Volume 33 Number 2
APRIL/MAY 2019

The Journal of
Information Integration and Management

database
TRENDS AND APPLICATIONS

In Pursuit of AGILITY:

How Organizations Are Transforming Their Data Environments

4

WWW.DBTA.COM

Data Quality: The Root of Sound Analysis 16

For Big Data Insights, Start Small 18

Three Important Database Security Features 28

Address the ELEPHANT IN THE ROOM

Bad address data costs you money, customers and insight.

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

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

See the Elephant in Your Business -

Name it and Tame it!

melissa

www.Melissa.com | 1-800-MELISSA

Free Trials, Free Data Quality Audit & Professional Services.

The Journal of Information Integration and Management

database
TRENDS AND APPLICATIONS

Volume 33 | Number 2
APRIL/MAY 2019

CONTENTS

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

DATABASE TRENDS AND APPLICATIONS (ISSN: 1547-9897; USPS: 16230) is published bimonthly (Feb./Mar., Apr./May, Jun./Jul., Aug./Sep., Oct./Nov., and Dec./Jan.) by Unisphere Media, a division of Information Today, Inc., 143 Old Marlton Pike, Medford, NJ 08055 USA; Phone (609) 654-6266; Fax (609) 654-4309; Internet: infotoday.com. Registered in U.S. Patent & Trademark Office. Periodicals postage paid at Vincentown, NJ, and additional mailing offices.

POSTMASTER Send address changes to *Database Trends and Applications*, P.O. Box 3006, Northbrook, IL 60065-3006.

ONLINE ACCESS Visit our website at www.dbta.com

Contents also available online under direct licensing arrangements with EBSCO, NewsBank, ProQuest, and Gale and through redistribution arrangements with information service providers including, Dow Jones Factiva, LexisNexis, OCLC, STN International, and Westlaw.

SUBSCRIPTION INFORMATION
Subscriptions are available free to qualified recipients in the U.S. only. Nonqualified subscribers in the U.S. may purchase a subscription for $79.95 per year. Delivery outside North America is $145 via surface mail per year. All rates to be prepaid in U.S. funds. Subscribe online (circulation@dbta.com) or write Information Today, Inc., 143 Old Marlton Pike, Medford, NJ 08055-8755.

Back issues: $17 per copy, U.S.; $22 per copy, Canada and Mexico; $27 per copy outside North America; prepaid only. Missed issues within the U.S. must be claimed within 45 days of publication date.

Change of Address: Mail requests, including a copy of the current address label from a recent issue and indicating the new address, to *DATABASE TRENDS AND APPLICATIONS*, P.O. Box 3006, Northbrook, IL 60065-3006.

Reprints: For quality reprints of 500 copies or more, call (908) 795-3703 or email reprints@dbta.com.

EDITORIAL OFFICE 121 Chanlon Road, New Providence, NJ 07974

List Rental: American List Council. Contact Michael Auriemma, Account Manager, (914) 524-5238 or email Michael.auriemma@alc.com

FEATURES

RESEARCH@DBTA

2 **DATABASES SHOW GROWING PAINS AS DIGITAL ENTERPRISES EXPAND**
By Joe McKendrick

FEATURE STORY

4 **IN PURSUIT OF AGILITY: HOW ORGANIZATIONS ARE TRANSFORMING THEIR DATA ENVIRONMENTS**
By Joe McKendrick

DEPARTMENTS

TRENDS

16 **DATA QUALITY: THE ROOT OF SOUND ANALYSIS**
By Joe Hellerstein

18 **FOR BIG DATA INSIGHTS, START SMALL**
By Wolf Ruzicka

MULTIVALUE SOLUTIONS

24 **ROCKET SOFTWARE BOOSTS ROOMS TO GO INVENTORY AND BUSINESS**

24 **IT WORLD CANADA CHOOSES ENTRINSIK INFORMER TO MAINTAIN CAMPAIGNS**

25 **BLUEFINITY ADDS A VARIETY OF NEW FEATURES IN LATEST EVOKE UPDATE**

COLUMNS

26 NEXT-GEN DATA MANAGEMENT > BY ROB MANDEVILLE
THE PROBLEM WITH DEFINING DATABASE WORKLOAD

27 EMERGING TECHNOLOGIES > BY GUY HARRISON
SPANNER STRETCHES THE CAP THEOREM

28 DBA CORNER > BY CRAIG S. MULLINS
THREE IMPORTANT DATABASE SECURITY FEATURES

30 IOUG OBSERVATIONS > BY SETH MILLER
TIPS FOR TRANSITIONING TO THE ORACLE CLOUD

32 DATABASE ELABORATIONS > BY TODD SCHRAML
FACT TABLES AND THEIR LYING PRIMARY KEYS

MEDIA PARTNER OF THE FOLLOWING USER GROUPS

IDUG The Worldwide DB2 User Community
International Informix Users Group
IOUG independent oracle users group
isug tech
OAUG oracle applications users group
PASS Professional Association for SQL Server
SHARE
U2UG

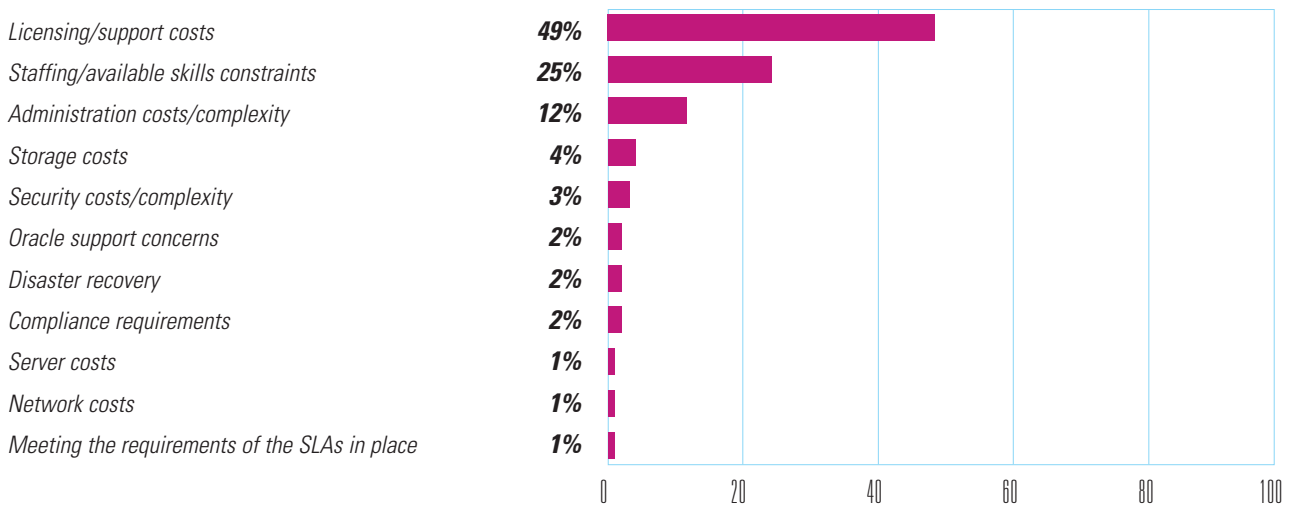
Databases Show Growing Pains as Digital Enterprises EXPAND

By Joe McKendrick

ARE TODAY'S ENVIRONMENTS ready to scale with the business? Data-driven attributes that businesses are relying on for growth in the digital economy—AI, machine leading, and the Internet of Things—require databases that are robust and flexible. However, many enterprises are encumbered by the licensing and support issues that typically accompany database systems, resulting in potentially high and unexpected costs, as well as skills shortages. While enterprises are turning to the cloud and automation solutions to enhance their capabilities in backup and recovery, the challenge is that many data managers subscribing to cloud services are not making licensing costs enough of a priority.

These are some of the key takeaways from the latest in a series of surveys of data managers and professionals, conducted by Unisphere Research, a division of Information Today, Inc., in partnership with VMware. This survey, which covered 260 members of the Independent Oracle Users Group, explored the infrastructure concerns and best practices shaping data management in today's fast-evolving business environments ("2018 IOUG Data Environment Expansion Survey").

Figure 1: What is your top challenge when seeking to expand the number of Oracle databases or applications?

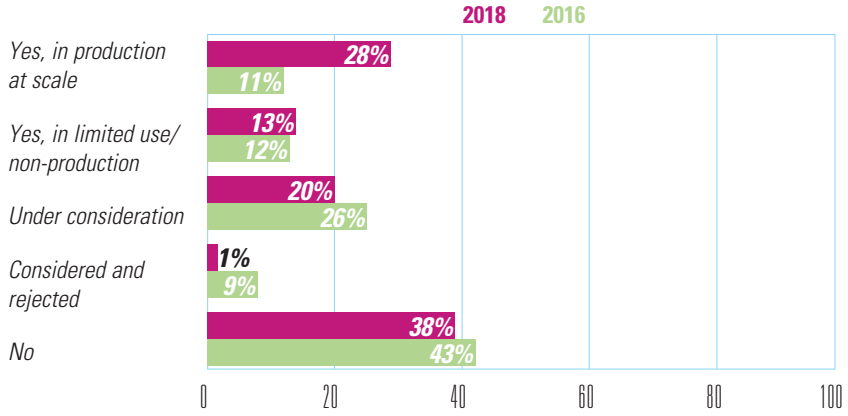


The survey found that licensing and support is the number-one challenge for organizations seeking to expand the number of Oracle databases and applications (see Figure 1). Additional issues also stand in the way of growth. When asked for the top factors holding back data environment expansion, licensing and support were far and away the most pressing challenges cited. As businesses increasingly rely on data to provide critical analytics as well as advance in AI and machine learning, demand for added database and testing capacity and related infrastructure requirements will only grow. In addition, the number of end users will continue to rapidly grow.

Public cloud services are increasingly seen as a way to address the need for greater capacity and scaling to new enterprise requirements (see Figure 2). According to the survey, the use of public cloud services at Oracle sites is growing. There has been a noticeable rise in cloud computing adoption among database teams, the survey found. Forty-one percent reported having cloud in production at scale or in limited use, up from the 2016 survey. Notably, 28% have cloud in production at scale, well over double the level seen 2 years ago (11%). One-third indicated their use of cloud is growing, with 20% reporting their cloud growth as "significant"—again, a rise over just 2 years ago.

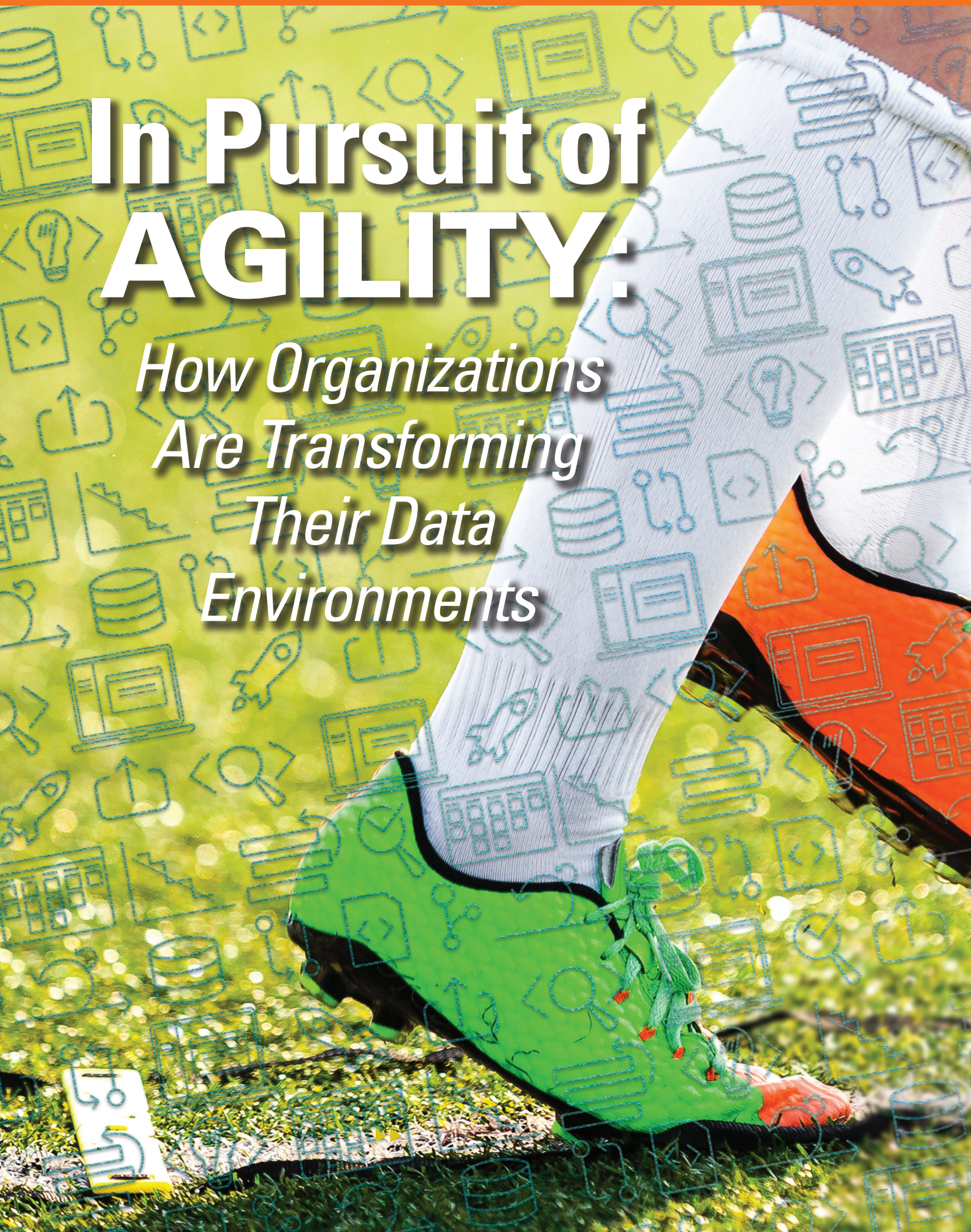
There has been a marked decrease in the percentage of data managers and professionals reporting they "considered and rejected" a cloud-based approach, from 9% 2 years ago to 1%. This points to the increased confidence in cloud across the enterprises, as well as among members of the data management community.

Figure 2: Does your organization use any public cloud services for Oracle databases and applications?



Public cloud growth is driven by backup and recovery supporting transaction environments. A total of 23% respondents delegate a significant share (one-quarter or more) of their backup and recovery processes for transactional environments to the public cloud. For additional processes affecting data environments, close to one in five relies on cloud for a considerable share of their business continuity, monitoring, and provisioning processes. For analytical data environments, there is less commitment to public cloud at this time—at most, 17% reported that they are dedicating a meaningful share of their backup and recovery process workloads to public cloud environments.

Cost reduction is the main benefit anticipated with cloud, but agility and capacity are more likely to be realized in existing deployments. Data managers and professionals seek the cost advantages of public cloud—which may form the basis of business cases, at least initially. Six in 10 foresee the cost reductions as the main benefit sought with cloud computing. As deployments mature, however, the additional agility and on-demand resources that cloud brings are also seen as leading benefits. In addition, the benefits of public cloud computing are far more apparent now than 2 years ago. When asked about benefits already seen, three in four respondents revealed they have experienced greater agility, and a majority of respondents cited the on-demand capacity clouds offer. ■



In Pursuit of AGILITY:

How Organizations Are Transforming Their Data Environments

By Joe McKendrick

THERE ARE MANY technology trends sweeping the business data space, but the single goal of all of them is to achieve a more agile operation and organization. Enterprise agility isn't a single initiative but rather a collection of activities and technologies that lead toward that goal. This includes adoption of microservices, containers, and Kubernetes to increase the flexibility of systems, applications, and data by releasing them from underlying hardware. In addition, practices such as DevOps are helping to increase the level of collaboration possible for fast-moving enterprises.

Simply put, it's all about the business—pursuing greater agility is important in light of the need to transform the enterprise data space. "Regardless of size, organizations need to be nimble enough to compete with established players as well as born-in-the cloud organizations to meet a new level of business demands," said Gerald Venzl, senior principal product manager for Oracle.

*Pursuing **greater agility** is important in light of the need to transform the enterprise **data space**.*

Does IT help agility or stand in the way? All too often, it's the latter. "Traditional IT systems were not built with agility necessarily in mind," said Pete Brey, director of product marketing for Red Hat Data Analytics Infrastructure. As a result, these systems tended to be cumbersome to use and update, he added.

Traditional databases "were built to support traditional monolithic applications residing in data centers, not distributed applications that would need and use data in the cloud—let alone out at the edge of the network in the cases of mobile and IoT," said Lewis Carr, senior director of product marketing at Actian. "In real time, the vast majority of these applications manipulated relatively little data compared to today's applications that rely on data to automate processes, derive insights, and deliver decision support."

RETHINKING ORG CHARTS

Still, IT and data professionals are at the forefront, addressing the challenge of getting data technologies and practices out of the way and bringing forth greater agility. Brian Jones, infrastructure cloud architect with the secure DevOps team at Liberty Mutual, is a believer in the DevOps movement, which aims to bring development and operations activity into sync and allow for rapid delivery of solutions. "The cloud has given startups the ability to disrupt markets that were once untouchable," he said. "This is one of the main factors leading enterprises to increase their agility. The implementation of DevOps practices allows enterprises to do this." However, such a transformation takes time, he cautioned. "DevOps doesn't just happen because one team decides it's the right thing to do. It requires a cultural change at all levels of your IT organization."

DevOps "is a response to IT's inability to meet the increasing demands of the business," said Mark Levy, director of strategy, software delivery at Micro Focus. "DevOps practices, such as continuous delivery, optimize and automate software deployments to ensure that software delivery is not a constraint in delivering business value. Database teams, which rely most on manual processes, are struggling to keep up with the flood of requests they receive to create, review, and deploy database changes in support of continuous delivery. This technology gap creates a constraint in the continuous delivery process and is a prime candidate for optimization and automation."

DevOps "isn't just for startups anymore, as more and more enterprise companies have realized they require agility to stay current and modern," said Ben Thrift, director of engineering with Dell Boomi. "New projects are being built with DevOps in mind, and tooling and automation are making it easier to retrofit legacy projects with DevOps principles."

Furthermore, DevOps is about more than speed. "It's about using speed to make the applications more intelligent and deliver optimal functionality that generates improved business outcomes," Carr pointed out. "DevOps is supposed to help agile teams 'fail fast,' thereby enabling them to perform rapid or even simultaneous A/B testing, and so forth. Implicit in this is the use of far more data and analytics."

RAPID DEPLOYMENTS

Other initiatives moving organizations toward greater agility include deployments of microservices and containers, which enable applications, services, code, and even data to be moved or reused in the most expedient technology environments, such as cloud platforms. However, there is also still much work to be done in these areas. Most enterprises have made some transition to DevOps, containerization, and microservices, but long-established corporations are struggling to turn around their processes and methods, said Venzl. "New projects tend to adopt agile practices faster, benefiting from the clean-slate state over already rolled out projects running in production."

This all means that "it is a brave new world that is unfamiliar to many, so there are more questions than answers today," said Brey. "While progress is definitely being made, especially in markets where the level of competition is high, the access to agile techniques is ready, and there are lower barriers to change."

Julie Furt, senior vice president, worldwide consulting and training services at MarkLogic, sees "pockets" of progress with these next-generation technologies, but often it's simply lip service. "Some organizations pretend to be agile, while really just putting a wrapper on their traditional approach—or worse, have interpreted 'agile' to mean that you can figure out what you're building at the end, not establishing clear goals and then iterating toward them."

Todd Moore, VP of open technology and developer advocacy at IBM, has seen such a radical transition at his own orga-

nization. "Front-ending hundreds of gigabytes of data that lives both in the cloud and enterprise systems with APIs just goes with the territory these days," he said. "Microservices can automatically render that data into content and interactive services and this is exactly the case for one of IBM's most important and visible websites, IBM Developer."

There is also a risk seen with adoption of these new technologies and practices—an assumption that they alone will set organizations on the right track. "A common misconception I have seen with people who are new to Kubernetes, is that the Kubernetes orchestration layer and its self-healing properties will solve all of your problems," Frank Reno, senior technical product manager at Sumo Logic, said. "While there are many problems and headaches it does solve, you cannot abandon your core monitoring principles. It is still critical to monitor the orchestration layer, in addition to all your applications, to ensure the orchestrator is orchestrating. You need a robust, scalable monitoring solution that can scale with the volume of data that comes from running applications inside of Kubernetes."

IMPACT ON DATA ENVIRONMENTS

The database space in particular has been a challenge in moving to more agile approaches, due to greater centralization of "both system and application DBAs," said Jones. "Any type of request to the DBAs could take days, weeks, or even months. In order to improve agility, the developers need to be in control of their databases."

To fully embrace an approach such as DevOps, Jones continued, "talented DBA teams need to become passionate about helping developers take control of what the DBAs used to do. The system DBAs are now helping to provide automation to the development community. They do it in a way that allows developers to securely deploy databases and connect their applications to them in less than an hour. This is something that in large organizations used to take weeks or months. The application DBAs are now either on development teams or working with them to modernize their applications. Part of this modernization is allowing teams to use schema management libraries in their code to deploy schema changes. They are also teaching teams how to effectively and efficiently monitor and maintain their databases."

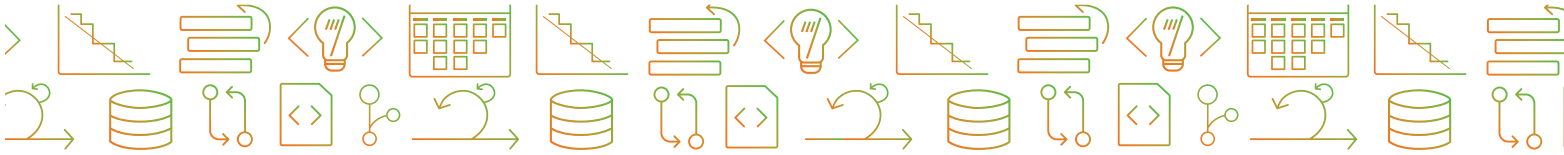
DevOps and agile practices can also butt up against established security mechanisms. DevOps culture is "the heart of agility and emphasizes frequent releases, highly automated—and often remote—build processes, constant configuration, and distributed teams," said Markku Rossi, CTO of SSH.com. "In DevOps environments, the development teams are directly in contact with production systems. There are daily releases of code directly into the live production servers, and the traditional boundaries between development, test, and production environments have eroded." The challenge is to provide "the ability to secure, monitor, control, and audit the connectivity into sensitive data and systems; something that has been in place and optimized in a traditional software environment for years," Rossi continued. Traditional security systems are not suited for today's accelerated and rapid pace of software delivery, he added.

What happens, Rossi related, is "developers and database admins typically game the system, bypassing the enterprise privileged access management by throwing in their own access keys unmonitored and unmanaged. Legacy privileged access management is also an agile administrative and operational roadblock, and source of friction. Managing the lifecycle of hundreds or thousands of access identities over time creates unmanageable complexity, which increases the risk of misconfigured and unmanaged, untracked access. It also means admins, developers, and expensive consultants are either constantly managing or waiting for access instead of doing productive work."

Another challenge has been the nature of applications found in the data center. To deal with massive scale, the alternative microservices approach emerged, but it quickly became apparent there was a technology gap in supporting communications for cloud-native applications, said Arsalan Farooq, chief of strategy for Netifi. "For example, people wanting to scale while retaining usage of relational databases were cut off from reactive programming due to existing standards. These challenges require new technologies, such as R2DBC, an API that allows asynchronous, non-blocking code that works efficiently with relational databases."

Microservices, while showing great promise in helping to break down and replace monolithic systems and applications, also still require caution. The microservices approach has almost a 9-year history and has many success stories and solid practices, but there's still a lot of hype around microservices and too many projects which fail, said Jeff Fried, director of product management at InterSystems. "I've experienced lots of operational and security challenges in stitching together microservices across multiple cloud services—applications which could be built quickly but were very challenging to run. I've also seen too many cases of the abstraction obscuring the reality: data-intensive applications where a microservices implementation resulted in huge, unnecessary data transfers or data duplication."

Data managers can also make greater strides toward agility by elevating ▶▶



their roles and visibility within organizations. "By looking at more ways to automate across the stack, data managers can spend less time on maintenance and instead help their organizations become more agile through new insights from data instead of just keeping the lights on," said Venzl. Venzl pointed to autonomous database technologies as an example of a way to "help reduce the time needed to monitor, repair, and scale critical business assets to eliminate the IT complexity that can slow the roll out of new applications."

Furt urges that "data managers advocate for the ability to dynamically harmonize their data into a consistent 360-degree view that can be flexible enough to respond to changes in data and requirements. They should strive to integrate all the silos of data into this consolidated view."

Additionally, it is important that data managers not build walls of accountability in their delivery pipeline, Thrift emphasized. "Teams must be equipped with the skills and resources to develop, test, deploy, and host their solution without passing accountability onto another group."

Involvement across the enterprise is critical. "Building a community around DevOps and databases has been a huge help in advancing the DevOps culture in our organization," said Jones. "The developers and DBAs are able to share their experiences. We're seeing that the more experienced teams are excited to share what they've learned. Again, I believe DevOps is a culture, not just a group of concepts. In a large enterprise, it cannot be accomplished by just one development team. It requires changes from all levels of the information technology organization."

PREPARING FOR THE AGILE FUTURE

The transformation to agile infrastructure and organizations doesn't happen overnight—it requires not only technology but also training and development. Agile processes and methodology training are essential to managers and professionals, said Furt.

"The adjustment to agile processes is a paradigm shift for most, and training is just the first step on that journey of change," she said. To adopt the new technology optimally, people need to fundamentally change how they think about data, architectures, and the organization of teams.

Brey compares such preparation and training to "team building skills, but on a whole different level because agility will require cross-team collaboration and even a certain amount of blurring the lines of traditional organizational structures." To create differentiated value, organizations will need to focus on the cultural aspects of training, he noted.

Initiatives moving organizations toward greater agility include deployments of microservices and containers, which enable applications, services, code, and even data to be moved or reused in the most expedient technology environments.

In addition, developers and DBAs will need to learn more of each other's roles. "Developers will need to see data management as something they can directly own and embed into their applications," said Carr. "The training isn't just the nuts and bolts of APIs—it's also facilitating a cultural shift as the developer becomes decoupled from substantial data management and a transition is made to a mindset that the developer is their own data manager, data architect, and DBA."

The challenge of applying DevOps, containerization, and microservices to databases "is that data is much more fluid than code," said Fried. "However, data managers can put themselves ahead by applying these agility practices to both their software applications and then ultimately their data. For instance, containerization enables organizations to package an application in a portable container to speed up and simplify the deployment and configuration process. Code and data are stored separately, allowing customers to package their code cleanly in the container. When making application updates, new code can be up and running in the cloud very quickly, and data is then naturally used with the new version."

It's important "to be aware that you are applying a new mindset and simply have patience," Fried continued. "DevOps practices are based on automation, meaning that you design processes to be replicable. It takes more time and resources to build the system initially, and it can be very frustrating as you need to go into the user interface and specify all the code up front. Once established, the ability to react quickly to incoming data, and the resulting insights pulled from it, increases exponentially. It's also important to cut through misconceptions and hype by being persistent and asking a lot of questions." ■

Actian
PAGE 12
OPERATIONAL ANALYTICS AT THE SPEED OF BUSINESS

Denodo
PAGE 13
THE LOGICAL DATA WAREHOUSE AS THE NEW STANDARD FOR DATA ANALYTICS

SlamData
PAGE 14
THE MISSING LINK IN ETL

Wipro
PAGE 15
NEXT-GEN DATA WAREHOUSES TO POWER INTELLIGENT ENTERPRISES

THE FUTURE OF DATA WAREHOUSING

Best Practices Series

database
TRENDS AND APPLICATIONS

RETHINKING THE FUTURE OF DATA WAREHOUSING

Best Practices Series

Does the data warehouse still have a place in today's fast-moving, real-time digital enterprise? Many pundits, analysts, and vendors have proclaimed the impending demise of data warehousing, suggesting that it has become too slow, isolated, and cumbersome to deliver insights at the push of a button. However, the data warehouse has proven the doomsayers wrong, with evidence that it is evolving into an integral and essential piece of the big data landscape.

For decades, companies have invested millions of dollars designing, implementing, and updating enterprise data warehouses as the foundation of their business intelligence systems. And they are still just about everywhere. Most larger enterprises still maintain data warehouses, and small to medium-businesses are also finding data warehouses a cost-effective option, thanks to the cloud.

Now, data warehouses are poised to play a leading role in next-generation initiatives, from AI to machine learning to the Internet of Things. While data warehouses do not appear frequently in marketing literature or analyst reports on these emerging technologies, data warehousing will remain a critical cornerstone of the foundation of the digital era ahead.

"If you can't build a data warehouse, you shouldn't do AI," Andrew Ng, noted computer scientist and co-creator of Google Brain, has said. This requirement runs deep through every digital engagement.

"Every single company I've worked with and talked to has the same problem without a single exception so far—poor data quality, especially tracking data," according to Ruslan Belkin, vice president of engineering for Salesforce.

For further proof of the continuing importance of data warehousing, look to skills demand in the IT workforce. Tellingly, data warehouse engineer has ranked sixth of the top 10 jobs in demand, a recent analysis by Indeed found.

There has been pressure on today's legacy data warehouses to evolve—both architecturally and technologically—to deliver

the agility, scalability, and flexibility that business need to thrive in today's data-driven economy. Alongside new architectural approaches, a variety of technologies have emerged as key ingredients of modern data warehousing, from data virtualization and cloud services to Hadoop and Spark and machine learning and automation.

Here's the shape of the future of data warehousing:

Data warehousing is going to be cloud-based. What was unimaginable just a decade ago is no longer the working reality today—enterprises are turning to cloud to power and store their data warehouses. It will be versatile, providing both real-time and historical insight. The data warehouse will work in unison with other components of the environment. Information from data warehouses will increasingly be the source of insights for both real-time and analytical actions to provide customer service at the time it's needed, while also serving as a repository for historical data. There has been rapid growth and excitement in recent months and years in cloud data warehouses hosted by leading internet companies such as Google and Amazon, which is essentially putting a stamp of approval on the concept of data warehouses in the cloud. In addition, traditional cloud providers also offer their capabilities as a cloud service, along with their traditional on-premise products.

Data warehousing is being extended into modern analytics ecosystems through the use of data virtualization. By federating multiple data warehouses, data virtualization can augment traditional ETL and data replication processes by acting as a virtual data source while also isolating applications from the complexity of disparate and changing underlying data sources.

Data warehousing is going to be analytical. The data warehouse world has blended with the analytics world to the point where they are one and the same. Data warehouses, for all intents and purposes, are data analytics platforms. Companies recognize that data analytical power is crucial to every aspect of their operations and products, and data warehouse technology is already delivering this power.

Data warehouses are poised to play a leading role in next-generation initiatives, from AI to machine learning to the Internet of Things.

Data warehousing is going to empower users like never before. The key advantage to data warehouse environments is the emphasis on self-service. Business end users have long had the capability to build queries or ask questions of their data that had never been asked before, due to the limitations of data silos. Data environments are only growing more diverse and complex, and budgets for IT staffing are getting tighter. The platform data warehouses provide for building queries is proving invaluable at a time when decision makers can't afford to wait on their IT or data management departments for answers.

Data warehousing is going to feed into data lakes, Hadoop, and Spark—as well as the other way around. There has been a great deal of discussion about the future of data warehouses in a world increasingly served by data lakes and about how traditional that extract, transform, and load environments are encumbrances when data needs to tapped on-the-fly for any and all applications.

Data warehousing is going to require fewer people to populate and operate. As with many other elements of the data environment, data warehouses have increasingly become autonomous. These environments were originally designed to be run with as little DBA time as possible.

Data warehousing is going to support AI and machine learning to deliver results. Not only will data warehouses be the foundation of datasets for AI, but AI will also enhance the operations and capabilities of data warehouses. For example, Google has incorporated machine learning into its BigQuery data warehouse.

Data warehousing is still going to occupy a central place in delivering the customer experience. The heritage of the data warehouse is built on understanding the customer in new and profound ways. No other environment maintains data that is so vital to CX. Data warehouses have long been the established repositories for not only historical customer data and demographics, but also can be blended with real-time data streams to provide on-the-spot services and responses to customers.

The data warehouse—as a system, as a concept, and as a way to delivery insights about customers, markets, and operations—isn't going away anytime soon. Data warehouses are increasingly becoming an even more critical part of the digital world. ■

—Joe McKendrick

Operational Analytics at the Speed of Business

Accelerating analytics to operate in-the-moment. From strategic decision-making to low-level operations and customer experience, your entire company must have up-to-date information and insights to keep pace with the speed of business. It isn't okay for your business to be waiting on daily batch updates.

LEADERS NEED REAL-TIME INSIGHTS TO MAKE INFORMED DECISIONS

Technology innovations, customer preference, global economics and market changes are causing the environments in which companies operate to change quickly and dramatically. Business agility is a necessity to survive and thrive in modern commerce. Market opportunities are short-lived, and threats are more impactful than ever. For leaders to be effective in recognizing changes in the environment and make informed decisions that lead to favorable outcomes, they need not only complete and accurate data, but also current data, so they can respond to changes in the moment.

MANAGEMENT NEEDS REAL-TIME INSIGHTS TO ACHIEVE PRODUCTIVITY, PROFITABILITY AND QUALITY GOALS

Sales, customer service, HR, finance, manufacturing and logistics—almost every business process in modern companies is technology-enabled. This can be good if the systems and people involved in operations are working smoothly together and everything is going well. Managers depend on data-driven insights about these business processes to understand operational performance, process quality and cost drivers, enabling them to see where problems exist that require attention.

EMPLOYEES NEED REAL-TIME INSIGHTS TO DO THEIR JOBS EFFECTIVELY

Modern businesses are complex, with operations spread across teams, IT systems and often geographic locations. For employees to be effective in their individual roles, they must understand what is occurring in the other parts of the company with which they interact. Manufacturing employees and planners need visibility of the sales-and-order-management pipeline. Sales teams need visibility to delivery schedules and logistics. Customer-service agents need visibility of customers' orders. To manage this complexity and make informed, tactical decisions, these employees need accurate and real-time data insights.

CUSTOMERS EXPECT REAL-TIME INSIGHTS AS A PART OF THE MODERN CUSTOMER EXPERIENCE

Employees and company leaders aren't the only people who have a need for real-time data insights. Modern customer experiences are highly automated, and customers expect the data they view on the company's Website to be current. Product availability, order status, shipping data and returns processing are where real-time operational data drive digital customer experiences. If there is a change, then customers expect to see the change reflected immediately—they have little tolerance for waiting until the next day for data to be refreshed.

Businesses evolve quickly, in big strategic ways and in small tactical ways. Real-time data and information insights are what enable all parts of your business to identify, understand and respond to changes quickly and decisively. Actian Avalanche—Cloud Data Warehouse Service that enables you to collect and harvest data insights in near real-time and at enterprise scale. This can help you accelerate your business-process execution, monitor and better respond to opportunities and threats and provide employees and customers with the data they need to be informed and effective. ■

ACTIAN
www.actian.com

The Logical Data Warehouse as the New Standard for Data Analytics

Data warehouses are a great tool to consolidate data from a variety of operational systems to become the reference for corporate reporting. They are specifically shaped for analytics and run on specialized hardware.

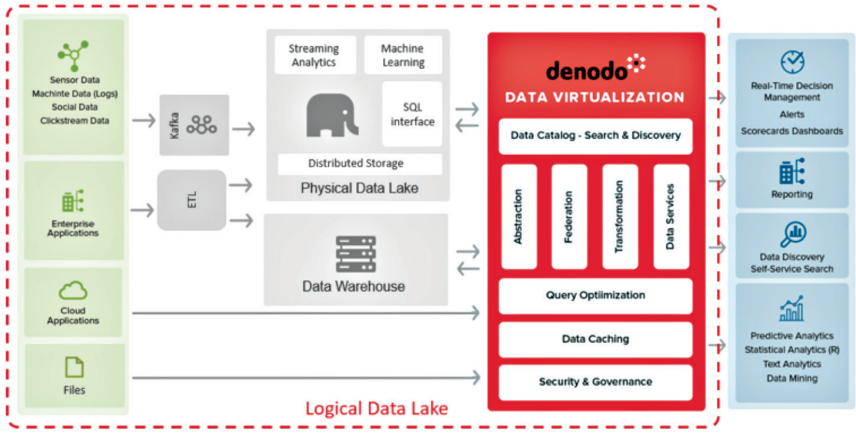
However, especially in the last few years, some of its core principles have been challenged:

- The rise of data driven decision making required storage of vast amounts of raw data. Traditional EDW appliances, with an elevated cost per stored byte, were too expensive. Cheaper distributed storage solutions (HDFS, S3, etc.) took the lead.
- The star/snowflake schema of an EDW is not the best way to store data for certain problems. Key-Value pair, graphs, and other NoSQL systems are designed to address specific challenges.
- Cloud vendors dominate the market. Specialized Software as a Service applications are the reference in many sectors and cloud mega-vendors are driving infrastructure to the cloud.

Although these factors provided huge advantages they also broke the premises of the data warehouse. The data landscape is fragmented, not just in location, but in shape and processing paradigms.

Physical re-consolidation, although possible, is less attractive than before. Volumes are too high, and replication to multiple systems creates brittle point-to-point connections. Out-of-synch data and uncontrolled replication leads to "data swamp" scenarios. End users pay the cost of a fragmented landscape in the form of extended time to market (or, more accurately, "time to data").

Thus, it seems that a logical approach is more feasible: a logical layer that connects different systems and exposes them as one. The complexity of the back-end systems is hidden from the end user. Security, governance and auditing are again centralized.



Data virtualization software like Denodo follows the ideas of relational databases. It provides a metadata catalog and an execution engine. It allows for the definition of derived views and data models. But unlike a database, it does not provide storage. Instead, connections to different systems will feed the data models in execution time. A virtual layer is focused on data delivery, not on storage.

How does execution work in a system like this? Underlying databases usually provide an execution engine, therefore, the virtualization engine takes advantage of them . This is called query push-down. It serves a double purpose: reduces processing in the virtual layer and network traffic. If all the data required for a query is in a single system, the virtual layer does the SQL dialect conversions and completely delegates the query to the source.

However, when data comes from multiple sources, the optimizer needs to come up with a multi-source execution plan. The plan is split into multiple branches that bring partial results from each source, and combines and aggregates them together in the virtual layer. Optimization techniques, although similar to those in relational engines, have evolved differently to deal with the nature of this problem. Techniques like complex query rewriting, on-the-fly data movement between sources, and MPP capabilities provide the processing muscle to perform efficiently.

The value propositions for these logical architectures is simple:

- There is one place to get data. Data exploration and "time to data" are greatly simplified
- Replication needs are significantly reduced, which reduces HW and operation costs
- Data governance is improved. Data is logically consolidated, traced to the source, and secured

As you can imagine, the benefits of a logical data layer go beyond warehousing and reporting, and can be applied to other scenarios like Logical Data Lakes to feed data scientists. ■

DENODO
www.denodo.com

SLAMDATA

REFORM

The Missing Link in ETL

ETL (EXTRACT, TRANSFORM, LOAD) has been around for decades. Its primary purpose is moving data from source locations to data warehouses so analytics and data science teams can perform analysis across a range of critical data sources with standard tools. More recently, with the rise of low-cost cloud object storage, like AWS S3, Azure Blob storage, and others, this process has morphed into ELT (extract, load, transform). In this process, the data transformations are pushed further down the pipeline which somewhat streamlines the problems and also lowers overall costs. ETL/ELT tools have flourished in the last decade as the volume and variety of data sources that enterprises need to handle has exploded. However, there is one obvious gap in the solution space, complex JSON data, which coincidentally is also one of the most popular and rapidly growing kinds of data we see in the market. Unlike traditional relational or tabular data, JSON does not have a one-size-fits-all data model. In fact, it can range from very simple to unbelievably complex depending on the whims of the developers building the applications that create the JSON data. When existing ETL/ELT vendors say that they support JSON they mean they support VERY simple flat JSON. As soon as complexity goes up, they go down, and revert to the familiar approach: Start writing custom code! Some vendors don't even try to avoid code; they actually build a coding engine in their platform to handle complex JSON. So, the harsh reality is that complex JSON is the last ETL problem to be solved.

BUT IS JSON DATA DIFFERENT?

JSON data is some of the most common data created today. Virtually all SaaS applications, Mobile applications, and IoT have JSON as the default data model. And the majority of Web APIs provide a JSON payload. JSON data is unlike traditional relational data in many ways, including non-fixed variable schema, variable data types, and the ability to have "nested" data structures. This presents a major hurdle when companies need to access this data for analytics purposes.

Analytics tools expect the data to be in a fixed tabular form (think spreadsheet) of rows and columns. In order to do this, the data needs to be transformed from the JSON model to the tabular model. Traditional ETL/ELT solutions cannot handle complex JSON well, if at all. To be clear, they all claim that they handle JSON, but what they really mean is that they can do some very simple things, and then require engineers to write complicated code to solve the rest. Most companies don't even bother trying to use commercial ETL/ELT software and simply have highly paid data integration engineers write custom code to transform their JSON data. This approach is slow, complicated, expensive, and not self-service in any way.

FINALLY, A SOLUTION FOR TRANSFORMING JSON

SlamData REFORM is a revolutionary solution lets ANY user visually prepare analytics-ready tables directly on the JSON data, regardless of complexity. This means ZERO CODING and no waiting on Data Integration Engineers. Users can curate out custom data sets in minutes, and then iterate over them at any time as their data needs change. These tables can be streamed into all popular data warehouses, including Redshift, Snowflake, and Teradata, or pushed to any other destination you choose.

- Zero coding solution
- Any users, not just engineers can make complex JSON data analytics-ready
- Any data (JSON,CSV,XML), regardless of complexity
- More agile and accurate than custom coding
- Fast high-performance streaming engine for large amounts of data
- Adjusts automatically to changes in data

WHO CAN USE SLAMDATA REFORM?

Data Integration Engineers—Makes their job easier, less coding, ability to respond to users' needs faster

Data Architects—Makes their job easier, less coding, ability to respond to users' needs faster

Business Analysts—Lets them have REAL self-service against complex JSON data (nobody else can really say this)

Data Scientists—Lets them have REAL self-service against complex JSON data (nobody else can really say this)

FAST AND EASY TO INSTALL

SlamData REFORM is a software tool (so no added SaaS compliance or security issues) and is also available in the AWS Marketplace. Users can install the solution within their existing infrastructure and use it as they need, securely.

REFORM supports JSON data stored in AWS S3, Azure Blob Storage, Wasabi, and MongoDB. We can add a new connector to any JSON data source quickly with our advanced Lightweight Connector Technology (LWC).

Learn more about SlamData REFORM at http://slamdata.com or see it in action in this informative video. ■

SLAMDATA
https://slamdata.com/

Next-Gen Data Warehouses to Power Intelligent Enterprises

In today's digital era, consumers around the world are driving organizations to transform themselves into intelligent enterprises by embracing technological innovation in artificial intelligence (AI), cloud, and Internet of Things (IoT). These innovations can radically impact businesses with adoption of right strategy to harness the power of data and analytics to aid digital transformation. The need of the hour is to move from a "system of records" to "actionable insights" through successful delivery of intelligent data platforms that can aid real-time analytics, providing the right data, on demand. The foundation of a successful, intelligent enterprise will be next-generation data warehouse platforms, which can enable any kind of data provisioning in a digitally disrupted world.

Traditional data warehouses served the need of descriptive analytics on core transactional systems capturing only 20-25% of all enterprise data. These warehouses cannot keep pace with business disruption and are a big impediment to agile business analytics and digital computing.

Some fundamental limitations to the traditional data warehouses include:

1. Increased operational risk and threat of data breach
2. Lack of scalability, affecting business agility and time-to-market
3. Increased latency issues as data volumes grow with complexity
4. Lack of accuracy in ROI quantification
5. Tightly coupled platform and integration affecting agility
6. Provisioning for structured data only

Today, data processing has become more evolved and complex with mobile, social media, cloud, machine, and sensor data integration. These new data sources have tremendous business value to be unearthed and monetized. Business need has evolved from descriptive/diagnostic to predictive/prescriptive analysis. This change in analysis is possible only when data is captured in its most native form through streaming, in near real-time, and merged with historical data amounting to massive volumes of data in terabyte/petabytes. Such volumes facilitate in-depth analysis and computing on a large scale to build various forecasting models, empowering businesses with actionable insight. Harvard Business Review Analytic Services recently published a report on the advantages real-time data and analytics can bring to an enterprise, helping to build a truly data-driven intelligent enterprise.

Next-generation data warehouses are on-demand, secure, and scalable self-service data centers that fully automate the provisioning, administration, tuning, backup, and recovery of data. This accelerates analytics and actionable insights while minimizing administration requirements. Next-generation data warehouses also provide real-time, complete access from surface-level analytics components to the core in-memory platform. This allows businesses to ingest and store structured and unstructured data, and also transform raw data assets. A complete portfolio of data exploration, reporting, analytics, machine learning, and visualization tools can be enabled on the data for accelerated analytics without replicating data. With next-generation data warehouses, organizations do not need an innovation-limiting, pre-defined schema that limits their ability to harness insights from available information.

THE ADVANTAGES OF NEXT-GENERATION DATA WAREHOUSES

Cloud is the cornerstone for next-generation data warehouses, given the advantages in cost, scalability, performance, anytime/anywhere access, security, and ease of administration. Many enterprises have started their data-to-decision transformational journey enabled by hybrid, public, and private clouds. With the advantage of hybrid and cloud-native platforms, next-generation data warehouses are becoming smarter in all three dimensions—storage, computing infrastructure, and services. Additionally, built-in resiliency, enterprise-grade security, and protected data-sharing capabilities are making them intelligent enough to empower users for generating insights in a self-service consumption model. With the advent of AWS, MS Azure, and Google Cloud, immense business benefits can be realized that include:

- Creation of a data-driven customer journey, resulting in increased customer satisfaction
- Enhanced business agility and faster time-to-market, enabling improved and faster decision making
- Reduced infrastructure, maintenance, and admin overhead costs, resulting in improved ROI
- Anytime/anywhere access, enabling self-service BI capabilities
- Automation based on AI/ML

With the tremendous growth that analysts are predicting in analytical database management over the next three years, the next-generation data warehouse market will be shaped by the following forces:

- The emergence of data warehouses in the cloud or data warehousing-as-a-service (DWaaS)
- The need for data warehouse infrastructure to support big data
- Increasing demands for low latency and high-speed analytics
- The increased role of business intelligence in enterprise management
- The commoditization of data warehouse software and hardware

With the evolution of data warehouses in the cloud, it is time to take away the complexity traditionally associated with business intelligence infrastructure and democratize data. Next-generation data warehouses have the ability to truly enable a big leap forward in enterprises, allowing on-demand access to make informed business decisions. ■

WIPRO LTD.
Visit: https://www.wipro.com/analytics/
Email: ask.analytics@wipro.com

TRENDS ▸▸

Data Quality:

The Root of Sound Analysis

By Joe Hellerstein

Competitive business strategy increasingly relies on data analytics. The core techniques of data analysis are more accessible thanks to commodity business intelligence packages, modern open source AI tools, and cloud services. Given this level playing field for software and algorithms, competitive advantage typically lies in the unique data that a business can gather and feed into its analytics pipelines.

The Underlying Challenges of Data Quality

It is important to note that the definition of data quality varies depending on the use case. What qualifies as high quality in one instance may not be the same in another. For example, consider "completeness" of temperature readings: Is an analysis dependent on distinct temperature readings every day? Every hour? Every second? Microsecond? This depends on what the readings are being used to assess. So there's no one-size-fits-all notion of data quality, and the use case for the data determines the quality requirements.

Keeping the specifics of a use case in mind, the following are some of the common data quality pitfalls that enterprises face:

- **Questionable statistical validity:** It's common to end up with extreme values in your data that can significantly impact analysis. In one use case, those extreme values might be best treated as "dirty data" and eliminated for the purposes of analysis; in another, they might actually be exactly what you're looking to uncover. On one hand, an outlier in temperature readings might skew your analysis of typical temperature patterns; on the other hand, it might tell you that you have a mis-calibrated sensor, or that someone has been lighting matches near one of your sensors. The statistical properties of data—combined with an understanding of the use case—can dictate whether data is dirty or clean for a given purpose.
- **Not meeting regulatory standards:** Depending on your organization and industry, your data may need to conform to certain standards, as dictated by industry-wide or region-specific (such as GDPR) regulations. These regulations are constantly changing, which can add further complexity and customization into the mix.
- **Missing values:** Depending on how data is collected, it's not uncommon to face issues with completeness. Missing information or values might introduce bias that leads to ineffective decision making. And the fact that data is missing might itself be important—remember Sherlock Holmes and the dog in the nighttime: The "curious incident" was that the dog did not bark. It is all too easy to overlook the data that is not there.
- **Outdated data:** If data is not refreshed often enough, the data, and therefore the subsequent analysis, can be outdated and potentially generate results that are no longer relevant. And in fast-moving organizations, old data is often a small piece of the bigger picture: Estimates suggest that 90% of the world's data was generated in the last 2 years.
- **Non-standardized encodings:** Real-world entities such as people and even dates often have varied encodings and need to be made canonical. Data analysis often stalls until the names of basic entities conform to a single standard.

Today's data professionals are increasingly using new and modern solutions to tackle the various flavors of data quality challenges, but many organizations' data efforts are still hampered by legacy technologies and processes for managing, ingesting, cleaning, and, ultimately, using their data.

New Tactics to Ensure Data Quality

So what can organizations do? Here are three actionable steps your company can take to improve data quality and produce better analyses.

1. Activate collaboration between business teams and IT and collaboration among team members.

- **The challenge:** Individual business teams have an in-depth understanding of how they use data and therefore what their specific quality requirements are for different use cases. Without their active engagement in the data preparation process—better yet, their ability to prepare the data themselves—they cannot have adequate input into how the data is prepared. This brings inefficiencies, potential errors, and an overreliance on IT resources when their time would be better spent serving as a central clearinghouse to align needs across business units and to monitor data quality and practices over time.
- **The solution:** Collaboration between business teams and IT. Many forward-thinking organizations are shifting the responsibility of data quality toward business users, who are closer to the data and understand it better. This brings new efficiencies to data preparation and use. Allowing business users to see and interact with the data sooner enables them to add valuable context to help address data quality issues—they bring a deeper understanding of the use case and can better navigate the quality issues that are actually pertinent to the analysis. Meanwhile, IT's role should be to work across business units, manage data quality tests, and transform processes over time on behalf of the business units.

2. Empower business users to leverage external data resources.

- **The challenge:** New use cases for analytics often motivate business groups to onboard data from external sources. IT professionals do not typically know the use case, related internal data, or the ultimate business user; hence, they're likely ill-equipped to be as creative as the business user about what external data sources could help. Meanwhile, business users who do deeply understand the data and its use case may have the creative ideas but likely can't easily get that external data into pipelines because it's not sanctioned by IT.
- **The solution:** self-service. Business users should have the ability to explore and onboard useful external data, and they should have the ability to incorporate that data into processes. They should also have a way to let IT know what data sources need to be tapped on an ongoing basis.

3. Unite people and AI to help assess and fix data quality.

- **The challenge:** No one would call Excel a modern—or sufficient—tool for preparing or working with data. Yet Excel is still being used as the primary tool for data preparation among 37% of data analysts and 30% of IT professionals. Manual processes such these hinder collaboration and efficiency.

  Algorithmic automation—especially modern AI—can bring massive improvements to data quality. Many aspects of data quality can be automated. For example, AI techniques can be used to flag outlier values, standardize values, or integrate data sources. But AI techniques based on machine learning bring their own data quality requirements: They need clean, labeled "training data" to build their models. Also, almost by definition, AI never gets things entirely right—AI is the technology we reach for when hard-and-fast rules do not work. This is familiar from well-known forms of AI such as Google searches: We typically don't just press the "I'm feeling lucky" button; we review the ranked list of suggested matches and browse through to see what looks most useful.

  As a result, even though AI can help with data quality transformation, it is critical for humans to stay in the loop of automation in efficient ways, assessing data quality before and after algorithms run, and having the ability to observe, steer, and override automated choices at the right level.
- **The solution:** Increasingly, our tools and processes should lean on AI-driven solutions for automation of data quality assessment and transformation. But those solutions must couple AI with a strong human-facing component for efficient and intuitive training, assessment, and override. Advanced tools leverage AI into interactive visual experiences for data quality that ensure quick feedback loops between algorithms and human domain experts. And with machine learning, the accuracy of AI-generated suggestions can improve over time based on that user interaction. This collaboration between people and computation is key to accelerating data preparation and landing on data outputs that inspire confidence in the analytic process.

Improved Data Quality for Better Results

In any data lifecycle, there are many entry points where data quality problems can creep in. From incomplete or improper data entry, incorrectly blending data from different sources, selecting incomplete datasets or missing fields for analysis, to even the analysis process—there is no shortage of chances for data quality problems to arise. Data professionals need to work to lessen and address the quality issues that occur.

Implementing these best practices will help organizations improve data quality and their resulting analysis and ultimately drive better-informed, more strategic business decisions. ■

Joe Hellerstein is co-founder and CSO at Trifacta (www.trifacta.com).

TRENDS ▸▸

For Big Data Insights, Start Small: Follow the Data Breadcrumbs

By Wolf Ruzicka

The data on big data indicates that up to 60% of analytics projects fail or are abandoned, costing companies an average of $12.5 million. That's not the result we seek from data lakes. Instead, companies are increasingly finding themselves mired in data swamps that are overfilled and too muddy to offer any useful visibility. Or are they?

Many companies and government organizations have jumped on the data bandwagon and are hoping to analyze as much data as possible to discover something that might give them a competitive edge or change their operations. Instead of big insights, they have petabytes of big data. They've become compulsive data hoarders while struggling to develop an actionable plan for that data.

The motivation to gain insights from data is good but the approach isn't. Big, traditional, analytical projects are expensive, take years to implement, and aren't as effective as the marketing materials would lead you to believe. A survey of more than 500 executives by McKinsey Analytics found that 86% of respondents believed that their organizations were only somewhat effective at meeting the goals they set for their data and analytics initiatives.

A better approach to big data is to start small. Instead of taking a giant bite of your data, you nibble at it, keeping risks and expenses low. Time and again, we have seen some of the most valuable business insights derived from surprisingly small datasets. We call this approach "minimal viable prediction," or MVP.

MVP is an approach replicated from best practices in software product development. The idea is to focus on delivering a minimum viable prediction as fast as you can and iterate from there.

- Focus only on the most-pressing problem you want to solve.
- Instead of assembling all your data, assemble only the data that correlates with that problem.
- You'll work "backward" as you identify other related data, essentially following the data breadcrumbs that lead to actionable outcomes.

For example, let's say you are an ecommerce grocer. Your customer experience culminates in customers adding items to their shopping carts. You know that product affinity changes with each customer's context, such as the time of day, day of the week, weather, and location. What if you wanted to suggest, in near-real time, items that appeal based on the customer's specific context? Not generic buying patterns based on all Amazon users or the like, but personalized for an individual customer and real data about their real environment when shopping virtually.

You might start by gathering historical, broad market basket data. You may discover that, in general, diaper purchases have the highest affinity with beer purchases. The next data breadcrumb might be a customer's gender data, if available. Another set of data could be weather data by ZIP code. Suddenly, you are starting to see some trends and you've barely started! Perhaps it turns out that product affinity of diapers and beer applies to male customers, but significantly more often during later hours of the day and has overwhelming statistical relevance when it is unusually hot. This is your first MVP that can get immediately validated in the top and bottom lines.

If the first dataset was product affinity, customer gender, and weather data, a second tier of data may include a list of brands and their profitability to you. You dig into that data, and simple A/B testing may uncover that regional brands have higher success rates. While you're probably not going to get rid of the major consumer packaged goods suppliers, you can keep that correlation in mind, and move on to the next dataset. The key point is to establish a methodical process that itera-

PAGE 23 ▶▶

USE CODE **DBT19** TO **SAVE $100 OFF** THE COST OF YOUR CONFERENCE PASS TODAY!

DATA SUMMIT

UNLEASH THE POWER OF YOUR DATA

FEATURING THESE SPECIAL EVENTS

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

MAY 21–22, 2019

PRECONFERENCE WORKSHOPS
MONDAY, MAY 20

HOT TOPICS INCLUDE:

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

HYATT REGENCY BOSTON

BOSTON, MA

dbta.com/datasummit

DIAMOND RESEARCH & KEYNOTE SPONSOR: Pythian

DIAMOND KEYNOTE SPONSORS: GEMINI DATA, ORACLE

PLATINUM SPONSORS: Actian, Aerospike, Attunity, Cambridge Semantics, CloverDX, Data Kitchen, Quest, SQream, Vertica

ASSOCIATION SPONSORS: Cognitive Computing Consortium, IOUG

BROUGHT TO YOU BY: database Trends and Applications, BDQ Big Data Quarterly

MEDIA SPONSORS: BESTSEOs, BDQ, CRM, CrowdReviews, database, DATAFLOQ, DMF, KMWorld, TOPSEOs, Visibility

ORGANIZED AND PRODUCED BY Information Today, Inc.

CONNECT: #DataSummit

DATA SUMMIT
UNLEASH THE POWER OF YOUR DATA

dbta.com/datasummit

CONFERENCE AT-A-GLANCE

MONDAY MAY 20 PRECONFERENCE WORKSHOPS (separately priced or with an All Access Pass)

9:00 a.m. – 12:00 p.m.	W1 ▮ Cognitive Computing 101	W2 ▮ Data Ops 101
1:30 p.m. – 4:30 p.m.	W3 ▮ Data Science Best Practices	W4 ▮ Machine Learning Best Practices

TUESDAY MAY 21

8:00 a.m. – 9:00 a.m.	**CONTINENTAL BREAKFAST**	
9:00 a.m. – 9:45 a.m.	**WELCOME & KEYNOTE** ▮ **Big Data, Technological Disruption, and the 800-Pound Gorilla in the Corner** ▮ ***Michael Stonebraker,*** *MIT & Tamr*	
9:45 a.m. – 10:00 a.m.	**SPONSORED KEYNOTE** ▮ Oracle	
10:00 a.m. – 10:45 a.m.	**COFFEE BREAK** ▮ In the Data Solutions Showcase	
	TRACK A ▮ Modern Data Architecture	**TRACK B** ▮ Competing on Analytics
10:45 a.m. – 11:45 a.m.	**A101 ▮ Building a Modern Data Architecture**	**B101 ▮ Taking Your Analytics to the Next Level**
12:00 p.m. – 12:45 p.m.	**A102 ▮ The New World of Database Technologies**	**B102 ▮ Data Science Best Practices**
12:45 p.m. – 2:00 p.m.	**ATTENDEE LUNCH** ▮ In the Data Solutions Showcase	
2:00 p.m. – 2:45 p.m.	**A103 ▮ Understanding Cloud Licensing**	**B103 ▮ Analytics in Action**
2:45 p.m. – 3:15 p.m.	**COFFEE BREAK** ▮ In the Data Solutions Showcase	
3:15 p.m. – 4:00 p.m.	**A104 ▮ Overcoming Big Data Integration Challenges**	**B104 ▮ Delivering Trusted Data**
4:15 p.m. – 5:00 p.m.	**A105 ▮ Securing the Internet of Things**	**B105 ▮ Everyday Chaos**
5:00 p.m. – 6:00 p.m.	**NETWORKING RECEPTION** ▮ In the Data Solutions Showcase	

WEDNESDAY MAY 22

8:00 a.m. – 8:45 a.m.	**CONTINENTAL BREAKFAST**	
8:45 a.m. – 9:30 a.m.	**OPENING KEYNOTE** ▮ **Digital Transformation Is Business Transformation** ▮ ***Michelle L. Gregory,*** *Data Science, Elsevier*	
9:30 a.m. – 9:45 a.m.	**SPONSORED KEYNOTE** ▮ Pythian Group Inc.	
9:45 a.m. – 10:00 a.m.	**SPONSORED KEYNOTE** ▮ Gemini Data	
10:00 a.m. – 10:45 a.m.	**COFFEE BREAK** ▮ In the Data Solutions Showcase	
	TRACK A ▮ Building the Data-Driven Future	**TRACK B** ▮ Digital Transformation
10:45 a.m. – 11:30 a.m.	**A201 ▮ Winning With a Modern Data Strategy**	**B201 ▮ Achieving a 360-Degree Customer View**
11:45 a.m. – 12:30 p.m.	**A202 ▮ Supporting Modern Applications**	**B202 ▮ Digital Transformation in the Real World**
12:30 p.m. – 2:00 p.m.	**ATTENDEE LUNCH** ▮ In the Data Solutions Showcase	
2:00 p.m. – 2:45 p.m.	**A203 ▮ Designing for Speed & Scalability**	**B203 ▮ Tapping Into New Data Sources for Business Value**
3:00 p.m. – 3:45 p.m.	**A204 ▮ The Rise of Knowledge Graphs**	**B204 ▮ Emerging Applications for Blockchain**
4:00 p.m. – 5:00 p.m.	**CLOSING KEYNOTE** ▮ **Bring It Home: How to Advance Your Analytics Strategies** ▮ ***John O'Brien,*** *Radiant Advisors*	

KEYNOTES

TUESDAY MAY 21

9:00 a.m. – 9:45 a.m.

WELCOME & KEYNOTE

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

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

9:45 a.m. – 10:00 a.m.

SPONSORED KEYNOTE

ORACLE

WEDNESDAY MAY 22

8:45 a.m. – 9:30 a.m.

OPENING KEYNOTE

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

Michelle L. Gregory, *SVP, Data Science, Elsevier*

4:00 p.m. – 5:00 p.m.

CLOSING KEYNOTE

Bring It Home: How to Advance Your Analytics Strategies

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

DATA LAKE BOOT CAMP

C101 ▮ Building a Data Lake for the Enterprise	
C102 ▮ Taking Your Data Lake to the Cloud	
C103 ▮ PANEL: Data Lakes: Challenges and Opportunities	
C104 ▮ Data Lakes in Action	
C105 ▮ Frameworks for the Future	

Cognitive Computing & AI Summit

	CS101 ▮ The Rise of Artificial Intelligence
	CS102 ▮ Machine Learning in the Real World
	CS103 ▮ AI in Action
	CS104 ▮ AI Success Factors
	CS105 ▮ Exploring Machine Learning

DATAOPS BOOT CAMP

C201 ▮ Succeeding With DataOps Today	
C202 ▮ The Rise of Containers	
C203 ▮ Operationalizing Big Data Workloads	
C204 ▮ Unlocking the Power of Data Wrangling	

Cognitive Computing & AI Summit

	CS201 ▮ Machine Learning Best Practices
	CS202 ▮ Diving Into Deep Learning
	CS203 ▮ AI Use Cases Today
	CS204 ▮ PANEL: Cognitive Computing

MONDAY MAY 20

PRECONFERENCE WORKSHOPS

Data Summit 2019 *offers four half-day preconference workshops on Monday, May 20 that provide immersive training for data professionals. Access to two workshops is included when you register for an All Access Pass. Workshops may also be registered for separately.*

9:00 a.m. – 12:00 p.m.

W1 ▮ Cognitive Computing 101

Hadley Reynolds, *Co-Founder & Executive Director, Cognitive Computing Consortium*
Sue Feldman, *President, Synthexis*

A new era of cognitive computing is unfolding, and its impact is already being felt across industries, from preventative maintenance at manufacturing plants and patient diagnosis at hospitals to the rise of sophisticated chatbots ready to assist us across the connected world. The goal of cognitive computing is straightforward: to simulate human thought processes in a computerized model. However, building cognitive systems and applications that can perform specific, humanlike tasks in an intelligent way is far from easy. Attend this workshop to get a full understanding of how cognitive computing works, popular use cases, and best practices IT leaders and practitioners can apply today.

W2 ▮ Data Ops 101

Mark Marinelli, *Head of Product, Tamr*

DataOps has emerged as an agile methodology to improve the speed and accuracy of analytics through new data management practices and processes, from data quality and integration to model deployment and management. By leveraging automation, data democratization, and greater collaboration among data scientists, engineers, and other technologists, DataOps can help organizations improve the time-to-value of their data. Attend this workshop to hear about the key supporting technologies, real-world strategies, and success stories and how to get started on your DataOps journey.

1:30 p.m. – 4:30 p.m.

W3 ▮ Data Science Best Practices

Joe Caserta, *Founding President, Caserta*

Data science, the ability to sift through massive amounts of data to discover hidden patterns and predict future trends, may be the "sexiest" job of the 21st century, but it requires an understanding of many different elements of data analysis. Extracting actionable knowledge from all your data to make decisions and predictions requires a number of skills, from statistics and programming to data visualization and business domain expertise. Attend this workshop for a deep dive into the fundamentals of data exploration, mining, and preparation, applying the principles of statistical modeling and data visualization in real-world applications.

W4 ▮ Machine Learning Best Practices

Chelsey H. Hill, *Assistant Professor of Business Analytics, Feliciano School of Business, Montclair State University*

Machine learning (ML) is on the rise at businesses hungry for greater automation and intelligence with use cases spreading across industries. At the same time, most projects are still in the early phases. From selecting datasets and data platforms to architecting and optimizing data pipelines, there are many success factors to keep in mind. The advantages that ML offers organizations–the ability to automatically build models that can analyze huge volumes of data and deliver lightning-fast results–have also led to a growth in the availability of both commercial and open source frameworks, libraries and toolkits for engineers. Attend this workshop for a hands-on course in the enabling technologies, techniques, and applications you need to know to succeed in today's environments.

Preconference workshops are practical and hands-on. Please bring a laptop with you so you can participate in the exercises. Laptops will not be provided on site.

BRING YOUR TEAM & SAVE 40%! LEARN HOW ON OUR REGISTRATION PAGE.

DATA SUMMIT
UNLEASH THE POWER OF YOUR DATA

UNLEASH THE POWER OF YOUR DATA AT DATA SUMMIT 2019

AT DATA SUMMIT 2019, you'll hear the innovative approaches the world's leading companies are taking to solve today's key challenges in data management. Whether your interests lie in the technical possibilities and challenges of new and emerging technologies or using Big Data for business intelligence, analytics, and other business strategies, Data Summit 2019 has something for you!

dbta.com/datasummit

TRACKS AT DATA SUMMIT 2019

At the Data Summit conference, we have tracks and special events designed for everyone in involved in data and information management from your organization. You and your team can choose from the following:

Cognitive Computing & AI Summit

A new era of cognitive computing has already begun, and its impact is being felt across industries, from healthcare and financial services to manufacturing and education. However, building cognitive systems and applications that can perform specific, humanlike tasks in an intelligent way is far from easy. This one of-a kind-event is an intense, 2-day immersion into the leading cognitive computing and AI use cases, strategies, and technologies that every organization should know about. If you are on the front lines of AI and cognitive computing, this summit is for you.

Designed for chief information officers, chief data officers, data scientists, data engineers, enterprise architects, systems architects, application developers, and tech-savvy business leaders.

DIGITAL TRANSFORMATION

The Digital Transformation Track helps you navigate the core technologies disrupting the business world today, from Big Data and cloud computing to artificial intelligence and the Internet of Things. Come learn how these technologies work and how businesses are harnessing them to increase efficiency, agility, and innovation. Attend this track to understand the game-changing technologies and how you can use them to succeed.

This track is designed for business executives, line of business managers, marketing professionals, sales professionals, innovation strategists, and new information technologists.

COMPETING ON ANALYTICS

The Competing on Analytics Track is your guide into the most important trends in analytics today, from dark data and machine learning to data science and prescriptive insights. Attend this track to understand the latest trends in data discovery and visualization, how new data preparation techniques are changing the analytics game, emerging best practices in statistical modeling, and the impact of cloud computing and new data platforms.

This track is designed for chief information officers, chief data officers, data scientists, data analysts, BI architects, BI developers/engineers, digital analytics directors/managers, and data analytics technologists.

BUILDING THE DATA-DRIVEN FUTURE

The Building the Data-Driven Future Track equips you with new insights and practical advice on the key technologies and emerging best practices in building modern systems and applications, from containers and microservices to recommender systems, mobile IoT applications, and cloud security. Attend this track to navigate the latest trends in Big Data and cloud engineering and agile application development.

This track is designed for chief information officers, systems architects and engineers, software developers and engineers, application developers, database developers, and associated IT managers and directors.

MOVING TO A MODERN DATA ARCHITECTURE

The Moving to a Modern Data Architecture Track takes you through the latest advancements in data management, from NoSQL and Hadoop to in-memory computing, cloud migration, data lakes, and real-time architectures. Attend this track to gain a deeper understanding of the latest technologies and techniques underpinning the hybrid future of data management and how you can succeed in this rapidly changing world.

This track is designed for data scientists, data engineers, enterprise architects, systems architects, application developers, data administrators, and associated information professionals.

DATA LAKE BOOT CAMP

From centralized data acquisition and offloading to data discovery and data science projects, data lakes are on the rise at enterprises today. Data Lake Boot Camp offers attendees a deep dive into the latest supporting technologies, best practices, real-world success factors, and expert insights.

Hadoop Day is designed to appeal to Hadoop experts and novices, business intelligence professionals, technologists, and data analysts.

DATAOPS BOOT CAMP

The world of data management, with its rigid schemas, silos, and manual processes, has historically been at odds with the fast, automated, highly iterative world of DevOps. At DataOps Boot Camp, you hear about the key supporting technologies, strategies, real-world success stories, and how to get started on your DataOps journey.

Designed for data scientists, architects, and engineers, as well as technology decision-makers and administrators. Both DataOps veterans and novices are welcome.

DID YOU KNOW?

There are a variety of combination passes and standalone passes available to allow you to attend the whole event, or focus on content geared toward your particular needs. Visit our registration page for details.

tively examines your top headache … or opportunity.

The approach is simple and straightforward, but keeping things small is actually difficult. It's a mindset that requires organizations to answer one question at a time and follow the data path to a conclusion. Armed with these little insights, it's tempting to want to dive deeper.

Here are a few tips to stay the course and embrace MVP:

Collect thoughtfully. Data becomes dated—fast. Unless your questions focus on historical changes, spring-clean your data regularly and eliminate data clutter.

Iterate. Don't waste time by taking too big a bite. Predictive analytics is a process of trial and error. So limit your focus and iterate, trying new paths until you find one that leads you in the right direction.

Prioritize. List all the problems or questions you would like to solve with data analytics. Rank them by potential business impact and ease of assembling the first set of data. Then pick the highest-ranked problem, and follow the breadcrumbs.

Stay focused. If you hit a dead end, retrace your steps and follow the next set of data breadcrumbs. You could easily assemble another set of data but don't—it will muddle your quest to answer question number one. The more self-discipline and focus you demonstrate, the easier it is to generate insights and achieve a better outcome. As you continue to iterate, include more data, where and when it's relevant, and only then.

Refine. Once you've answered your number-one problem and have your first MVP, refine it again and again so you have an ongoing stream of actionable, affordable predictions.

The cloud is your friend. Cloud-based pay-as-you-go tools, such as Microsoft's Power BI, are ideal companions for this approach. Data management, analysis, and visualization tools in out-of-the-box, cloud-based packages make it feasible for organizations to become data savvy, regardless of budget or analysts on staff.

Once you have your first MVP, your project is already more successful than 50% of the big data projects out there. We can attest to seeing results in 2–3 months—rather than 2–3 years. And that's where the MVP approach excels. By only focusing on the relevant data breadcrumbs and starting small, it becomes much easier to gauge whether you're on the right track and to correct any mistakes relatively cheaply.

The MVP approach requires a mindset shift inside the organization—one that values agility and accepts trial and error. It's as much a cultural change as it is a data approach. Initially, the MVP approach's small scope might feel too small for organizations experiencing a revenue-impacting pain point. The approach was challenged when it was first used in software development too. Have faith that starting small—with limited questions, investment, and time—will produce results faster, less expensively, and without the risk of big-bite data approaches. ■

Wolf Ruzicka is chairman of EastBanc Technologies (www.eastbanctech.com).

MV SOLUTIONS

ROCKET SOFTWARE BOOSTS ROOMS TO GO INVENTORY AND BUSINESS

Rocket Software recently helped Rooms to Go (RTG) with managing inventory as well as improving and enhancing customer interaction in order to accelerate sales growth.

Founded in 1990, RTG is an innovator of personalized furniture shopping. RTG pioneered the concept of displaying and packaging furniture in complete room settings, allowing customers to easily visualize how to transform their homes with the perfect furnishings. RTG designers create pre-packaged rooms with coordinated colors, fabrics, and accessories in order to provide a simplified and convenient shopping experience for their customers.

The company deployed Rocket UniVerse early in its history to help manage inventory as well as improve and enhance customer interaction.

Partnering with Rocket has helped RTG grow from three showrooms in Florida in 1991 to more than 200 locations and seven advanced distribution centers across the U.S.

In the early 2000s, RTG found that the green-screen applications supporting its sales staff weren't delivering information fast enough to increase sales growth. The sales team wanted to deliver an even more intimate, personalized shopping experience.

Continuing to evolve, the company, with the support of Rocket Software, debuted a new tablet-based retail application in 2012 that sales associates now use everywhere in the showroom.

The RTG merchandising team uses its product data management system to load configurations and manage inventory while UniVerse pulls the data and serves up targeted descriptions of each asset.

The combination of the UniVerse back end with the new Windows interface has increased RTG's ability to deliver more personalized customer experiences. Not only are sales associates freed from the kiosk, they're now able to answer a range of questions by simply entering a customer's ZIP code, as well as provide views and availability of different variations of sofa designs using their "iSofa" application.

The RTG team is also able to call up all inventory items and offer additional items that are not physically in the showroom. The application showcases furniture designs and combinations of complementary accessories, such as rugs and pillows, that ultimately lead to increased sales and higher levels of customer satisfaction.

RTG has accelerated sales growth by 30% as a result of helping customers visualize a complete range of products and complementary accessories. The company continues to grow as a dominant furniture retailer in the Southeastern U.S. and Puerto Rico with annual revenues exceeding $2 billion.

For more information about this news, visit www.rocketsoftware.com.

IT WORLD CANADA CHOOSES ENTRINSIK INFORMER TO MAINTAIN CAMPAIGNS

IT World Canada recently used Entrinsik Informer to manage all of its campaigns and deliver contracted registrations on time.

More than 75,000 IT executives and professionals in Canada rely on IT World Canada's technology information. As the Canadian affiliate of International Data Group (IDG), IT World Canada is the leading Canadian online multimedia information provider with digital titles including CanadianCIO.com, IT Business.ca, ComputingCanada.com, ComputerDealerNews.com, and Directioninformatique.com.

It is critical, from a customer satisfaction perspective, to constantly monitor all of its clients' campaigns to ensure the high-quality leads coming from registrations are delivered to every client on time.

IT World Canada uses Informer for report generation, decision support, and business process management. Informer's Dataset enables the company to consolidate the different campaign Datasources while Dashboards provide up-to-the-minute reporting and act as a hub to link all the different software tools used to the individual campaigns for easy access.

IT World Canada also faced additional challenges, such as clients requiring individualized formatting for their lead reports to align with their CRM system configuration. In addition, campaign managers and staff needed a convenient way to easily access the different software tools and files used in managing each individual campaign.

Using Informer, IT World Canada now produces an overview of all the campaigns it is running for clients, so they can easily see the status of the campaigns and identify ones that need immediate attention.

According to Arlo Murphy, director of business intelligence and audience services at IT World Canada, "We rely on Informer's powerful reporting and dashboard functionality to keep us on top of all of our campaigns so that we can deliver every registration to every client on time."

Using Informer's Dashboards, IT World Canada set up a Campaign Monitor to allow campaign managers to understand the status of each campaign and exactly where they should prioritize their resources to ensure the success of each. This enabled them to eliminate wasted time in non-critical areas.

The Campaign Monitor also shows campaign managers the departments that are currently working on specific campaigns and their associated outstanding tasks. In addition, the Campaign Monitor makes it easier for campaign managers to

ensure the success of campaigns by acting as a central hub that not only provides status reporting but access to all the software tools used to manage each campaign.

For more information, go to www.entrinsik.com.

BLUEFINITY ADDS A VARIETY OF NEW FEATURES IN LATEST EVOKE UPDATE

BlueFinity's Evoke platform is adding a series of updates and enhancements, including chatbots, a signature panel, and multiple-developer support.

Evoke is a rapid app development platform that provides the complete environment for existing staff to design, develop, and deploy business apps across multiple devices (iOS, Android, and Windows phones and tablets, plus Windows, Apple, and Linux desktops).

This update introduces a range of new graphs and charts that can now be included as an integral part of apps developed using Evoke. The inclusion of graphics in any application is a very effective way of visualizing important information quickly and clearly.

As with other aspects of Evoke, the company says graphs are made available across all platforms and device types. Evoke features such as automated resizing according to the screen size, adaptations to design the screen for optimum screen layout, and WYSIWYG design capabilities to further customize the layouts for each screen are all made available when incorporating graphs and charts into an app.

Evoke supports the inclusion of data from multiple files in multiple databases for use within a single app. Users can configure the data fields (columns/attributes) associated with the data, as well as any required calculation and manipulation of the information.

Evoke offers a low code or no code route which can evolve into fully customizable apps as required. It allows users to create web, hybrid, and even native apps and to integrate and synchronize with existing back-end systems and SQL and MultiValue databases.

For more information about these updates, visit www.bluefinity.com.

ROB MANDEVILLE

Rob Mandeville is a senior DBA at SolarWinds, experienced with major RDBMS vendor engines.

The Problem With Defining Database Workload

Great! Wait ... How do we define database workload? Good question! I've researched many approaches to this question and still don't have a definitive answer or way to measure. Let's dive into some of the research and personal thoughts I've had around this topic.

Here are the Definitions of Workload, According to www.merriam-webster.com:

1. The amount of work or of working time expected or assigned
2. The amount of work performed or capable of being performed (as by a mechanical device) usually within a specific period

I like the second definition a little bit more as it applies to IT concepts. Extrapolating that for our purposes, we can say that database workload is the amount of work performed or capable of being performed by a database within a specific period. I like the idea of it being within a specific period though, so that's a start. For our purposes, it is useful to stick to what can be observed or measured, so I'm going with the actual amount of work performed for my definition, rather than the potential work that could be done. My kids are capable of getting straight A's, but that's not what I've observed.

What is database work? Again, another good question! During my research, I found one concept I agreed with—database work should be looked at without regard to resources available or resources consumed. The reason I agree with this is that we should try to remove as many variables as possible. Every system is different, including the number of cores plus clock speeds of the chip architecture; the amount of RAM; disk speeds (spinning versus SSD); local disk versus SAN or NAS; network bandwidth; additional load on the system from non-database processes; virtualization (where other VMs are running on the same physical server) versus physical; and whether it is cloud-based.

Excluding external dependencies from the equation, what is the actual database work? We could say it is the aggregate "asks" placed on the database engine such as application calls (queries); maintenance jobs; internal database engine management (memory management, parses, optimizer activity, etc.); ad hoc (i.e., administrative queries, data modifications, access control changes, etc.); release changes (i.e., schema changes, data migrations, data definition language, etc.); and anything else asked of the database.

Devilish Details

Conceptually, this is a decent proxy to database work. However, the devil really is in the details. What about this work do we want to capture or use to define database workload? Resources consumed? No, that's not quite right as that's more of a byproduct of the workload playing out against our system and is subject to a lot of variables. A tally of database calls regardless of the type of call? That's not quite right either, as all calls are not created equal. Number of active sessions? That's still not quite right, plus how should we handle parallel executions? Statistics (logical reads, executions, physical reads, parses, etc.)? Maybe stats can be part of the overall workload definition and quantification, but stats would be dependent on resources allocated to my system (an example being a physical read on one system could be a logical read on another system if you have enough memory to hold the data you need in cache). Some amalgamation of these things?

Is Workload Scaling Linearly?

This may be late in the game, but some reading this article may ask: Why do we want to define workload? We can run a specific static workload set against different systems that can help us determine benchmarks. We can amp up that workload to see where the system breaking point is. However, those scenarios are using a known workload to define other things. What I'm talking about is measuring actual workload with an eye on the prize of knowing when that workload is no longer scaling linearly. This is likely a review but just to level-set, an example of linear scaling is doing 100 units of work in 1 hour, then doing 200 units of work in 2 hours. You are no longer scaling linearly if it then takes 6 hours to do 300 units of work. Somewhere between 200 and 300 units, there was an elbow (started taking more time than previously to do one more incremental unit). That's where I'm headed—I want to know based on actual workload, where is my elbow or where should I expect it to be?

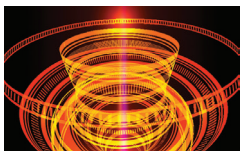
The Wrap-Up

Another approach could be just to look for inefficient workload. I'll save that for another article. So, here at the end of this post, there are a lot of ideas around database workload but no definitive answers (in my mind). Please comment with your thoughts around how to define database workload—I would love to hear your thoughts and feedback! ■

GUY HARRISON

Guy Harrison, a software professional with more than 20 years of experience, is a partner at Toba Capital and the author of *Next Generation Databases* (Apress). Contact him at guy@tobacapital.com.

Spanner Stretches the CAP Theorem



EMERGING TECHNOLOGIES

In distributed databases, no principle is quite as famous as "CAP" Theorem (also known as Brewer's Theorem). CAP Theorem states that a distributed database can at most support two of the following three desirable characteristics:

- Consistency: Everybody gets the same view of the database.
- Availability: The database stays online through a failure of at least a minority of nodes.
- Partition Tolerance: The database stays available during network partitions.

CAP Theorem sounds complex, but it is pretty easy to understand in practice. If I have a database that has nodes in the U.S. and in Australia, what happens if the network between the U.S. and Australia fails? In some distributed databases, this is called the "split brain" scenario. Oracle's RAC clustered database chooses consistency in this scenario: The smaller of the two partitions will shut down (probably Australia). Amazon's Dynamo chooses availability: Both Australia and the U.S. stay online, but Australians and Americans may see slightly different views of the database until the network is restored. You simply cannot serve up identical views of an active database to two locations if there is no network connectivity between those two locations.

The implications of CAP Theorem, more than anything else, led to the schism in modern database management systems. With the rise of global applications with extremely high uptime requirements, it became unthinkable to sacrifice availability for perfect consistency. Almost in unison, the leading Web 2.0 companies such as Amazon, Google, and Facebook introduced new database services that were only "eventually" consistent but globally and highly available.

Google has tried to resolve this database schism with its Spanner SQL database. Google has not claimed to have overthrown the CAP Theorem but has instead adopted a "You don't need to worry about that" philosophy.

CAP Theorem assumes that network partitions are inevitable in a wide area network. And in the universal wide area network of the internet, this is undoubtedly true—you simply can't assume network availability when the network is constructed of so many varied service providers. But Spanner runs exclusively on Google's global network. Google's network has sufficient redundancy to eliminate hardware failure as a likely cause of a network partition and has adopted procedures designed to minimize the possibility of a human error-driven network failure.

Google doesn't say a network partition is impossible; rather that it has made it much less probable than other possible failure scenarios. Therefore, for most applications, network partitions should no longer be a major concern. Nevertheless, it is worth noting that should a network partition occur, Spanner chooses consistency over availability, which means it has more in common with traditional databases such as Oracle than with next-generation databases such as Dynamo.

Another other novel feature of Spanner is its TrueTime system. Distributed databases go to a lot of effort to return consistent information from replicas maintained across the system. Locks are the primary mechanism to prevent inconsistent information from being created in the database, while snapshots are the primary mechanism for returning consistent information. Queries don't see changes to data that occur while they are executing because they read from a consistent "snapshot" of data. Maintaining snapshots in distributed databases can be tricky: Usually there is a large amount of inter-node communication required to create agreement on the ordering of transactions and queries.

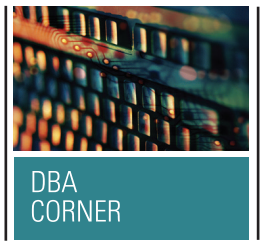
Google Spanner simplifies the snapshot mechanism by using GPS antennas and atomic clocks physically installed in each server. GPS provides an externally validated timestamp while the atomic clock provides high-resolution time between GPS "fixes." The result is that every Spanner server across the world has the same clock time. This allows Spanner to order transactions and queries precisely without requiring inter-node communication.

Google engineering was once described as "Ph.D.s driving tanks." Spanner is a typical Google combination of smart architecture, pragmatism, and brute-force engineering. It minimizes the impact of the CAP Theorem on distributed systems by providing a database that—almost—lets you "have it all." ■

CRAIG S. MULLINS

Craig S. Mullins is president of Mullins Consulting, Inc. He's an IBM Gold Consultant, IBM Champion for Analytics, and the author of three best-selling books, *DB2 Developer's Guide, Database Administration: The Complete Guide to DBA Practices & Procedures,* and *A Guide to Db2 Application Performance for Developers.*
Website: www.mullinsconsulting.com

Three Important Database Security Features



One of the most important and rapidly changing areas of database management is security and protection. The major DBMS vendors have been adding security features and improving the way you can protect your precious corporate data. But it can be difficult to keep up with these features, so let's take a brief look at some of the more interesting and useful database security options.

Encryption

Although encryption has been around for quite a while, it has only recently become an important aspect of database security for protecting sensitive data. When data is encrypted, it is transformed using an algorithm to make it unreadable to anyone without the decryption key. The general idea is to make the effort of decrypting so difficult as to outweigh the advantage to a hacker of accessing the unauthorized data. There are two situations where data encryption can be deployed: data in transit and data at rest. In a database context, data "at rest" encryption protects data stored in the database, whereas data "in transit" encryption is used for data being transferred over a network.

*A growing number of **DBMSs** offer label-based access control, which delivers more fine-grained control over authorization to **specific data**.*

Encrypting data at rest is undertaken to prohibit "behind the scenes" snooping for information. When the data at rest is encrypted, even if a hacker surreptitiously gains access to the data behind the scenes, without the decryption key the data is meaningless. Data at rest encryption most commonly is supported by using built-in functions, a DBMS feature such as Oracle Transparent Data Encryption, or through an add-on encryption product.

Encrypting data in transit protects against network packet sniffing. If the data is encrypted before it is sent over the network and decrypted upon receipt at its destination, it is protected along its journey. Anyone nefariously attempting to access the data en route will receive only encrypted data. And again, without the decryption key, the data cannot be deciphered. Data in transit encryption most commonly is supported using DBMS system parameters and commands or through an add-on encryption product.

Label-Based Access Control

A growing number of DBMSs offer label-based access control (LBAC), which delivers more fine-grained control over authorization to specific data in the database. With LBAC, it is possible to support applications that need a more granular security scheme. LBAC can be set up to specify who can read and modify data in individual rows and/or columns.

LBAC is not for every application; it is geared more for top-secret, governmental, and similar types of data. For example, you might want to set up an authorization scenario such that each column and row has specific rules pertaining to which employees can see and manipulate the data. Setting up such a security scheme is virtually impossible without LBAC. An administrator configures the LBAC system by creating security label components, which are database objects used to represent the conditions determining whether a user can access a piece of data. A security policy, composed of one or more security label components, is used to describe the criteria for determining who has access to what data. The security administrator defines the policy by creating security labels that are composed of security label components. Once created, a security label can be associated with individual columns and rows in a table to protect the data held there. When a user tries to access protected data, that user's security label is compared to the security label protecting the data.

Any attempted access to a protected column when the LBAC credentials do not permit that access will fail. If users try to read protected rows not allowed by their LBAC creden-

tials, the DBMS simply acts as if those rows do not exist. This is important because sometimes even having knowledge that the data exists (without being able to access it) must be protected.

Consult your DBMS documentation for where and how to establish this hierarchy and how to use LBAC.

Data Masking

An additional technique for protecting database data is to deploy data masking and obfuscation. Data masking is the process of protecting sensitive information in databases from inappropriate visibility by replacing it with gibberish or realistic but not real data (in the case of production data used in test systems). The goal is that sensitive personally identifiable information is not available outside of the authorized environment. Protecting sensitive data using data masking can prevent fraud, identity theft, and other types of criminal activities. A common usage of data masking is to comply with PCI-DSS regulations to show only the last four digits of a payment card number on a receipt.

Data masking can be done while provisioning test environments so that copies created to support application development and testing do not expose sensitive information. Valid production data is replaced with usable, referentially intact, but incorrect or obfuscated data. After masking, the test data is usable just as with production data but the information content is secure.

__Data masking__ is the process of protecting sensitive information in databases from inappropriate visibility by replacing it with gibberish or realistic but not real data and its goal is that __sensitive information__ is not available outside of the authorized environment.

It is possible to mask data using a variety of techniques. A good data masking solution should offer the ability to mask using multiple techniques. Common techniques include substitution, shuffling, number and data variance, nulling out, encryption, and table-to-table synchronization. Data masking is supported by many DBMS offerings as well as by third-party products.

Staying Up-to-Date

Be sure to keep up-to-date on the latest security requirements and capabilities of your DBMS. Understand what is available to you, and what you may need to augment with additional tools. And keep in mind that the items covered here are not the only security features available to you for protecting your database data. ■

SETH MILLER

Seth Miller, senior principal software engineer with Veritas Technologies, has been working with Oracle technologies since 2005 and specializes in database administration, and solutions integration. He currently serves as the director of communications for the IOUG.

Tips for Transitioning to the Oracle Cloud

Transitioning to the cloud can be a daunting thought. Logging into a cloud portal for the first time and not knowing where anything is or how to execute the most basic tasks can elicit self-doubt in even the most experienced technologist. For some, it may feel similar to starting over. This article aims to squelch that self-doubt for those moving into the Oracle Cloud by providing enough knowledge to get started with Oracle Cloud Infrastructure Database Services—quickly and with confidence.

Classic Versus OCI

There are two versions of Oracle Cloud. The current "second generation" platform is called Oracle Cloud Infrastructure (OCI). When Oracle first entered the public cloud market, the product was called Oracle Public Cloud (OPC). Many of the OPC services are still active and run separately from OCI. In most cases, the way to differentiate between them is to look for the term "Classic." This indicates that the service is part of the original OPC infrastructure. OPC was later renamed OCI-Classic. Anything not referred to as Classic is part of OCI.

Portals

When you first sign into your Oracle Cloud account, you will probably be on the "Dashboard" or "My Services" page. This is the top-level portal which has access to both OCI and OCI-Classic services. The slide-out pane on the left of the dashboard has a list of all of the services available to your account. Selecting a Classic service such as "Compute Classic" will bring you to a separate portal for the Compute Classic service. Selecting any service that is not a Classic service such as "Compute" will bring you to the Compute OCI service page in the OCI portal. The difference is subtle but important. The Classic services have their own isolated portals. The OCI services are all part of the same OCI portal.

The slide-out pane on the left changes, depending on the type of portal you are in. The Dashboard portal will show the OCI-Classic and OCI services. Once you are in any of the OCI services, the pane will no longer show any OCI-Classic services but will include the paths to all other OCI services.

Tenancy

The original Oracle Cloud, now called OCI-Classic, referred to accounts as "identity domains." But that term hasn't really carried over to OCI. Instead, everything is based on the concept of a "tenancy" within OCI. For all intents and purposes, the tenancy can be considered the "account."

Compartments

Multitenancy in OCI extends far beyond the database. There are myriad ways of isolating OCI components and one of the highest levels is "Compartments." Compartments are a logical grouping of cloud resources. Privileges (officially called IAM Policies) can also be applied at the compartment level, meaning that there are also practical benefits of compartments beyond logical separation and organization. Compartments are part of OCI only. None of the Classic services will utilize or be constrained by compartments or any other OCI governance components.

All OCI service components are isolated by Compartment. You can see the Compartment in which you currently reside in any service by looking under the "List Scope" header on the left side of the screen. If at any time you don't see the service components you are looking for, make sure the correct Compartment is selected.

The path to Compartments administration can be found in the left slide-out pane under the "Identity" menu. A "root" Compartment is created when your account is first provisioned. Before you start provisioning services, you will have to create at least one additional Compartment.

Virtual Cloud Networks

A Virtual Cloud Network (VCN) is a group of networking components used by most OCI services. A VCN closely resembles a traditional network, with firewall rules and various types of gateways to communicate across internal networks and over the internet.

Most services require a VCN to exist before they can be provisioned. Some services, such as the Compute service, will give the option of creating the VCN as part of the provisioning process.

The minimal creation of a VCN will only provision some of the components required. The easiest way of making sure the VCN has everything needed for a functional and accessible instance is to select "Create Virtual Cloud Network Plus Related Resources" when creating the VCN. This will add an Internet Gateway to allow access to your services over the internet, a subnet for each of the Availability Domains in the region, DNS resolution, and dynamic IP addresses.

This is where OCI differs from some of the other major cloud providers. With OCI, you must set up at least one VCN and the

associated resources before you can start provisioning and using other infrastructure and resources. This contrasts with some other major cloud providers where the network basics are done automatically or by default.

SSH Keys

It is required that you create a shared SSH (Secure Shell) key pair before provisioning an instance. A common way to do this is to use the PuTTY suite of SSH tools, specifically the PuTTY Key Generator. Many of the SSH key generator tools do not save the public key in the format that OCI services expect. To create the public key file, simply copy the text from the key generator under the heading, "Public key for pasting into OpenSSH authorized_keys file," and save it to a file. Don't forget to save the private key as well.

Compute

With a VCN in place and an SSH key pair created, a Compute Instance can be provisioned. OCI goes beyond OCI-Classic by adding the ability to provision Bare Metal Machines, which gives you a dedicated physical server.

For example, to create a basic Linux Virtual Machine (VM), select "Instances" under the Compute menu from the slide-out pane on the left. When choosing the SSH key file, select the file onto which you copied the OpenSSH formatted public key. All of the networking components should default to the VCN components you created.

It won't take long for the VM to be in a running state, at which point you can log into it with SSH. When setting up your SSH connection, make sure to use shared key authentication and choose the SSH private key file. The username for Oracle Linux compute instances is "opc," which is one of the many remnants left over from the original Oracle Public Cloud. The opc user has permission to performed privileged operations using sudo.

Database

The OCI Database service has prerequisites similar to the Compute service. This is mostly due to the fact that compute instances are created to support the database.

Select "Bare Metal, VM, and Exadata" under the "Database" heading from the slide-out pane on the left. Make sure the correct Compartment is selected and select "Launch DB System."

"Shape" indicates the size of the compute instance(s) that will support the database. The shape number indicates the compute capability of the instance. The first number indicates the hardware platform on which the instance will run: 1 indicates that the instance will run on an Oracle X5 server, while 2 indicates the instance will run on an Oracle X7 server. The second number indicates the number of CPU cores that will be allocated. Choosing a shape with more than one CPU will allow you to select up to two nodes for an optional RAC (Real Application Clusters) database configuration.

"Software Editions" indicates that certain database features are enabled. The following list indicates which version and features are enabled for each option. Each Enterprise edition includes the features of the lower Enterprise editions.

- Standard Edition—Oracle Database Standard Edition 2
- Enterprise Edition—Oracle Database Enterprise Edition, Data Masking and Subsetting Pack, Diagnostics and Tuning Packs, and Real Application Testing
- Enterprise High Performance—Multitenant, Partitioning, Advanced Compression, Advanced Security, Label Security, Database Vault, OLAP, Advanced Analytics, Spatial & Graph, Database Lifecycle Management Pack, and Cloud Management Pack for Oracle Database
- Enterprise Extreme Performance—Active Data Guard, In-Memory Database, and RAC

Special consideration should be taken when selecting "License Type" because the price difference is significant. For example, using "Pay as You Go" pricing for a single Extreme Performance VM, the Bring Your Own License (BYOL) cost is approximately $0.29 per CPU per hour versus the included license cost of $2.52 per CPU per hour. Including the license makes the instance almost nine times more expensive.

The "Available Storage Size" option is a little misleading because it only accounts for a portion of what Oracle will actually charge for storage for this VM. The "Total Storage Size" reflects the total storage consumed by the VM. The available storage size is the size of the +DATA ASM disk group. The +RECO disk group will also be created with 256GB of space. The storage allocated to the OS makes up the rest. Selecting 256GB will result in a total storage allocation of 712GB. Selecting a larger storage size only affects the size of the +DATA disk group.

The available database versions as of February 2019 are 18.0, 12.2.0.1, 12.1.0.2, and 11.2.0.4. With the exception of version 11.2, the created database will be a multitenant database, regardless of whether you choose to create a pluggable database (PDB) during the provisioning process.

The Path Forward

Navigating Oracle Cloud Infrastructure Database Services can be a little overwhelming at first. However, creating Compute and Database Instances in OCI is relatively easy as Oracle has limited many of the options available when configuring an Oracle Database system. Some of the terminology may be unfamiliar but it should not be difficult to master for a seasoned Oracle DBA transitioning to a cloud database architect.

Managing OCI Database Instances once they are created poses additional challenges. The next article in this series will examine database management and administration tasks such as backups, pluggable databases, Data Guard, and client networking. ■

TODD SCHRAML

Todd Schraml has more than 20 years of IT management, project development, business analysis, and database design experience across many industries from telecommunications to healthcare. He can be reached at TWSchraml@gmail.com.

Fact Tables and Their Lying Primary Keys

When working on a multidimensional design, every fact table within scope should be handled with care. In an ideal world, each low-level fact table represents the metrics related to a business event. The meaning of a fact table, ideally, should be evident based on the table name and the composition of the fact table's primary key. Deciding on a primary key for a fact table is an important choice. Some designers simply create a surrogate key to play the role of a primary key. The surrogate key, when used as a primary key, ensures that no semantics could ever be derived from a simple review of the table's design. The surrogate-keyed fact table has no specific meaning at all, as even 100% duplicated data can be added with impunity to the fact as long as inserts generate a new surrogate value. The logic within the code that determines when to insert is the actual home of the table's meaning. Hopefully, that table-insert logic is documented elsewhere as well.

When a data architect chooses not to use a specially generated surrogate key, the usual and expected fact table primary key is a composite of foreign keys pointing to the dimensions associated with a fact. Under this composite primary key approach, the combination of dimensions describes the grain of the given fact table. A row cannot physically be inserted that is at any lower grain than that of the combined values from all those dimensions. However, it is not unusual to find dimensions associated with a fact table that are *not* contributing to the grain of the fact. When a designer knows that a dimension is not impacting a fact table's grain, the designer should not include the foreign key for that dimension as part of the primary key, as doing so may mislead people about the nature of the fact. This non-contribution-to-the-grain condition most often arises when dimensions are correlated. Data modelers should stop and give some consideration to why they are including these dimensions. A worst-case scenario might be that the designer has a hierarchy that has not been included within another dimension. It could be these hierarchy-component dimensions were added directly to the fact as a work-around to avoid snowflaking. The correction would be to expand the appropriate dimension to include the additional levels of the correlated hierarchy. Even so, some circumstances may still cause the design to include dimensions not contributing to the fact table's grain. If too many of these kinds of relationships exist, one may have a centipede fact with an unusually large number of dimensional relationships. Designers should seek a balance to avoid an excess.

When only a sub-set of the foreign keys comprise the full primary key of the fact table, it is not unusual for developers, or query writers, to gloss over the documentation and assume that *all* foreign keys are the primary key. Functionally, they have assumed a non-intended super key. This may not cause them to fall down a wrong rabbit hole all the time. After all, if [A] is a unique set of values, then [A, B] is also a unique set of values. But in some circumstances, unnecessary work might be done by these over-achieving query writers. Some database designers may feel that it's OK to arbitrarily assign all dimensional foreign keys as part of a fact's primary key even when one or more do not actually contribute to the fact table's granularity. Such arbitrary labels are only exposing the designer's sloppy approach to their models. Sloppy approaches are often found in weak designs that likely contain flaws. In such cases, these super-keyed facts are lying to users about the grain of the fact table. ■

APRIL/MAY 2019 Ad Index

Kore Technologies 25
Melissa . . . . . Cover 2
Revelation Software . . . . . Cover 4
Wisconsin Madison 23

Best Practices Sponsors

Actian 12
Denodo 13
SlamData 14
Wipro 15

Cognitive Computing & AI Summit

MAY 21-22 2019

HYATT REGENCY BOSTON | BOSTON, MA

A featured event at

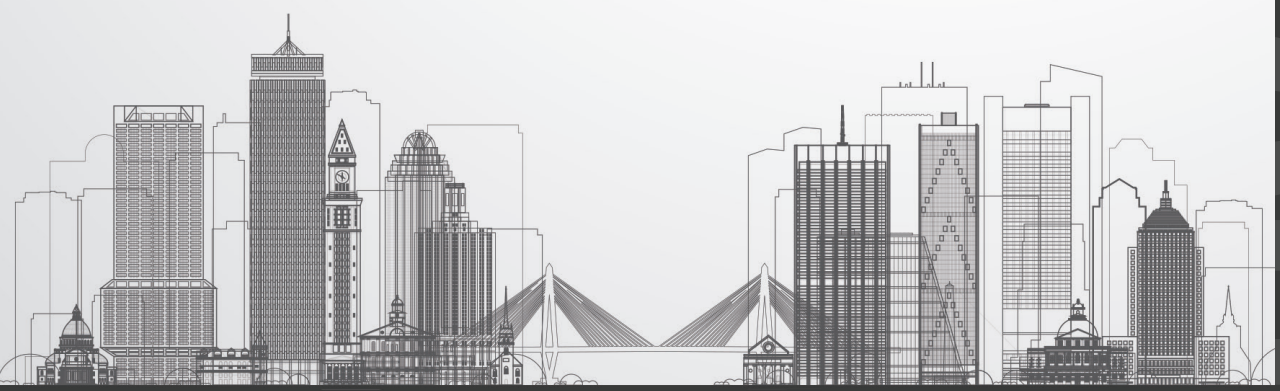
DATA SUMMIT
UNLEASH THE POWER OF YOUR DATA

REGISTER TODAY!

Use code DBT19 to save $100!

A new era of cognitive computing has already begun, and its impact is being felt across industries, from healthcare and financial services to manufacturing and education. However, building cognitive systems and applications that can perform specific, humanlike tasks in an intelligent way is far from easy. The Cognitive Computing & AI Summit is an intense, 2-day immersion into the leading cognitive computing and AI use cases, strategies, and technologies that every organization should know about. Whether you are a developer, engineer, executive, entrepreneur, or product manager, if you are on the front lines of AI and cognitive computing, this summit is for you. **Reserve your seat today to join your peers in Boston!**

dbta.com/cognitivecomputingsummit



REGISTER TODAY!