. MongoDB, Hive, and other NoSQL and Big Data sources have been available since 2013. The latest release adds Visual Explain Plans for Hive, Spark, and Impala, along with MongoDB support for Views, Decimal128 Data Type, and Collation.

IDERA's database lifecycle management solutions allow database and IT professionals to design, monitor and manage complex data systems with complete confidence.

IDERA Software
www.idera.com

RedPoint Global

Dale Renner, CEO

Digital transformation is forcing brands to reimagine their customer engagement models at lightning speed or risk falling behind. Accomplishing digital transformation requires a deep understanding of customers that can only be achieved by collecting multiple types of data at multiple cadences and connecting it all so that organizations can deliver the most relevant, consistent, and contextually aware engagement as a driver of profitable revenue growth.

The RedPoint Customer Data Platform™ (CDP) is an award-winning data technology that provides a customer golden record, yielding a unified view across all data silos and technologies. Regardless of source—batch or streaming, internal or external, structured or unstructured, transactional or demographic, personal or general—the RedPoint CDP provides an always on, always updating golden record and makes it continually available at low latency to all touchpoints and users across the enterprise when needed, and ready for purpose.

This includes RedPoint's real-time customer view capabilities, which empower marketers and business users with an easily accessible single view of activity across all touchpoints coupled with prioritized next-best actions. RedPoint's data-driven technology enables hyper-personalized interactions that reflect each consumer's unique needs, preferences, and styles at their speed; businesses are also able to access RedPoint's advanced matching algorithms and Master Data Management (MDM) functionality to resolve anonymous-to-known customer identities with unparalleled speed, precision and efficiency.

For more information about the RedPoint Global Customer Data Platform and all of RedPoint's data technologies, visit www.redpointglobal.com.

RedPoint Global
www.redpointglobal.com

SnapLogic

Craig Stewart, VP of Product Management

As companies hit the wall with on-premise systems, they are looking to the cloud to take advantage of promised cost savings, nearly limitless data processing power, and instant scaling options. But connecting cloud-based big data environments with diverse data sources, while also creating Apache Spark pipelines to transform that data, has typically required highly technical knowledge and continuous coding resources, resulting in soaring operational costs and long time-to-value.

To help enterprises overcome these barriers and harness data as a strategic asset, SnapLogic introduced SnapLogic eXtreme earlier this year enabling big data engineers and technical business users to process large volumes of data, without complex code. Just as SnapLogic democratized app and data integration for IT and citizen integrators, SnapLogic is extending its capabilities to bring the same time and cost benefits to data engineers integrating big data services.

Via SnapLogic's graphical user interface, data engineers can land data in storage services like Amazon S3 or Azure Data Lake Store using 450+ pre-built connectors. The engineers can then quickly create transformative Apache Spark pipelines with SnapLogic's ephemeral plex capabilities to easily process large volumes of data from various endpoints using managed big data services like Amazon EMR and Microsoft Azure HDInsights. This results in a substantially lowered barrier to creating cloud-based big data architectures for data engineers while empowering companies with massive cloud compute power and cost efficiencies.

SnapLogic eXtreme benefits include:

- Managed data architecture in the cloud: Automated, fully managed cloud-based big data runtime environment includes integration (iPaaS), processing (BDaaS), and data storage.
- Self-service: A drag-and-drop interface empowers technical business users, or citizen integrators, to save time and eliminate the IT bottleneck.
- Scale: The SnapLogic platform populates the cloud data lake by leveraging 450+ pre-built Snaps for Hadoop, Kafka, Cassandra, MongoDB, AWS Redshift, and other applications and data stores.
- Accelerated time-to-value: Move big data to the cloud quickly to generate business value faster.
- Compatibility: Easily connect managed Hadoop services like Amazon EMR, Microsoft Azure HDInsight, and more

SnapLogic
www.snaplogic.com

Sqream

David Leichner, CMO

Due to exponentially growing data, organizations today are facing slowdowns, with analytics taking hours or even days. Time-consuming preparation is needed for each change in perspective, and some complex analytics simply cannot be done.

SQream has redefined big data analytics with SQream DB, a complementary SQL data warehouse harnessing the power of GPU to enable fast, flexible, and cost-efficient analysis of massive datasets of terabytes to petabytes. SQream's powerful technology breezes through trillions of rows of data, getting you results up to 100x faster.

SQream DB integrates seamlessly into enterprises' MPP ecosystems—whether on-premise or on the cloud—drastically reducing query times and enabling previously unobtainable business intelligence. With standard SQL syntax as well as ODBC, JDBC, .NET, Node.js and Python connectivity, SQream DB is already supported.

With SQream DB, raw data is analyzed directly, enabling data scientists and BI analysts to ask more questions about more data from a variety of perspectives without the need for arduous preparation.

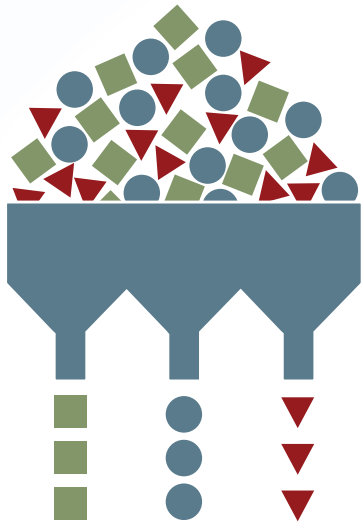
Leading organizations in telecom, retail, healthcare, finance and additional industries around the world use SQream DB to accelerate business intelligence and gain access to a world of never-before-seen insights.

Sqream
https://sqream.com/

BIG DATA BY

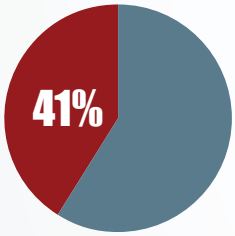
DATA GOVERNANCE: BRINGING ORDER TO CHAOS

Increasing data volumes can add rich insights but also exacerbate the complexity of finding and protecting data. These issues are propelling the growth of the data governance market, which is estimated to be $1.31 billion in 2018 and expected to expand to reach $3.5 billion by 2023, representing a 22% compound annual growth rate, according to a MarketsandMarkets report released in April 2018.

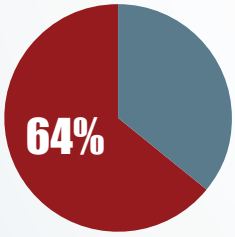


RISK AND REGULATORY EXECUTIVES IN THE AMERICAS AND ASIA-PACIFIC (APAC) REGIONS HAVE IDENTIFIED DATA LINEAGE, DATA GOVERNANCE, REGULATORY CHANGES, AND COMPLIANCE RESOURCES AS THEIR BIG FOUR CHALLENGES.

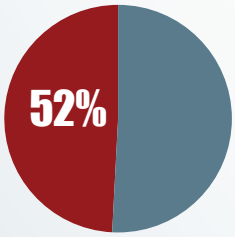
41%
of America's executives say they trust the accuracy of their data, representing an 11-point fall-off from 2017



64%
of respondents in the Americas say there is a need to improve data aggregation, lineage, and reporting



52%
of America's executives identified a need to strengthen data governance as one of their biggest challenges



Source: AxiomSL's 2018 survey, focused on financial services executives' outlook on the regulatory landscape

TO COMPLY WITH THE GENERAL DATA PROTECTION REGULATION (GDPR), ORGANIZATIONS MUST KNOW THE TYPE, VALUE, AND LOCATION OF THE INFORMATION THEY RETAIN, AND ALSO BE ABLE TO DELETE, CHANGE, OR PROVIDE DETAILS ON THE INFORMATION THEY HOLD.

34% of executives will sometimes let operational and cost concerns take precedence over compliance

57% train staff on data protection compliance

50% identify internal staff and practices as their biggest data threat

38% cite external hackers as the biggest threat to data security

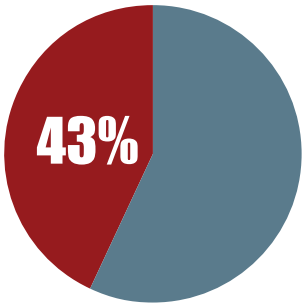
41% have no system in place to determine data origin and quality

3% have fully automated processes with audit trails

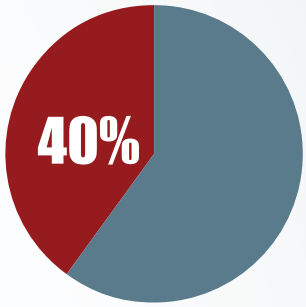
Source: "Top Corporate Data Protection Challenges" from the Compliance, Governance and Oversight Council

THE NUMBERS

A RECENT SURVEY UNCOVERED A LACK OF STAFF WITH GDPR EXPERIENCE.



43% cite a lack of expert staff as their primary GDPR compliance challenge



40% cite a lack of budget

Source: "2018 GDPR Compliance Report" released by Crowd Research Partners, in partnership with Alert Logic, AlienVault, Data443, D3 Security, Haystax Technology, and Securonix, April 2018

DATA MANAGEMENT AND GOVERNANCE SHORTCOMINGS AT ORGANIZATIONS HAVE RESULTED IN POOR DATA QUALITY, LACK OF CONFIDENCE, AND THE SCARCITY OF COLLABORATIVE FRAMEWORKS.

20% ..... of business and IT executives say users can find trusted data on their own

18% ..... say data lineage can be tracked to a source

48% ..... say that users spend at least 61% of their time finding and preparing data for analytics

50% ..... have no formal data governance strategy

38% ..... are only somewhat confident about the data lineage of the data used in reports and analytics and 18% have no confidence in the data lineage

Source: "Reducing Inefficiency and Increasing the Value of Analytics and Business Intelligence," a survey report created by TDWI and commissioned by Datawatch, April 2018

THERE IS BROADER AWARENESS AND FOCUS ON DATA GOVERNANCE BUT THERE ARE STILL CHALLENGES IN TERMS OF UNDERSTANDING, EXECUTIVE SUPPORT, AND FINANCING.

98% view data governance as important or very important from a business perspective

46% don't have a formal data governance strategy in place

63% don't have a budget for data governance or don't know if they have one

40% say the IT department still takes responsibility for data governance costs

21% are just getting started with data governance

Source: "2018 State of Data Governance Report" commissioned by erwin through UBM, February 2018

TRENDING NOW

Enter the Chief Data Officer

Q&A With Accenture Applied Intelligence's Ramesh Nair

Ramesh Nair,
North America Financial Services Leader

TODAY, ORGANIZATIONS EVERYWHERE want to become data-driven—directed by business knowledge rather than intuition or experience from the past. But to reach that level of sophistication, they need someone at the helm who is capable of forming a strategy for data usage and governance. Enter the "chief data officer," or CDO. Recently, *Big Data Quarterly* spoke with Ramesh Nair, North America Financial Services leader at Accenture Applied Intelligence, about why more companies are putting executives in that role.

—*J.W.*

'The financial services industry is the leading sector embracing the adoption of the CDO ... within financial services, data is the product—unlike other industries where data is in support of a "physical" product.'

Are more companies putting chief data officers (CDOs) in place?

Absolutely, we have seen a sharp rise in the number of CDOs throughout organizations regardless of size in recent years. Business executives continue to have even greater expectations from their data, but many are challenged to keep up with today's data demands and sophistication of analytics and technologies. They need a seasoned and knowledgeable senior executive to lead them through this change. This is why the role of CDO becomes critical and why many more companies will put a CDO in place—it offers true strategic advantage about data at the enterprise level.

What is their role and responsibility?

It's an interesting question as the role is still relatively new and definitely still evolving. A couple of years earlier, I would have said that the CDO would be focused solely on data quality, data management, and regulatory/compliance. If you look at the role today, especially with the proliferation of big data, CDOs are increasingly responsible for much more—innovation and growth agendas, transformation from legacy systems to new big data ecosystems, and looking ahead to position their companies for next-generation technologies such as AI and blockchain. I think that this continual evolution will be the case for the foreseeable future.

Is there consensus among organizations on this role or does it vary substantially at different companies?

Although we're starting to see some common themes emerge, the role of the CDO often changes substantially from company to company as it is still relatively new and evolving. Earlier this year, Accenture Applied Intelligence hosted a 2-day CDO Roundtable for Financial Services. We had representation from companies across the financial services industry—banking, capital markets, insurance—all with CDOs in place, all of which had different responsibilities and different bosses. For example, we had some CDOs who reported to the chief information officer, some to the chief operations officer, while others reported to the chief risk officer. Within each of these groups, their responsibilities varied. Some were responsible for data management and governance, others for technology and operations, while others only focused on data management and big data analytics. It was definitely interesting to see how much each CDO's responsibilities varied just within this group we hosted.

Is there an industry sector that is embracing adoption of CDOs?

We are seeing that the financial services industry is the leading sector embracing the adoption of the CDO. It makes sense when you take a step back and look at the data ecosystem for financial services companies. They process massive amounts of data on a daily basis, have to plan and navigate through an ever-changing regulatory landscape, need to protect data from sophisticated attacks on a daily basis, and have to figure out how to keep evolving their business strategy. The need for a CDO in these companies is very real, and this industry has taken the lead on embracing this role. As some CDOs noted, within financial services, data is the product—unlike other industries where data is in support of a "physical" product.

How does this job compare to that of the CIO?

The two roles should be complementary to each other. The CIO has traditionally been, and will continue to be, focused on the IT aspects of data—storage, volume, recovery, security, ingestion, access, and availability—and enabling the rest of the C-suite. The CDO is focused on taking data and turning it into a corporate asset to drive business growth. We also found that a number of the CDOs at the roundtable were actually working closely with their

technology counterparts to jump-start their efforts. Ultimately, it will be important that these two work in tandem in order to achieve a successful digital transformation for their organization.

Why are CDOs needed now?

CDOs are needed more than ever. As the value, centrality, and complexity of data to the enterprise continues to increase, organizations need someone to create a strategy for optimizing data for competitive gain, executing on the delivery of this strategy, and extracting value from their data strategy. The CDO will be the point person to accelerate their company's journey from the traditional data warehouse to next-generation data capabilities.

What are the challenges CDOs are facing?

Our roundtable participants provided some practical insight into what's keeping them up at night. These challenges fall into three broad categories—getting the foundational data capabilities right—metadata management, data quality, data lineage, and data governance; extracting value from their current big data investments; and preparing for the future. Speaking of the current big data investments, two major challenges that were discussed at length were how to decommission and migrate legacy data and analytical environments, and how to transform talent and culture. Although these are significant challenges, they can stand out as great opportunities for CDOs to demonstrate value.

How does architecture factor into the success of executives in this position?

The CDOs at the event actually weighed in on this exact question. The group agreed unanimously that enterprises are headed toward the data lake construct—a move which brings performance and scalability powers unseen until now. From a conceptual architecture point of view, the fundamental notion is that of a shared concept. The group, however, questioned whether this current approach will be sustainable through the increasing explosion of data volumes and types, the wider dispersion and adoption of artificial intelligence and machine learning capabilities, and the ultra-fast performance required for companies to operate in real time. At the end of the day, the group agreed that architecture will play a big role in their continued success and that we need to start thinking about an enlightened approach to building future-proof architecture.

Do you see the role of the CDO becoming more important in the future?

I think the role will definitely become more important in the future—and sooner than you might think. We found this to be consistent with the group that we hosted at the roundtable, where roughly 50% of these CDOs were reporting directly to business leaders in the C-suite. The CDO will be the role directly accountable for guiding their organization and accelerating it to the "new." The importance of this role will be highlighted as companies prepare for and make this journey.

What needs to change in order for CDOs to enact the changes organizations seek?

CDOs will need to take a holistic approach to their mandate for fixing the core foundation—data, technology, processes, and talent—while leading a transformation across the enterprise. Ultimately, though, to be successful in this role, CDOs must rely on their ability to step outside their comfort zones and seek broader remits and decision rights.

melissa®

Unison Makes Master Data Management Secure, Fast, and Hassle-Free

Unison is a data steward's best friend. Melissa built Unison as an ideal centralized solution to establish and maintain contact data quality—without programming. The holistic platform brings data standardization, validation and enrichment together for end-to-end data quality. With Unison, you can cleanse sensitive customer information throughout the enterprise—fast, easy and hassle-free. Development time is completely eliminated, and data never leaves your organization.

COMMAND A UNIFIED ARCHITECTURE FROM A SINGLE CUSTOM CONSOLE

The platform is a web-based client-server application. The login portal is accessible through local intranets in any modern browser. Unison unifies all of Melissa's data cleansing technologies through its straightforward, modern and powerful user interface. Unison's elegant interface is organized by the steps of the data management workflow and is fully customizable.

COMPREHENSIVE TOOLS FOR RECORD VERIFICATION AND ENRICHMENT

Unison validates U.S. and Canadian addresses and performs aggressive address corrections in a single console. It leverages a SERP and CASS Certified™ address engine to match and correct spellings and naming mistakes for cities and streets, and to add the correct street name suffix, prefix and ZIP+4 information. It can append latitude and longitude coordinates with Census data, validate and standardize email addresses and phone numbers, plus validate and parse through full names. Unison streamlines the data preparation process and offers the tools to perform thorough dataset analysis.

RAPID DEPLOYMENT, ZERO DEVELOPMENT

For those who are concerned about implementation and development time, there's no need to worry. Because no coding is required, Unison is effortless to install on premise. Melissa built everything inside a handy platform, so you can implement this solution fast. Unison requires no technical knowledge. Once installed by IT on your network, updates to Unison are distributed directly to your IT team by Melissa. If your Unison installation has internet connectivity, updates will prepare themselves on your servers, and you'll get notified once its ready.

ENJOY FLEXIBLE SCALABILITY

Unison supports horizontal scaling, so you can maximize existing hardware and any additional hardware you choose to add in the future. Scaling horizontally allows you to spread work across multiple servers and begets huge leaps in processing speed. Docker Swarm tears the roof off any performance ceiling by easily spreading Unison containers across as many processors as you choose to configure. The only limiting factor will be your network speed.

BUILT-IN COLLABORATION ENHANCEMENTS

Ideally, a number of data stewards can address data quality issues across the enterprise quickly and securely. Administrators can choose to integrate Unison within the company's LDAP system for preexisting logins, or create Unison account logins for data stewards to begin processing projects. Accessing Unison by multiple stewards is that simple!

Role-based capabilities and configurable user rights ensure secure access and management as well as effective project collaboration. Connect to multiple RDBMS platforms and schedule jobs to process during off hours to maximize performance, then visualize reports with analytics to better collaborate on projects.

ADVANCED, INTEGRATED REPORTING FOR SUPERIOR ANALYTICS

Beautiful, easy-to-read reports are automatically generated per job to provide a high-level overview of how your data improved. This gives you both the high level overview of operations on your data, as well as a detailed report on how your data changed. Unison's reporting system offers a drill-down at the record level for each section of reports, demonstrating exactly which record sets contain specific result codes. You can now schedule jobs to run as often as needed, and have the ability to filter data to multiple locations based on result codes.

SACRIFICE NEITHER SPEED NOR SECURITY

Because Unison works completely offline, data can be managed confidently on-site to meet compliance and security requirements. Even within the platform itself, Docker containerization security features leave assets isolated and self-contained, making Unison ideal for industries with sensitive data that must remain on-premise.

Streamline data prep workflows, reduce analytics busy work, gain more insights and increase efficiency with Unison.

Try a product demo. Connect with Melissa's global intelligence team at www.melissa.com or call 1-800-MELISSA.

Data Operations Problems Created by Deep Learning

Thanks to the dramatic uptick in GPU capabilities, gone are the days when data scientists created and ran models in a one-off manual process. This is good news because the one-off model was typically not optimized to get the best results. The bad news is that with an increase in the total number of models created—including iterations over time—the amount of data used as inputs and generated by the models quickly spirals out of control. The additional bad news is that there are a variety of complexities associated with model, data, job, and workflow management.

The typical starting point for any deep learning application is with the data sources, and as the number of sources grows, so does the complexity of the data management problem. Over time, each data source may be expanded with the addition of new data, or enriched with metadata or data sources. To be clear, data cannot be versioned in the way that software or deep learning models can by using a version control system such as Git. In addition, the data sources must be versioned in lockstep with both the software and the models. It is imperative that the data be versioned over time in order to reproduce past results and to have an explanation as to what was done at a given point in time.

If the data is not stored where the compute workload will occur, then there will be data movement issues. Deep learning frameworks and GPU-based workloads do not support the HDFS (Hadoop Distributed File System), because it is not a standard file system. While they do support storage area network/network attached storage, the problem is that the distributed workloads require data to be copied back into HDFS. Deep learning software needs a POSIX (Portable Operating System Interface) file system and the distributed analytics workloads need a file system which supports the HDFS API. Without a system that supports both POSIX and HDFS APIs, data must be copied out of HDFS to a POSIX file system to use with a GPU. Upon completion of the job, the data must be copied back into HDFS to perform distributed analytics.

If this sounds crazy, it's because it is. It also requires a lot of work to manage, since every time data is copied in or out of HDFS, steps such as the application of security controls must be repeated. Versioning this data is critical, and it cannot be accomplished in-place within HDFS. The crux here is to reduce the data storage and management into a single context for both deep learning and all other workloads simultaneously.

The code that is used to create the working models is another point of pain. Languages such as Python are often used, as are notebooks, such as Jupyter or Apache Zeppelin, to create the models. Versioning code is pretty straightforward with great tools such as Git. However, when the generated models and parameters are coupled with the data, there is a complex issue at hand. For this, there is no "easy button." However, using point-in-time consistent snapshots within your distributed storage alongside the GPUs is a great option, as it can handle the data in-place with no data copies required.

After the models are created, they must be performance tested. Data must be run through all the models and analyzed in aggregate to find the best-performing model to take forward into production. Development and testing work are predominantly batch-style workloads. Because of this, new issues arise when preparing for production, such as the handling of real-time workloads, and testing and upgrading models while running. Iterating over and testing changes to existing models require a deployment model such as the rendezvous architecture. This brings with it the concepts of a decoy and canary as ways to test and validate models before going all in for a production environment.

Performing data preparation workloads (distributed processing), GPU workloads, data analysis workloads (distributed processing), and even real-time learning and scoring into a single platform causes the abundance of problems discussed here to no longer be serious issues. It is critical to use a platform that supports both HDFS and POSIX. Now, if this platform also supports event streaming, then double the bonus, because real-time scoring is the end goal of most deep learning applications. This removes the need to move or copy the data multiple times and addresses the data operations problems associated with deep learning.

Jim Scott, director of Enterprise Strategy and Architecture at MapR, is the co-founder of the Chicago Hadoop Users Group (CHUG), where he coordinates the Chicago Hadoop community.

AD INDEX

Melissa ..... Cover 4

BEST PRACTICES

Aerospike 16

MariaDB 17

BIG DATA 50 TRAILBLAZERS

BackOffice Associates... 26

Denodo 26

erwin 27

Franz Inc. 27

Idera 28

RedPoint Global 28

SnapLogic 29

Oracle DBAs Versus SQL Server DBAs

Recently, VMware delivered a unique technology event as a component of an even more unique program. What made this event so unusual was the particular group of technologists attending and the approach VMware has taken to win them over. Since 2013, VMware has offered the VMware Experts Program, to which it invites recognized technology experts in the specific disciplines of Oracle, Microsoft SQL Server, and big data and high-performance computing. The experts become part of this special program to learn and collaborate with other VMware and external experts. The program initiation occurs when the selected experts travel to a location around the world for 3 days of executive briefings, architectural and engineering discussions, customized labs, and social activities all centered on the particular application or database of focus running on VMware technology. The program enables these industry experts to ask the tough questions, and, since they have all agreed to non-disclosure, they are given straight answers and are provided direct access to the technology to try to push it past its limits. This esteemed program has been held in Palo Alto, Calif.; Sofia, Bulgaria; and Cork, Ireland; the latest event was held in Sydney, Australia.

The amazingly unique aspect of the recent event held in Sydney was that both Oracle experts and Microsoft SQL Server experts were in the same room for the majority of the event. In this article, we explore what makes these two groups different and what makes them similar.

Baby Boomers Versus Generation X

The year was 1977, the company was Software Development Laboratories (SDL) and this is where the first version of the Oracle Database began development. In 1979, SDL changed its name to Relational Software, Inc., then in 1982 to Oracle Systems Corp., and later on, to Oracle Corp. In 1979, Oracle Version 2 was released as its first commercial version of the database. Oracle has always shied away from releasing a version 1 of any its products. The only version 1 of an Oracle product we can remember was FastForms that was released with Oracle Database Version 4.

In the early days, DBAs did not exist. Oracle professionals were developers who performed the many roles that we now know as the sacred realm of the DBA. It is possible but also rare to find an Oracle DBA with almost 40 years of experience.

It was nearly 10 years later when, in 1988, Microsoft joined forces with Ashton-Tate and Sybase to create the first Microsoft SQL Server release, a 16-bit version for the IBM OS/2 operating system. This release allowed Microsoft to enter the enterprise database market and to compete against the Oracle Database and IBM DB2. In 1993, about the same time, the new Windows NT operating system was released, and Sybase and Microsoft parted ways.

At the Sydney event, there was a noticeable difference in age between the Oracle DBAs and their Microsoft counterparts. It was interesting to note that the older Microsoft DBAs all began their careers on Oracle Databases.

The Oracle DBAs present were hard-core DBAs while the SQL Server DBAs were mostly developer DBAs or originally Windows system administrators. This was reflective of the complexities of Oracle versus Microsoft. SQL Server is tightly integrated with the operating system, allowing the typical SQL Server DBA to become more intimate with the application development aspect of the organization. The Oracle RDBMS DBAs tends to use a more siloed approach. The production DBA would not also be expected to be a development DBA.

Loyalty to the Product

Both groups were fiercely loyal to their database and its capabilities. Yet the love for the company was very different. The Oracle DBAs loved the Oracle database, but they were not always very happy with Oracle and how they were treated. This should come as no surprise to anyone who has ever done a Google search using the key term "Oracle aggressive."

No matter where the event is held, the topic of Oracle licensing on VMware is always one of the most asked for and animated discussions at the experts event. At this event, VMware invited two firms that specialize in licensing Oracle on a VMware virtualization platform, LicenseFortress and House of Brick. ▶

Michael Corey, co-founder of LicenseFortress, was recognized in 2018 as one of the Top 100 people who influence the cloud. Corey is a Microsoft Data Platform MVP, Oracle Ace, VMware vExpert, and a past president of the IOUG. Check out his blog at michaelcorey.com.

Don Sullivan has been with VMware since 2010 and is the product line marketing manager for Business Critical Applications.

The Microsoft DBAs, in contrast to the Oracle DBAs, loved the database and loved the company. They see Microsoft as a partner and feel comfortable criticizing the company to help make it better.

Operating Systems

Oracle early on adopted the C programming language, and this enabled it to port the database very quickly to different operating systems. As a result, the Oracle DBAs matured in a world of multiple operating systems, and they were able to understand the effect that had on the database tuning and configuration. While the Oracle Database is portable, there were subtle nuances a DBA would have to master. To save an organization development costs, it is common for an Oracle application to be built on one operating system and then ported and deployed in production on another. The skepticism on how the database would perform was clearly reflective in the questions the Oracle DBAs asked during the program. The tenor was reminiscent of a famous statement made in the 1890s by Congressman Willard Vandiver: "I'm from Missouri, and you've got to show me."

The SQL DBAs were raised in a world of one database tightly coupled with a single operating system—Windows. As a result, the database required much less tweaking to get optimal performance out of the database. The processes were well-known and well-understood. When moving from development to production, it was apparent what to expect for performance and behavior. It is very common for a SQL Server DBA to spend as much time being a developer DBA as a production DBA. In the Oracle world, the DBAs are much more specialized.

From a technology perspective, Oracle DBAs were required to develop skills outside the realm of the database. Oracle Parallel Server (Oracle Real Application Clusters RAC) required the experienced DBA to become a network administrator, and Oracle "Automatic Storage Management" required the DBA to develop skills traditionally restricted to the storage administrator. As a result, the Oracle DBA role became pre-eminent in the data center. In the book *Oracle on vSphere*, Oracle DBAs are compared to the Roman Praetorian Guard in that the Oracle DBA, similar to the elite legionnaires who were specially selected to protect the Roman emperors, had influence way above their paygrades.

Out of necessity, Oracle DBAs have become more specialized. Will this happen to SQL Server DBAs now that the database is offered on Linux?

The Communities

Both Oracle and SQL Server have very well-established communities. Yet, while many of the Oracle DBAs were aware of each other's prominence, the SQL DBAs were a much tighter family globally. This is a direct result of the plethora of events run by PASS (Professional Association for SQL Server), including the loosely connected SQLSaturday events that are held every week around the world.

While the two communities are different, they are also similar in many ways. All DBAs worry about the performance and security of the data and the database. Out of necessity, Oracle DBAs have become more specialized. Will this happen to SQL Server DBAs now that the database is offered on Linux? Both are fiercely loyal to their technology stacks, yet each feels very different about their relationship with the vendor. And, after 40 years of change and growth, the old cliché is still valid: The more things change, the more they remain the same.

Data Governance: We Are Programmed to Receive

Summer vacation has ended in the Northern Hemisphere. From camping and hiking to glamping and relaxing, many of us had our fair share of fun again this year. This past summer, I had the opportunity to stay in a wide variety of hospitality establishments for both personal and professional travel, which, for me, has generated another fun way to look at data governance that I am excited to share. Because, let's face it, without good analogies, data governance on its own can be, well, kind of dry.

Welcome to the Hotel Data Governance. Such a lovely place. (The tune is already in your head isn't it? You're welcome.)

1. Reservations recommended, but not required.

It would be nice if we knew everything that was going to be included in our data governance program day to day, but that's not reality.

Similar to hotel regulars, there are some data governance initiatives that are clearly defined and planned. They provide plenty of notice, and they are very specific about what they want in their accommodations, how long they expect to stay, and when they will return. Initiatives that are compliance-driven or regulatory in nature are typical Hotel Data Governance regulars.

But, there are also projects, requests, and needs that view data governance as a pit-stop on a long trip. Heck, these guests probably didn't even know they were stopping by until the last minute. For them, data governance is an afterthought, and they don't make reservations. You want to garner your customers' attention, even if it is at the last minute. Build it into your framework to attract and accept these weary travelers. Think of part of your communication plan as an interstate billboard seen on the road as stomachs start growling and eyelids are getting heavy. They can't afford not to stop. And, as long as they enjoy their stay, they will likely stop again on another trip. They might even make a reservation next time.

The trick is that, similar to a hotel, there is limited vacancy. With existing reservations, you can only accept walk-ins until you reach capacity. But the billboard is still up. You will have to say no sometimes because you are limited by your resources. You can't accept all last-minute projects or solve all the organization's data problems at once. When you reach the point where you can no longer accept walk-ins, it is time to add on to the hotel and expand your data governance program.

2. "We are programmed to receive. You can check out any time you like. But you can never leave!" *(Eagles, "Hotel California," 1977)*

Perhaps an interesting twist and new theory for the meaning of these treasured lyrics, but these are ideal words for data governance. (How's that tune in your head now?)

Intake should be easy for data initiatives. Don't make onboarding for data governance (and related data management) onerous or they will never want to stay. Bring projects into the data governance fold with few constraints. And be ready to push the responsibility for compliance with overarching enterprise governance down as quickly as possible. Data decisions belong at the lowest level of autonomy within an organization. Enable an environment in which your patrons don't have to subscribe to the "bureaucracy" of having others govern them and are able to govern themselves. They check out, but they never leave. Really, the lyrics are brilliant!

3. Housekeeping and/or maintenance?

If a guest has a problem in the Hotel Data Governance, do they call housekeeping or maintenance? It depends. Data governance needs both to keep the light on for you (thanks Tom Bodett).

Housekeeping and maintenance are often used interchangeably by hotel guests. When a problem arises, they don't care who fixes the problem, they just want it fixed. Data governance patrons are no different—they just want solutions (and to complain about the problem, of course). Streamline the process for customers to log their issues, but recognize that housekeeping and maintenance are not the same. While they share outcomes of guest comfort and safety, what housekeeping and maintenance do and how they do it are vastly different.

Housekeeping is a daily activity with a keen focus on cleanliness and hygiene. While cleaning, housekeepers also look for potential safety concerns, ensure a comfortable guest environment, and supply all room amenities. Housekeeping follows a very rigorous process that is repeated throughout the entire establishment. While some rooms may receive different daily treatment at the guest's request, there is a predetermined process for guests to request additional services or decline services each day. Ideally, housekeeping is a finely tuned process with little to no room for ad hoc requests other than when predetermined room amenities need to be replenished. The process and rigor of housekeeping are necessary for effective and efficient data governance.

Maintenance operates differently: Maintenance is not a single daily task or process. Maintenance can happen at regularly scheduled intervals which vary from task to task. While maintenance is also intended to ensure guest safety and comfort, it is more periodic. Maintenance is also available 24/7 to address ad hoc requests related to functional operations (such as light bulbs, remote controls, and toilets) and privacy/safety.

Data decisions belong at the lowest level of autonomy within an organization.

Because the expectations of housekeeping and maintenance are unique, they are typically different roles. Make sure you are able to provide both housekeeping and maintenance within your data governance program.

There is plenty of room at the Hotel Data Governance. Housekeeping and maintenance will ensure your safety and comfort. It will be clean, and it will be functional. Housekeeping will ensure there shouldn't be a problem, but if one should arise, maintenance will make sure it is resolved. Reservations are recommended but not required. You can check out, but you will never leave. Are you livin' it up at the Hotel Data Governance?

Anne Buff is a business solutions manager for SAS Best Practices, a thought leadership organization at SAS Institute. She specializes in the topics of data governance, MDM, data integration, and data monetization.

Good Habits Light the Way to IoT Innovation

In my last article, I discussed how the Internet of Things market is showing early signs of maturity, but that many projects still can stumble. I identified seven "habits" that successful projects have in common, which, when used together, are powerful enough to set your IoT project on the right path.

I described habit 1, lead with use cases, which has proven to be a very effective way of modeling software systems. Let's now consider habits 2 and 3.

Habit 2: Working in Multi-Disciplinary Teams

Remember this joke: How many software engineers does it take to change a light bulb? The answer is: "None, that's a hardware problem."

Well, with the arrival of IoT, this joke can be binned. Technically, IoT stands for Internet of Things—but it really means the merger of operational technology and information technology. (I love the fact that the acronym literally shows the coming together of IT and OT).

IoT projects, for many companies, trigger an epiphany in their understanding of digital transformation. As such, many of the IoT projects they embark upon function as lighthouse projects, illuminating the path to the future for customers and employees alike.

With all the stakeholders in the room, and the high hopes accumulated, these projects are not the easiest to run. But that is not the only reason; IoT projects are hairy beasts for several reasons:

- They tend to cross boundaries, inside as well as outside the organization.
- They tend to touch a lot of different technologies, from hardware (sensors, devices) to telecommunications (networks) and software (security and integration).

A side effect is that the IoT projects have an impact on people and departments across the chain. So, is there a way to handle these challenges?

When you start your projects, make sure that you have the right mix between the different capabilities, as well as IoT platform developers, and make sure you have engineers to tackle the hardware-related problems (device integration). You also need architects to figure out how to design a scalable platform that can integrate into the larger IT landscape and UX (user experience) designers who can build appropriate dashboards and user interfaces. ▶

If the project is to be commercially viable, make sure marketing and sales are involved to come up with pricing schemes and ideas on how to pitch the unique value early in the process—and make sure that those pricing schemes can be measured in the platform. This may sound strange, but as IoT projects will be chartered to disrupt the competitors in the market, there might be whole new business models that need to be supported.

Your company might be thinking of renting your goods out on a per-usage basis, where previously they would have sold at a one-off price. You can imagine that if you are going to charge customers based on real usage that—besides measuring the consumption—there would be all kinds of additional billing processes to create.

Finally, if there is a component of intelligence, such as predictive maintenance, make sure you have a data scientist who can validate that the data that is collected can be analyzed and taken into account appropriately.

Getting all these teams on board and keeping them there is a challenge in itself.

IoT projects, for many companies, trigger an epiphany in their understanding of digital transformation.

Habit 3: Put the Platform at the Heart

As a kid I loved the story of the Three Little Pigs. In case you don't know, it is a fable about three pigs who build three houses of different materials. "The big bad wolf" blows down the first two pigs' houses, made of straw and sticks, respectively, but is unable to destroy the third pig's house, which is made of bricks.

Interestingly, many discussions that I have with prospects about applications versus platforms, in the realm of IoT, remind me of this tale. This is because the platform-versus-application discussion often favors opportunistically short-term quick gains over long-term needs.

Bart Schouw is IoT solutions director at Software AG. Based in the Netherlands, he has nearly 20 years of experience in IT in all areas.

The arguments given in favor of applications are often the same. "If the application is exactly tailored to address this specific need, why build it ourselves?" Or: "Why take the development risk?" Another argument heard often is: "We have been delaying this too long; I need it now—not next year." Even worse: "If IT steps in, they hinder innovation."

I don't necessarily disagree with the arguments (except maybe for the latter). However, an issue arises if there is no strategy within the company to bring all those solutions together and bind them. If the business side buys applications, and then leaves them to IT to organize and support, support costs will explode. And not only that; while parts of the organization benefit from the application, the organization as a whole could suffer security and reliability risks, difficulties in orchestration, an inability to upgrade assets or devices due to lock-in with application vendors, and much more.

While one platform might require the most effort in the beginning, the advantages in the long term may be undeniable for the organization. So, is it possible to overcome some of the obstacles mentioned before and justify starting with an IoT platform?

The first suggestion is to consider a staged approach. Take a cloud-based platform initially, but consider one that can grow to a PaaS (platform as a a service) or beyond if complexity increases. It will help accelerate the development in the beginning, allowing you to reduce risk and get some first findings—if the ideas that you started with are the right ones. This lets you get the confidence and buy-in from management for the more advanced solutions.

And, second, refuse the creation of a business case per solution; to place a tab on the cost and value of such innovation is nearly impossible, as the impact of IoT on the changing business models is often not understood enough. Instead, tie the introduction of an IoT platform to the digital transformation program and assure executive sponsorship.

To conclude, don't forget that any application that you can buy can also be purchased by your competitor. If you do that and build your IoT strategy with straw or sticks, I am sure that your competition will stand on your doorstep on an early morning and try to huff and puff your organization away. Be safe and build your IoT strategy from bricks; build it on top of an IoT platform.

In my next article, I will discuss habit 4, Work Backward, and habit 5, Be Obsessed with Data.

Data Warehouses, Data Marts, Operational Data Stores, and Data Lakes: What's in a Name?

MANY ORGANIZATIONS NOWADAYS are struggling with finding the appropriate data stores for their data. Let's zoom in on some key data structures to facilitate corporate decision making by means of business intelligence. More specifically, let's look at data warehouses, data marts, operational data stores, data lakes, and their differences and similarities.

A data warehouse was first formally defined by Bill Inmon in this way: "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."

Subject-oriented implies that the data is organized around subjects such as customers, products, sales, etc. The data warehouse is integrated in the sense that it integrates data from a variety of operational sources and a variety of formats such as relational database management systems, legacy database management systems, and flat files. Time variant refers to the fact that the data warehouse essentially stores a time series of periodic snapshots. Operational data is always up-to-date and represents the most recent state of the data elements, whereas a data warehouse is not necessarily up-to-date but represents the state at some specific moment(s) in time. Non-volatile implies that the data is primarily read-only and will thus not be frequently updated or deleted over time. Hence, the two most important types of data manipulation operations for a data warehouse are data loading and data retrieval.

A data mart is a scaled-down version of a data warehouse aimed at meeting the information needs of a homogeneous small group of end users such as a department or business unit (marketing, finance, logistics, or human resources). It typically contains some form of aggregated data and is used as the primary source for report generation and analysis by this end user group.

There are various reasons for setting up data marts. First of all, they provide focused content, such as finance, sales, or accounting information, in a format tailored to the user group at hand. They also improve query performance by offloading complex queries, and therefore workloads, from other data sources, such as a data warehouse. Data marts can be located closer to the end users, alleviating heavy network traffic and giving them more control. Finally, certain reporting tools assume predefined data structures which can be provided by a customized data mart. ▶

In order to denote the contrast with a data mart, a full-blown data warehouse is often called an enterprise data warehouse to emphasize the organization-wide aspect.

An operational data store (ODS) is another way of dealing with the disadvantage of data warehouses not containing up-to-date data. An ODS can be considered a staging area that provides query facilities. A normal staging area is only meant for receiving the operational data from the transactional sources for the sake of transforming the data and loading it into the data warehouse. An ODS also offers this functionality, but in addition, it can be queried directly. In this way, analysis tools that need data that is closer to real time can query the ODS data as it is received from the respective source systems, before time-consuming transformation and loading operations. The ODS then only provides access to the current, fine-grained and non-aggregated data, which can be queried in an integrated manner without burdening the transactional systems. However, more complex analyses requiring high-volume historical and/or aggregated data are still conducted on the actual data warehouse.

The data lake concept became known as part of the big data and analytics trend. Although both data warehouses and data lakes are essentially data repositories, there are some clear differences as listed in the table at right. A key distinguishing property of a data lake is that it stores raw data in its native format, which could be structured, unstructured, or semi-structured. This makes data lakes fit for more exotic and "bulk" data types that we generally do not find in data warehouses, such as social media feeds, clickstreams, server logs, and sensor data. A data lake collects data emanating from operational sources "as is," often without knowing upfront which analyses will be performed on it, or even whether the data will ever be involved in analysis at all. For this reason, either no or only very limited transformations (formatting, cleansing) are performed on the data before it enters the data lake. Consequently, when the data is tapped from the data lake to be analyzed, quite a bit of processing will typically be required before it is fit for analysis. The data schema definitions are only determined when the data is read (schema-on-read) instead of when the data is loaded (schema-on-write) as is the case for a data warehouse. Storage costs for data lakes are also relatively low because most of the implementations are open source solutions that can be easily installed on low-cost commodity hardware. Since a data warehouse assumes a predefined structure, it is less agile compared to a data lake, which has no structure. Also, data warehouses have been around for quite some time already, which automatically implies that their security facilities are more mature. Finally, in terms of users, a data warehouse is targeted toward decision makers at the middle and top management level, whereas a data lake requires a data scientist, which is a more specialized profile in terms of data handling and analysis.

	Data Warehouse	Data Lake
Data	Structured	Often unstructured
Processing	Schema-on-write	Schema-on-read
Storage	Expensive	Low cost
Transformation	Before entering the DW	Before analysis
Agility	Low	High
Security	Mature	Maturing
Users	Decision makers	Data Scientists

It is important to understand the differences and similarities between data warehouses, data marts, ODSs, and data lakes. All these data structures clearly serve different purposes and user profiles, and it is necessary to be aware of their differences in order to make the right investment decisions.

This article is based on the recent book *Principles of Database Management—The Practical Guide to Storing, Managing and Analyzing Big and Small Da*ta by Wilfried Lemahieu, Bart Baesens, and Seppe vanden Broucke. See www.pdbmbook.com for more details.

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

BRINGING YOU THE WORLD OF DATA MANAGEMENT

READY TO MAKE DATA WORK FOR YOU?

SUBSCRIBE TODAY!

FOR MORE THAN 20 YEARS, ***Database Trends and Applications*** **magazine has covered the technologies and processes involved in every aspect of the creation, management, application, and storage of structured and unstructured data.** ***DBTA*****'s content is original, factual, and uniquely valuable—providing clarity, perspective, and objectivity in an increasingly complex and exciting world, where data assets hold the key to organizational competitiveness.**

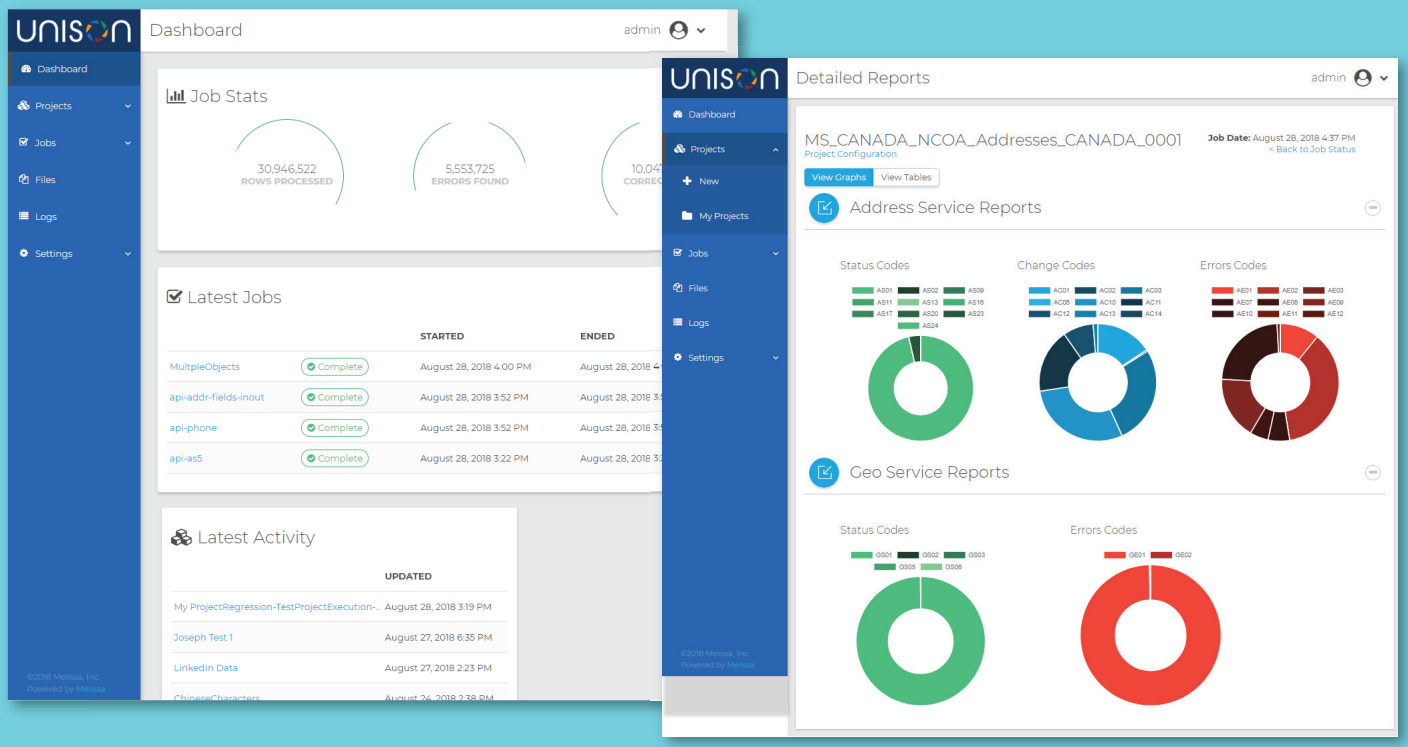
dbta.com/subscribe

MDM – Fast, Easy, Hassle-Free

Unison – end-to-end data quality in one platform with no coding required

Melissa's Unison is a data steward's best friend. It's the ideal centralized solution to establish and maintain contact data quality - without programming! Connect to multiple RDBMS platforms and collaborate on projects. Bring data standardization, validation, and enrichment together for end-to-end data quality. Schedule jobs, then visualize reports with analytics for deep insight. With Unison, you can cleanse sensitive customer information throughout the enterprise – fast, easy, and hassle-free. Data never leaves your organization! Streamline data prep workflows, reduce analytics busy work, gain more insights and increase efficiency.

- **Verify, standardize names, addresses, email addresses and phone numbers, plus append lat/long coordinates and Census data.**
- **Verify securely on-premise with regular automatic FTP updates from Melissa.**
- **Works offline, scalable across multiple servers & allows users to script batch jobs with various levels of data access.**
- **Check and monitor data quality over time with analytic reports and visualizations, plus enjoy project collaboration and more!**



Get Started Now!

Melissa.com/bdq-unison

1-800-MELISSA

melissa®