

Volume 7 Number 1 ■ SPRING 2021

# BDOQ

BIG DATA QUARTERLY

Getting More Data  
Value with DataOps

The \$100-Trillion  
Opportunity for IoT

Move Over Unicorn,  
It's the Rock Star's Time

## Enterprises Learn to Unravel the Complexities of Multi- Cloud

WWW.DBTA.COM

# DATA SUMMIT CONNECT

[dbta.com/datasummit](https://dbta.com/datasummit)

**MAY 10–12, 2021**

Dive into the latest strategies and technologies in data management and analytics this spring at *Data Summit Connect 2021*, a virtual event that will run May 11–12. We'll also be hosting a series of in-depth preconference workshops on May 10.


From modern data architecture and cloud migration to DataOps and analytics, *Data Summit Connect* will equip you with the technical know-how, practical advice, and expert insights to succeed in this evolving space. You'll hear about innovative approaches that the world's leading companies are taking to solve today's key challenges and emerging technologies, revolutionizing how data is stored, protected, integrated, enhanced, and acted upon.

Come see where the leading businesses, government agencies, and public institutions go to find the right strategies and solutions to become insights-driven enterprises.

**REGISTRATION IS OPEN!**

**VIRTUAL  
EVENT!**

**DON'T MISS**

**AI &  MACHINE  
LEARNING** SUMMIT

**DATAOPS  
BOOT CAMP**

**DATABASE  
DEVOPS  
BOOT CAMP**

**Stay Connected**



**#DataSummit**

DIAMOND  
SPONSOR



PLATINUM  
SPONSORS



LICENSEFORTRESS®



MariaDB

GOLD  
SPONSOR



LEAD WITH DATA



TigerGraph

vmware®



MEDIA SPONSORS



ORGANIZED AND PRODUCED BY



Information Today, Inc.

FROM THE PUBLISHERS OF



# BDOQ

## BIG DATA QUARTERLY

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

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

**CORPORATE HEADQUARTERS** 143 Old Marlton Pike, Medford, NJ 08055

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

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

Joseph McKendrick,  
Contributing Editor; Joseph@dbta.com

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

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

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

Lauree Padgett,  
Editorial Services

Tiffany Chamenko,  
Production Manager

Erica Pannella,  
Senior Graphic Designer

Jackie Crawford,  
Ad Trafficking Coordinator

Sheila Willison, Marketing Manager,  
Events and Circulation  
858-278-2223; sheila@infotoday.com

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

### ADVERTISING

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

### INFORMATION TODAY, INC. EXECUTIVE MANAGEMENT

Thomas H. Hogan, President and CEO

Roger R. Bilboul,  
Chairman of the Board

Mike Flaherty, CFO

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

Bill Spence, Vice President,  
Information Technology

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

### POSTMASTER

Send all address changes to:

*Big Data Quarterly*, 143 Old Marlton Pike, Medford, NJ 08055

Copyright 2021, Information Today, Inc. All rights reserved.

### PRINTED IN THE UNITED STATES OF AMERICA

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

No part of this magazine may be reproduced and by any means—print, electronic or any other—without written permission of the publisher.

### COPYRIGHT INFORMATION

Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by Information Today, Inc., provided that the base fee of US \$2.00 per page is paid directly to Copyright Clearance Center (CCC), 222 Rosewood Drive, Danvers, MA 01923, phone 978-750-8400, fax 978-750-4744, USA. For those organizations that have been granted a photocopy license by CCC, a separate system of payment has been arranged. Photocopies for academic use: Persons desiring to make academic course packs with articles from this journal should contact the Copyright Clearance Center to request authorization through CCC's Academic Permissions Service (APS), subject to the conditions thereof. Same CCC address as above. Be sure to reference APS.

Creation of derivative works, such as informative abstracts, unless agreed to in writing by the copyright owner, is forbidden.

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

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

### SUBSCRIPTION INFORMATION

Subscriptions to *Big Data Quarterly* are available at the following rates (per year):

Subscribers in the U.S. —\$97.95; Single issue price: \$25

editor's note | *Joyce Wells*

## 2 The Intersection of Data Democratization and Security

### departments

## 3 TRENDING NOW | *By Joe McKendrick*

### Data Still Not Aligned With Digital Transformation

## 6 INSIGHTS | *By Kendall Clark*

### Semantic Graphs and the New Data Integration Landscape

## 25 INSIGHTS | *By Tim VanTassel*

### Information and Wisdom: The Art of Building and Operationalizing an Analytic Model

### features

## 4 THE VOICE OF BIG DATA

### Getting More Data Value with DataOps: Q&A with DataKitchen's Chris Bergh

## 8 FEATURE ARTICLE | *Joe McKendrick*

### Enterprises Learn to Unravel the Complexities of Multi-Cloud

## 24 BIG DATA BY THE NUMBERS

### Hybrid and Multi-Cloud Deployments Expand

### columns

## 27 DATA SCIENCE PLAYBOOK | *Jim Scott*

### Insider Threat Detection With Accelerated Machine Learning

## 28 DATA DIRECTIONS | *Michael Corey & Don Sullivan*

### Denial of Service Attacks Can Come Directly From Silicon Valley

## 30 THE DATA-ENABLED ORGANIZATION | *Lindy Ryan*

### Move Over Unicorn, It's the Rock Star's Time

## 31 THE IoT INSIDER | *Bart Schouw*

### The \$100-Trillion Opportunity for IoT

## 32 GOVERNING GUIDELINES | *Kimberly Nevala*

### Rethinking the Case for Governance

# The Intersection of Data Democratization and Security

By Joyce Wells

FASTER DECISION MAKING ENABLED BY ACCESS TO role-appropriate information is the goal of organizations striving to become data-driven. At the same time, there is strong pressure on companies to ensure data quality and trustworthiness, as well as to maintain data security to avoid breaches and risk regulatory non-compliance.

The requirements for widespread data availability and the equally important issues surrounding data security and governance are addressed from a number of standpoints in this issue of *Big Data Quarterly*.

Data may be at the center of all digital engagements, but today, data use is too difficult, and decision makers often have limited insight into data-related problems or the benefits that can be accrued from investments, according to *BDQ* contributing editor Joe McKendrick who highlights new research from BARQ. The research notes that data producers need to understand the needs of data consumers, but, at the same time, data consumers must understand the requirements and restrictions of data production processes.

McKendrick also looks at the growth of multi-cloud deployments and the associated tools, technologies, and risks in his cover article. Using a variety of cloud platforms is a great way to ensure more services, but it takes some skill to get everything aligned, he notes.

Ways to improve data use are also considered from a range of perspectives by additional industry leaders. FICO's Tim VanTassel looks at the challenges posed by the democratization of analytics and then provides a practitioner's guide to analytic model development. Machine learning is useful, he emphasizes, but a combination of information, explainability, and wisdom is critical. In addition, data integration systems leverage semantic graphs and data virtualization to represent connectedness and unlock business value, observes Stardog's Kendall Clark in an article on the new integration landscape. "Thanks in part to the pandemic, we realize more than ever that connected networks are everywhere and the data systems and data silos must be united as a result."

Now is the time to shift the focus to people and processes, emphasizes DataKitchen's Chris Bergh in an interview about the rise of DataOps. "You really have to make the team, with all the tools and all the data, work together." The use of data and analytics is a team sport, not the result of individual acts of heroism, he states.

And in this issue, Radiant Advisors' Lindy Ryan debuts her column series, "The Data-Enabled Organization," making the case for greater recognition of the new data rock star: the business-oriented user. "An uneducated data user is a dangerous data user," she observes. Collaboration enables the full value of digital transformation, adds Software AG's Bart Schouw. Thinking of digital transformation as simply automating and integrating established interactions is not sufficient, he stresses. Real digital transformation turns your infrastructure into an infostructure.

Along with articles on how companies can better leverage their data through shifts in process and culture, the intertwined requirements for security and governance are also tackled in this issue.

LicenseFortress' Michael Corey and VMware's Don Sullivan look at the current data threat landscape and what companies should take into account, while NVIDIA's Jim Scott writes about improved processes for insider threat detection with machine learning.

And finally, SAS' Kimberly Nevala makes the point that it's time to change how we think about data governance. Too often, it is thought of as a means to impose restrictions to achieve compliance or eliminate risk, but it's important to remember that it improves the cadence and quality for organizations' collective and individual decisions—whether as a simple BI dashboard or a complicated AI algorithm.

We'll continue to explore all these topics and more during the upcoming Data Summit conference, which will be held May 10–12, 2021. Overcoming travel and geography challenges, Data Summit Connect 2021 will be held virtually. For more information, go to [www.dbta.com/Conferences/2021](http://www.dbta.com/Conferences/2021).



# Data Still Not Aligned With Digital Transformation

By Joe McKendrick

DATA MAY BE AT THE HEART OF ALL DIGITAL ENGAGEMENTS, but most enterprises are still behind the curve when it comes to effectively identifying and managing it. That's the takeaway from the latest survey of 419 enterprise executives from BARC, which finds continuing challenges with identifying and surfacing the data assets needed to succeed in today's digital economy.

While most organizations intend to digitally transform themselves, few are taking action at this time, the survey found. Ninety percent, for example, agreed that information has a high priority in enterprise decision making. But only 25% stated that decisions are predominantly based on data at this time.

"In principle, everyone agrees that data is important, and its targeted use can make a decisive contribution to improved company results," the survey's authors said. "But the fact is that data use is far too difficult today. Investing in improvements is not usually a real priority. Decision makers in particular have little insight into their data-related problems and the benefits of potential investment."

The challenge is that convincing decision makers to invest in data is a chicken-or-egg issues, the survey's authors added. Close to two-thirds of respondents, 65%, agree that the value of data is not sufficiently transparent, but only 23% believe that creating more transparency in this area is an important approach to improving the handling of data.

The survey report identified "best-in-class" companies that are functioning as data-driven enterprises. These leaders "have already created transparency about the value of data and what can be drawn from it. They have thus created the basis for convincing decision makers to invest."

## The Value of a Data Catalog

A data catalog will help meet the requirements of an enterprise seeking to run on data analytics, but this demands buy-in from business users. Tellingly, 60% of companies state they "waste a lot of time" asking the same questions about data or repeating work. The top three approaches to improve the handling of data include providing more information about data (59%), defining clear responsibilities (57%), and providing a business glossary (56%). "Data catalogs help meet these needs," the survey's authors pointed out. This type of technology is in use or planned by 72% of the companies in the survey. A majority of the "best-in-class" companies identified in the survey, 57%, already have a data catalog in use.

Data democratization is also on the table for many enterprises. A majority of respondents, 74%, stated that they already analyze a lot of data, "but conditions are not in place to use this knowledge in real-time processes." In addition, 58% said their data governance processes are still too immature to deliver data analytics in a widespread way. True data democratization requires a "new deal" on how data is handled across the enterprise, according to the survey's authors. "Data producers need to understand and take into account which data-related needs data consumers have. At the same time, data consumers must understand the requirements and restrictions of data production processes. Enterprises need a new deal between data producers and data consumers that effectively addresses the top three challenges to improving data handling—time spent, a lack of transparency of data value, and insufficient data quality."

## Automated Data Management

Best-in-class adoption of automated data management using machine learning offers a potential benefit. Thirty percent of leading companies have already taken this approach, and 43% are planning to do so.

Data quality is another issue that continues to hamper digital transformation efforts. "Insufficient data quality drives the need for individual data preparation, inevitably leads to an inflation of data silos, and undermines any governance efforts." Seventy-two percent of respondents agree that business users lack the time to develop new ways to use data, and 62% agree that business users lack the competence and skills to work with data.

"Enabling a data-driven enterprise requires a fundamental cultural change driven by the executive level," the report authors stated. "Technology is an enabler but not the driver for data-driven working. Individuals adapt to the corporate system. Corporate culture and organization must therefore be realigned. In this respect, the widely adopted bottom-up approaches to digital transformation are very limited in their impact. Measures such as establishing clear responsibilities for data in the line of business, investing in data literacy by carrying out targeted staff development and training, and developing the corporate data culture from 'need to know' to 'right to know' require strategic orientation and active support by the executive level. You will also need a cross-functional team of mid-level directors and managers who have a vested interest in becoming a data-driven organization."

*Data quality is an issue that continues to hamper digital transformation efforts.*



**Joe McKendrick** is a contributing editor and writer to *Database Trends and Applications* and *Big Data Quarterly* magazines, as well as lead research analyst for *Unisphere Research* at *Information Today, Inc.*

# THE VOICE OF BIG DATA

## GETTING MORE DATA VALUE WITH DATAOps

*DATAOps IS SEEN AS A KEY APPROACH FOR SUPPORTING INSIGHT-DRIVEN CULTURE AT ORGANIZATIONS SEEKING TO EXTRACT MORE VALUE FROM THEIR DATA.*

*RECENTLY, CHRIS BERGH, CEO AND FOUNDER OF DATAKitchen, A VENDOR THAT PROVIDES AN END-TO-END DATAOps SOFTWARE PLATFORM AS WELL AS ADVISORY SERVICES FOCUSED ON DATAOps TRANSFORMATION, TALKED TO BDQ ABOUT WHAT DATAOps IS, HOW THE METHODOLOGY HAS EVOLVED, AND HOW CUSTOMERS CAN USE IT TO GAIN GREATER BENEFIT FROM THEIR DATA.*

### **With the emphasis on using data more effectively, is DataOps becoming more widely understood?**

More people are learning about it and realizing that it's important, but it's certainly not a standard yet. It isn't about a new technology; it's not about "Let's buy a faster database" or "Let's buy that cool new visualization tool." They're starting to realize that, after all this investment in new databases; big data and streaming data; and IoT, AI, ML, and all the acronyms out there, they've got all the technology that they need. It's more about the people and the process that matters—the people who use that technology—and the journey to be data-driven is getting to its natural point of maturity.

### **How so?**

You really have to make the team, with all the tools and all the data, work together. I grew up in Wisconsin, and when I was in high school, American Motors, which made the Pacer, went out of business. My dad worked for Wisconsin Bell, and he drove a Toyota Corolla, and people didn't like that because he was in a union and it was an American car versus a Japanese one, but my dad said, "It's cheaper, it's better, and it lasts longer." And so why did Toyota make better cars than American Motors? You have to work on the people, and on the process that people work in. The journey to be data-driven is less about data and more about the people and the teams who are creating value from the data.

### **You have said that what you do is less important than how you do it. What does that mean?**

It reflects my own journey and starting to do data and analytics 15 years ago. I always look at it from the lens of a leader and a manager: How do you lead your people? What is the organizing principle to make it work? And how do you focus on the value that your customers are really receiving instead of getting into



**Chris Bergh, CEO and Founder, DataKitchen**

the trap of building something and expecting that a year later, wonderful things are going to happen? That has to be the shining light. In order to be data-driven, you have to be of service to the people who need the data, help them deliver value to them, and then work to improve upon it.

### **Descriptions of DataOps methodology vary by organizations. What is your definition?**

I think it is a set of technical practices and cultural norms for how people work to get value out of data. It is a technology and in some ways it's technical practices or architectural practices that go hand-in-hand with how people work. There is a philosophy in DataOps that is about iterative development and delivering things short and focusing on your customer. And those come from agile processes and even from lean manufacturing ideas. And then there is a technical part to it because it's really hard to do these things when you have constellations of tools running in different systems. Those technical practices and the cultural norms are intertwined.

### **How does someone start with DataOps and what does DataKitchen provide?**

Typically, it's a senior person in a large company who is saying, "Hey, we want to be data-driven, but it's not working." And they examine why after they have bought all the great tools, and they realize that they've got to work on the people and their process. And they stumble upon DataOps as a sort of philosophy, and then they bring us in. First, they want us to help them transform their team and lay out a set of steps for how they can go from zero to DataOps, and then they bring us in as a technology provider to help them create a central place where they can plug all the technology they have—their ETL tool, their data tool, all their people in all their different



---

*Insight doesn't just come from hiring a bunch of data scientists; it comes from building a team and a process that continuously creates value.*

---

locations—and that can provide a central point of kind of coordination for all those steps that are involved in creating value, while not replacing what they currently have.

### **Does that make DataKitchen both a technology provider and a consultancy?**

Yes, we offer advisory services and other assistance, as well as software. We realized that since this problem is both a technology problem and a people problem, you've got to work with both. We also have partners that are working to come up-to-speed and do this DataOps transformation work that we do, but right now we're doing it because people need the help.

### **It is frequently said that data scientists and analysts spend about 80% of their time cleaning and prepping data and only 20% analyzing it. Does DataOps address that problem?**

Looking at the 80% problem, the issue is that they are doing things other than what they really want to do. Whether you call them data scientists or data analysts, they want to create and invent, and sit between the data and the customer to figure out if it really helps with what their VP of marketing or their VP of ecommerce needs to know. And that's a really interesting, creative job, but they're spending a lot of time not doing things that they think are part of that role, and a piece of that is preparing the data.

### **What are the other challenges?**

The other part is just when they want to do something new. In a lot of organizations, it takes months for them to move from the time that they have an idea to getting it in their customers' hands because of all the meetings and coordination. Data and analytics are not individual acts of heroism; it's a team sport.

### **You are saying the tools are not the problem.**

Most people have the tools that they need to do their data work. In their tool chest, they have a tool that does data transformation or ETL, they have a database, a tool to do charts and graphs. They need those and they need the data that they can act upon. And, of

course, they have a customer who they're trying to understand, but, in addition to that, they need to start thinking about the tools to help them with the process. And that's automated testing, monitoring, and deployment, and managing the technical environments they do development and production in. Those are the sort of facilities that the DataOps market is providing.

### **Are there specific companies or market sectors where DataOps is being deployed most heavily?**

Yes, there are companies that are doing it in Silicon Valley that just have a lot of data and they've always worked in a very iterative, customer-focused way. And then second, there are companies now that want to be like that and feel threatened by the Amazons of the world, and they want to be more data-driven. Those are typically companies in financial services, healthcare, and manufacturing that have a lot of data that they can't get value from, so they've hired a chief data officer and they're trying to be data-driven.

### **Are there any roadblocks to DataOps that you see right now that are preventing people from putting their data to work?**

For people who are experienced in data and analytics, we're saying to them, "You can make changes to your systems fast, and you can do that with very minimal problems." And that is, for some people, almost heresy because they just got [their system] working and they don't want to change it. They are thinking, "It's running, don't tell me I can change it every week, or change it every day."

People have spent a lot of time building very brittle systems that require a lot of care and feeding, and they've done the best they could, but this change in focus to how you do it—as opposed to what you do—is my motivation. I suffered for years not working this way. And it is a challenge for people to see that this different mindset can help them and is not a threat to them; it's actually a way to alleviate their suffering.

### **Why is now the time to jump in to DataOps?**

In my experience running analytics teams, I just got phone calls when things went wrong or were going too slow. I'm seeing that same pattern happen again, and people are going through the same experience I had. The time is now because organizations have invested in data and analytics systems, and they want more change and more insight, but that insight just doesn't come from hiring a bunch of data scientists; it comes from building a team and a process that continuously creates value. I think most people want to build Toyota Corollas and not AMC Pacers.

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



*Semantic graphs are the only way to represent data that is natively stored in other structures while maintaining all relevant metadata and context.*



**Kendall Clark** is founder and CEO of Stardog ([www.stardog.com](http://www.stardog.com)), an enterprise knowledge graph (EKG) platform provider.

# Semantic Graphs and the New Data Integration Landscape

By Kendall Clark

Conventional data management systems are fundamentally ill-suited for the world of data as it exists today. These systems, based with few exceptions on the relational data model, are broken because they integrate based on data location at the storage layer. While this approach worked reasonably well for the past 25 years, the world today has far too much data to use data location in storage as the basic lever.

The ill-suitedness of traditional, relational data model-based data integration tools reveals itself in several ways. The most obvious difficulties occur when combining several data silos or sources together because, in nearly all cases, they were modeled differently and conform to their own independent sets of rules and constraints. Data integration breaks down for two reasons. First, a single shared data model has to represent a global view over the sources. Second, significant manipulation and transformation are typically required to transform between the source and target schema as well as make the source data conform to a set of standardized rules.

All of this manipulation, cleaning, and transforming is necessary because relational systems aren't very good at representing contextualized business meaning, and the fact that the relational model itself, as distinct from its dominant query language, SQL, is a leaky abstraction that does not lend itself very well to integration, especially for connection-rich data. The relational data model was never intended to support the complex business processes with changing requirements that, in today's data landscape, are dominated by heterogeneity and diversity.

## Older Integration Styles Are Falling Short

The data landscape is not only more heterogeneous than it used to be, but it has also expanded dramatically. When the relational model and SQL were being developed, semi-structured and unstructured data simply didn't count. Emails, social network data, and IoT were all either to be invented or weren't part of the enterprise world. In other words, relational data management systems worked reasonably well when the enterprise data landscape was itself predominantly structured—but not anymore. The enterprise data landscape is increasingly hybrid, varied, and changing.

In fact, the challenges to conventional data integration are proliferating. The emergence of IoT, the rise

in unstructured data volume, increasing relevance of external data sources, and the headlong rush to hybrid, multi-cloud environments are all impediments to wide-scale data integration based on data location in storage with a relational data model. So, while the data landscape itself has changed, what about the enterprise's requirements for data integration and analytics systems? Surely, if those requirements are relatively unchanged, then there must be some hope left for relational systems, right?

The truth is, these requirements have changed, along with the data landscape itself, creating two relevant pressures. First, the impact of globalization and an ever-shrinking world has created unprecedented awareness of the connectedness of the human world and, of course, of the modern enterprise. Thanks in part to the pandemic, we realize more than ever that connected networks are everywhere and the data systems and data silos must be united as a result. As the name suggests, enterprise data is largely about the enterprise itself. The connected enterprise, conversely, deals in business meaning and context, which may be why they seem to be winning everywhere we look. The second pressure has been created by the rise of AI, ML, and the various analytics systems. These are nothing more than incredibly intricate, powerful machines which run on data as their essential fuel. No data, no insight—and that's true no matter how clever the algorithm, AI team, learned model, etc.

## The Future of Enterprise Data Integration

To create business value within the enterprise, an organization must be able to connect all the data that matters. Yet because of changes in the data landscape, this information is literally spread everywhere and includes numerous formats, types, storage systems, applications, and computing environments. Worse, it resides everywhere, whether internally or externally via public or private cloud. Of course, some of this data exists as relational tables, but more and more of it will exist in some other form in the future, which may not be amenable to being represented by or with relational tables.

The real problem with conventional data integration systems based on the relational model comes down to representation. The new diversity of requirements and data landscape has finally burst the long-running



illusion that you can just jam non-relational information/assets into relational storage and it will all be fine. The NoSQL movement of data storage systems has already reckoned with that reality, leading to another proliferation, this time in the area of database systems.

However, the data integration space is still completely dominated by relational-first and relational-only systems that are stuck with the fundamental idea of data location in the storage layer. Given the current representational problem, many companies now understand that only the semantic graph data model is able to represent data that is natively stored in any other structures and to connect all relevant metadata and context.

Semantic graphs create meaning by mapping entities, their metadata, and their relationships in an evolving information network. By applying a fundamentally different approach to data integration, organizations can substitute the idea of business meaning in the compute layer for data location in the storage layer. In this way, next-wave data integration systems will leverage semantic graphs and data virtualization technology to represent the connectedness of data in a way that unlocks business value by shortening the gap between what data means and how its managed, queried, searched, analyzed, and connected.

Consider this actual use case as an example. Dow Jones is using a semantic graph as its data integration platform to connect multiple data silos and using advanced AI techniques to “reimagine the news.” With access to millions of facts derived from 50 years of digitized news media, Dow Jones is a great example of how the data landscape has changed. Internally, it also demonstrates how it has a unique data universe that is naturally unconnected. The business environment floods Dow Jones’s customers with endless noise, but there is a signal within the noise, and Dow Jones has built a personalized news sense-making application that focuses on what customers need to know and when.

## Representational Power Unlocks Reusability

Relational models also create missed opportunities because the relational model is a leaky abstraction, and it’s very unusual to see relational schemas or data models reused across many different applications. In fact, it only seems to ever really happen when reusing actual physical relational databases or the equivalent, including materialized views.

Yet, nowadays, multiple schemas, or data models, are required to manage an enterprise. Which is just to say that there are many different perspectives within an enterprise and there is no hope for one data model that will satisfy all those various perspectives, use cases, and requirements. This is true because a data model establishes the meaning of the data as well as the relationship of various entities to one another. At enterprise scale, it is inevitable that different business units define terms differently; it may even be required by regulation or law. And, with the increasing relevance of third-party data, it’s simply impossible to impose one definition upon all data producers or consumers.

Given the limits of conventional data integration, the typical way of overcoming this issue is to copy data for each new use case, creating a new and distinct data model in the process. Even with all the advances in IT over the past 20 years, the most common data integration technique is still batch or bulk copying of data. However, this practice leads to a proliferation of data within an organization, degrading its quality and causing uncertainty over which copy is the source of truth. Then, when faced with a new project that requires making existing applications speak to one another, effort is wasted on patchworks of otherwise unnecessary efforts to “reintegrate,” that is, to undo the copying that’s been going on, in lieu of an actual solution, to try to work back upstream to something like a source of ground truth.

It’s important to understand that the disease is actually caused by the purported cure; or, if that seems a bit extreme, the so-called solution is at a minimum making the disease worse instead of better. All of this storage-level reintegration leads to the enterprise being

very slow to respond to emerging threats, crises, and opportunities. When unanticipated questions or needs arise, work grinds to a halt as the data preparation starts anew. This reactive data strategy leaves teams flat-footed when the market shifts or new questions arise. As such, enterprises require a more responsive data strategy, one that keeps pace with the needs of the business itself.

---

*The real problem with conventional data integration systems based on the relational model comes down to representation.*

---

## Support for Data Reuse

Data integration based on a semantic graph can end the cycle of copies of copies of copies within the enterprise because it’s able to represent business meaning at a level of abstraction beyond the storage layer. Additionally, this representational power leads directly to reuse of data rather than copying copies of data. Reusing both data and data models means that each new application, or response to a new crisis, requires less time and energy because reusing previous work builds value incrementally over time. New project timelines plummet. Enterprise responsiveness increases.

In fact, the largest information integration projects on the planet already use this semantic graph model. Consider your web browser, for instance. The web itself contains a world of information, created by different contributors, and accessible through a single browser. Google Search, which includes a network of 500 billion facts about 5 billion entities, uses a knowledge graph as well. Both Google and the web are great examples of this large-scale, complex, and decentralized information integration style as it delivers information based on its meaning and relationships.

Traditional data integration based on the relational model originally arose in response to the creation of storage and database systems based on that same model. Today, integration must follow the intersection of the needs of the enterprise and the nature of the data itself. Those needs and the nature of data integration have changed dramatically, and it’s good to see that modern integration systems such as semantic graph technology are stepping up to capture the real-world context of data, regardless of where it resides and how.

By  
Joe McKendrick

# Enterprises Learn to Unravel the Complexities of Multi-Cloud

**MULTI-CLOUD IS CHANGING THE WAY** we manage data. It's seen as a way to build a more resilient diversity of services while ensuring a greater degree of independence from a single vendor. At the same time, it takes skill to get everything aligned. In recent years, multi-

cloud has become a popular approach, with 93% of enterprises using a multi-cloud strategy, according to the latest Flexera/RightScale survey on cloud adoption. The survey found that respondents use an average of 2.2 public and 2.2 private clouds.

It's important for data managers and their business counterparts to understand the benefits and implications of the cloud implementations they are using to run their businesses. "Enterprises don't choose multi-cloud," said Sriramkumar Kumaresan, head of service lines markets at Mindtree. "Multi-cloud chooses an enterprise as a result of keeping up with dynamic and multiple dimensional business transformations. An example is mergers and acquisitions. The advantage of a multi-cloud environment is ensuring availability of the right applications at the right place, at the right time, with the right speed and at the right cost."

## THE MULTI-CLOUD VALUE PROPOSITION

What is driving the appeal of multi-cloud arrangements? Avoiding vendor lock-in is a frequently cited advantage to multi-cloud strategies, said Matt Quinn, chief operating officer of TIBCO. "It allows for different architectural and technical benefits that wouldn't be available if the organization uses a single cloud vendor. Multi-cloud arrangements allow organizations to invest in different vendors for different purposes,

encourage architects to think about core use cases and make sure services are portable, and open up the possibility for new services that a cloud of choice may introduce in the future."

Embracing a multi-cloud strategy "enables true freedom and control to run an application, workload, or data on any cloud," agreed Paul Speciale, chief product officer at Scalify. Modern enterprises have sophisticated needs that can no longer be satisfied by one cloud provider, he added. A multi-cloud strategy ensures a higher level of data availability and durability because data can be replicated across multiple, fully autonomous clouds. Another benefit is cost savings, due to greater competition between cloud storage providers and the ability to instantly take advantage of new pricing for any given storage offering, added Speciale.

Increasingly, "the leading driver for a multi-cloud strategy is maintaining flexibility and choice of providers; enterprises want to avoid vendor lock-in and take advantage of best-of-breed solutions," said Vikas Mathur, senior vice president of products at Actian.

What does it mean to be multi-cloud? "Some think it is running an application across multiple cloud providers so that they can capture cost benefits or reduce risks," said Jim Walker, vice president of product marketing at Cockroach Labs. "It is pretty difficult to pull this off though, as technical complexities are pretty intense. Hybrid cloud for many is a reality and can be considered multi-cloud as well. Often in highly regulated industries or with high-value workloads, organizations will want better control."

There are availability concerns that make a multi-cloud strategy more preferable as well. "Not every cloud vendor offers services in every country or region, and having multiple cloud vendors gives customers more geographic coverage," said Quinn. Another way to think about multi-cloud is multiple regions, which can be complex but a very pragmatic reality for many, as they want failover or they need applications to live in different regions to meet latency requirements of their users, Walker said.

In the era of digital transformation, data is playing a more important role than ever,

said Rick Vanover, senior director of product strategy with Veeam. “A multi-cloud approach enables companies to take advantage of many benefits, including increased agility and cost management,” Vanover said. This also aligns to developer skillsets, since many organizations that have multiple lines of business or divisions will inherently have a multi-cloud strategy based on their different areas of expertise.

## APPLICATIONS

Applications that are typically deployed across multiple cloud environments tend to be data-intensive, said George Burns III, senior consultant for cloud operations at SPR. “More common multi-cloud implementations are with data science, business intelligence, and data warehousing engagements,” he said. “As high-growth sectors in recent years, these data-dependent areas have flourished by treating data location as an agnostic attribute, creating connections between platforms to use data in place without the need to migrate or copy data from its original location.”

According to David Tareen, director of AI and analytics at SAS, multi-cloud is a natural environment for analytics “due to the bursting and unpredictable nature of analytic workloads. Analytic applications are often deployed across multiple cloud environments to take advantage of scale. To eliminate data movement, we often see analytic applications being co-located with the data across multiple cloud environments.”

Multi-cloud also serves as a supportive environment for “application ecosystems that comprise multiple applications and/or tools,” said Mike Rulf, CTO of the Americas at Syntax. “For example, due to a geographically dispersed workforce, an organization may turn to virtual desktops as a way to ensure secure access to the corporate environment.” Once on that virtual desktop, a VPN or private connectivity can be used to access a corporate ERP system that could be on-premise or hosted with a private cloud provider, he explained. “In turn, that ERP system may tie to a SaaS solution for HR data and integrate with a data lake for dashboards and visualizations.”

Choosing multi-cloud application workloads requires understanding which deliver strategic and dynamic growth and prioritizing those for cloud deployment in a multi-cloud strategy, said Janine Corey, global head of cloud solution architecture for Qumulo. “The most critical step is to have a clear business value-based strategy and cross-functional leadership alignment around various workloads’ contribution to the overall business,” Corey said. The idea is to move to the cloud those workloads that benefit from scaling up and down dynamically. “Applications that drive compelling digital customer experiences and are core to a company’s brand, growth, and expansion strategy should be not only moved to the cloud but also modernized to take advantage of fast scale-up for unexpected big growth spurts, and the ability to be elastic for the best cost efficiency, offered by the cloud.”

## RISKS

Due to its highly distributed and diverse nature, multi-cloud adoption is not without its share of risks. For example, with the onset of the COVID-19 crisis, issues around automation, security, and visibility across multiple clouds quickly came to the fore, said Quinn. “As organizations expand to multiple clouds, their network is spread across additional providers and it results in blind spots that make it difficult to manage devices. This presents a huge risk to organizations, especially since the number of IoT devices has grown exponentially due to remote work.”

Having a multi-cloud strategy is similar to “having multiple data centers, each ... architected differently in terms of deployed tools, compute platform, storage platform, and networking, which increases administrative overhead,” said Rulf. “You will need to extend your management and monitoring tools to support the different APIs, native tooling, and each cloud platform.”

And “with multi-cloud strategies in place, organizations might experience complications like framework misalignment and differing security policies,” said Tareen. Vanover also warned that a “side

effect of today’s multi-location, multi-data center, multi-cloud strategy is hyperspawl, meaning that there is federation of data everywhere, across multiple clouds, databases, and devices.” Moreover, there are risks of “multi-cloud inefficiencies in some organizations, especially around purchase power, cloud overspend, and a potential increased security burden.”

In addition, pointed out Chander Damodaran, CTO of cloud engineering at Brillio, the “complex architecture required to address multi-cloud resilience is potentially fraught with higher costs, complicated delivery, and potentially longer time to market.” Another issue is the technical skills and resource availability needed to manage multi-cloud environments.

Observability is another requirement for multi-cloud sites, as enterprises will need to ensure performance and security from a number of vendors. “Without an observability platform in place, organizations attempting a multi-cloud approach are flying blind and can introduce a host of risks,” said Tareen. “It’s like turning off your radio before trying to land a jet at JFK. Make investments upfront to ensure you remain in control and have a clear, safe runway for your cloud applications.”

A multi-cloud deployment “will, most likely, require more operational effort to monitor and maintain,” said Burns. “Just as single-cloud deployments are handcuffed to changes on the platform they run on, multi-cloud deployments are subject to any changes to two or more different platforms. As a result, multi-cloud deployments face potentially exponentially increased workloads. Resource dependency between detached components can also become complicated, especially in data engagements where data sources are often being utilized across clouds. These connections are usually accomplished by using platform-specific connectors, which inherently tie down a multi-cloud deployment as well.”

## TOOLS AND TECHNOLOGIES

When it comes to managing data across multiple clouds, “it is essential to look for tools that will give the user access to all

# ENTERPRISES LEARN TO UNRAVEL THE COMPLEXITIES OF MULTI-CLOUD

platforms that host data—whether that’s private cloud, public cloud, or on-prem—providing one centralized API to manage in-band and out-of-band operations,” said Speciale. “The platform should be cloud-provider-independent, meaning users decide on the optimal cloud storage option for their data. The platform should act as an abstraction layer between data and the proprietary APIs used to place that data in any given cloud storage service. It should also be able to translate data into the native format of any given cloud storage service so that the data remains open and accessible to any system and not just the system that placed the data. This will prevent time-consuming and costly migrations and pave the way for a more flexible and efficient workflow to keep data more available and actionable.”

The rise of containers has also paved the way for multi-cloud adoption, industry observers said. The correlation between multi-cloud and containers is strong, since the ultimate goal of deploying software in containers is portability across environments such as on-premise (bare-metal servers) and cloud environments, said Speciale. “In the past, when we wrote monolithic applications for specific target environments—Windows, UNIX, Linux—there was really no hope of portability. Application developers would need to work hard to port the application to a new environment. Now that we have users, applications, and data being deployed using data center resources and the cloud, a central goal is to make applications intrinsically portable. If we really want to reap the benefits of multi-cloud, we absolutely will need the software to run anywhere, and that, in turn, is one of the key benefits of containers.”

Increasingly, enterprises “are betting on Kubernetes container technologies to drive their multi-cloud strategies and are deploying applications to easily orchestrate, manage, and scale while also retaining the flexibility to port applications across different clouds,” said Mathur.

This convergence is not without its issues, however. “Orchestration is necessary for any sort of cloud deployment at scale, whether multi-cloud or not,” said

Rulf. “That said, I have not seen container platforms easily deployed in a multi-cloud strategy. Most cloud platforms have unique feature or functionality built into their container platform that prohibits doing container deployments that span across cloud providers.”

To ensure greater transparency and control, Quinn advised putting a “cloud manager” in place to provide visibility across multiple cloud environments. “A good cloud manager combines data governance, data catalogs, and enterprise metadata management capabilities, providing customers visibility into their data landscape.”

**The rise of  
containers has  
paved the way  
for multi-cloud  
adoption.**

In addition, Rulf urged administrators to adopt robust monitoring and billing analysis tools to increase transparency of multi-cloud arrangements.

When enterprises embrace a multi-cloud strategy, “it becomes critical that data and applications are always available across all cloud types, including private, public, and hybrid clouds, to meet innovation and competitive demands,” Vanover said. That means enterprises must ensure availability is at the forefront of their digital transformation strategy, so that when applications and workloads are being moved across different infrastructures, there’s a backup and disaster recovery plan in place so that downtime does not become an issue.

## THE GREAT DECOUPLING

As enterprises move to multi-cloud environments, they need to understand which applications—or parts of applications—may be best-suited to deployments across multiple cloud vendors. Decoupling applications from underlying platforms may ensure the flexibility required. “When viewed through the lens of application modernization, new, modernized applications are really just collections of components deployed on top of a platform orchestration fabric,” said Burns. “These modern application components include containers, serverless functions, storage, relational databases, and non-relational databases.” Understanding which of an application’s components are spread across a hybrid environment is key to understanding cloud scale, Burns said. “It also points to why application decoupling will unlock the most potential from your cloud deployment.”

This changes the way applications operate as well, Burns continued. “You likely can’t span traditional monolithic applications across multiple service fabrics and still operate as expected,” he said. “The emergence of multi-cloud deployments has expanded the definition of cloud scale to include separately deployed components of an application. Applications deployed within your own data center, for example, can now leverage public cloud databases.”

Ultimately, flexibility and choice will encourage more competition, opening the door for innovative providers and solutions that are capable of running on the three major cloud platforms and incorporate private cloud and on-premise technologies as needed, said Mathur. More competition “will also encourage the dominant public cloud providers to double down on open interfaces, upgrades, and innovations to ensure their offerings integrate seamlessly with more platforms. The move toward multi-cloud adoption with an eye toward portability and compatibility across cloud platforms not only allows companies to select from best-of-breed solutions for improved business outcomes but also gives them more control over their cloud environments.”



**Qubole/WhereScape**

PAGE 14

A WAREHOUSE BY  
THE LAKE: HOW TO  
AUTOMATE THE IDEAL  
DATA INFRASTRUCTURE

**Qlik**

PAGE 16

THREE ESSENTIAL  
ATTRIBUTES TO  
MODERNIZE YOUR  
DATA REAL-ESTATE:  
A DATA WAREHOUSE,  
DATA LAKE, AND  
REAL-TIME STREAMING

**Ahana**

PAGE 17

THE RISE OF OPEN  
DATA LAKE ANALYTICS

**Matillion**

PAGE 18

A MODERN, CLOUD-  
NATIVE APPROACH  
TO ACCELERATING  
DATA INSIGHTS

**Action**

PAGE 19

CONVERGENCE OF  
DATA HUBS, LAKES,  
AND WAREHOUSES IS  
NEEDED TO UNIFY BI/AI  
TEAMS AND PROJECTS

**Pythian**

PAGE 20

ACCELERATING YOUR  
ANALYTICS JOURNEY

**Dremio**

PAGE 21

OPEN, NO-COPY DATA  
LAKE ARCHITECTURE

**Denodo**

PAGE 22

USING LOGICAL DATA  
FABRIC TO UNIFY  
DATA LAKES, DATA  
WAREHOUSES, AND  
CLOUDS FOR ANALYTICS

**BDOQ**  
BIG DATA QUARTERLY

# Modern Analytics: **DATA LAKES, DATA WAREHOUSES, AND CLOUDS**

Best Practices Series



# OPENING UP NEW FRONTIERS for DATA WAREHOUSES and DATA LAKES THROUGH the CLOUD

AT THE CORE OF EVERY DIGITAL TRANSFORMATION effort is data analytics. Of course, this requires data, and lots of it. That's why cloud computing is such a critical resource to build and sustain such efforts. Nearly 80% of data managers surveyed by *Database Trends and Applications* currently have digital transformation initiatives underway, and 55% of these organizations are deploying business intelligence or data analytics with cloud-based support ("DBTA Digital Transformation and Cloud Workloads Survey," 2021).

Cloud is a key foundation for these initiatives going forward. A separate survey conducted by Unisphere Research among PASS members found that 51% manage enterprise data in the cloud, and 16% planned to do so within a year's time ("DBAs Look to the Future: PASS Survey on Trends in Database Administration," January 2020). A majority, 52%, expected the volume of enterprise data in the cloud to grow significantly over the next 3 years.

The move to the cloud is about attaining the needed speed, agility, and insight to navigate today's rapidly changing digital economy. As a result, enterprises are embracing strategies to accelerate the movement of their data analytics capabilities to cloud platforms. Data warehouses and data lakes—once the exclusive domain of the world's larger enterprises—are now available and accessible to companies of all types and sizes.

For the incoming generation of data warehouses and data lakes, there is a growing array of choices when it comes to cloud platforms, deployment models, and features. At the same time, challenges remain. Data governance and security are still hot-button

issues. Real-time data requirements are on the rise at many organizations. And management and monitoring can be a concern whenever you add complexity onto an existing environment.

For organizations with growing data warehouses and lakes, the cloud offers almost unlimited capacity and processing power. However, transitioning existing data environments from on-premise systems to cloud platforms can be challenging.

Here are some key considerations for making the move:

**Plan ahead.** Undertaking a migration from an on-premise data environment to a cloud-based data warehouse or data lake is not an overnight process. It can be costly, both in terms of the budget and resources required. The organization may have deep investments in a data warehouse going back decades, and there need to be assurances that all data structures—and the ability to query them as before—remain intact. Of course, a cloud-based architecture calls for redesigning the data model.

**Define the business benefits.** Moving data analytics applications to cloud environments typically involves "greenfield" projects that have a fresh start, and are geared toward delivering superior customer or user experiences through mobile apps or highly responsive web-based interfaces. However, especially in the case of data warehouses, existing on-premise systems may have been built up over the years, reflecting millions of dollars of investments and time—and business users may perceive that the system is highly effective and delivering the needed results. In a 2019 survey of data managers conducted by Unisphere



## Best Practices Series

Research and the Independent Oracle Users Group (IOUG), 40% expressed concerns about the benefits being too small to justify the investment required to migrate.

There are many compelling business benefits for making the move to the cloud that go beyond simple upfront cost savings. There will no longer be constraints on the ability to pursue data-driven business ideas. The organization may be able to move more aggressively toward AI and more powerful vehicles of data analytics with the additional capacity available through cloud-borne data warehouses and data lakes. There will also be an enhanced ability to store data from various enterprise applications—ERP systems, financial systems, IoT-based systems—within a common environment. Cloud proponents will need to illustrate the vastly expended business pursuits that cloud environments support.

**Determine the best architecture.** There are several types of approaches to data warehouses and data lakes in the cloud. In the DBTA survey on cloud adoption, 59% of data managers said they are deploying applications and data to public cloud services, 55% are maintaining their data environments within private clouds, and 36% report using hybrid cloud arrangements.

A hybrid data warehouse or lake enables an organization to maintain data or applications on-premise while either gradually moving to the cloud or delineating functions that remain on-premise versus those that are cloud-based. Of course, this requires retaining skills for both cloud and on-premise systems. At the same time, it ensures greater resiliency, as well as greater flexibility in data

placement. A multi-cloud data lake or data warehouse leverages more than one platform while also requiring skills to integrate and manage two platforms that likely have differing protocols.

**Assess your skills and staffing requirements.** Within on-premise data warehouses and lakes, there is a need for database administrators as well as software engineers to build and scale such environments. While cloud reduces the need for such skills, it requires new skill sets, such as those of cloud engineers or cloud architects, as well as professionals with insights in cloud-based security and resource management. New roles need to be defined, and appropriate training provided. In the Unisphere-IOUG survey, one in five, 19%, cited skills availability as a challenge to moving to cloud environments.

**Establish ongoing data movement.** Getting a cloud data warehouse or lake up and running is only the beginning of the process. Data needs to be continuously synced and moved between sources and data environments. This ties into the overall performance of the data environment as it becomes increasingly cloud-borne. In the Unisphere-IOUG survey, 24% said concerns over maintaining the required level of performance in the public cloud was a challenge.

**Rethink security.** Moving to the cloud does not mean outsourcing security to a third-party provider. Security needs to remain a top priority for enterprises, regardless of how much data and how many applications are managed through cloud services. Due diligence is important, and enterprises and their data managers need to hold vendors closely accountable for the security of their corporate data assets. In the Unisphere-IOUG survey, 36% expressed concerns about data security as they moved data environments to the cloud.

**Reorient your cost structure.** While on-premise systems typically involve upfront capital expenditures, a cloud-based approach spreads costs across subscription plans. However, subscription costs can quickly add up, requiring a different methodology for calculating expenditures. Upfront investments will be minimal, but costs associated with increasing usage—as well as skills and resources still required to plan and build out capabilities—may escalate, resulting in sticker shock. In the Unisphere Research survey of IOUG members, 32% cited worries about hidden or unforeseen costs of cloud subscriptions as a challenge to cloud adoption, making this one of the leading concerns. Similarly, 28% cited higher licensing costs to run in a public cloud than in current on-premise solutions.

**Consider storage requirements.** The storage requirements for a cloud-based data warehouse or data lake may be massive. The increase in data and analytical capabilities can mean exponential increases in storage. It has always been challenging to scale onsite storage to meet the requirements of a booming data environment, which were often met by deploying strategies such as data compression, along with additional hardware. Cloud reduces the need for such strategy but also introduces new challenges, such as latency, integration, and the potential costs due to charges, whether they be per megabyte of data transfer or monthly fees.

Overall, cloud computing is proving to be a boon to adoption and expansion of data warehouses and data lakes. The key is preparing the organization for the endless possibilities this architecture brings.

—Joe McKendrick





# A Warehouse by the Lake: How to Automate the Ideal Data Infrastructure

IN SOCIETY AND IN OUR ORGANIZATIONS, we need a healthy balance between creativity and process. One prevents the other from stagnating or getting out of control. We must progress with structure and balance to innovate effectively. The same is true in our data ecosystems.

Data lakes and data warehouses are both methods of panning for gold, aiming to distill actionable data insight but using different approaches. We often hear people talk about lakes and warehouses as if we must choose one or the other. In reality, they serve different purposes and are complementary. While both provide storage for data that can be queried for analytical purposes, each has a different structure, supports different formats, and is optimized for different uses.

This article studies the benefits of data lakes and data warehouses before exploring how they can be harnessed most effectively using automation software.

## WHY USE A DATA WAREHOUSE?

Data lakes and the exploratory technologies that unstructured big data enables are only as useful as your company's ability to assimilate their findings into a structured environment. This is where the data warehouse takes over: a data lake can be added as a source to a data warehouse, and its data blended with other real-time and batch sources to provide rich, contextualized business insight.

Originally, the data warehouse was used to store and organize data, but now it supports and drives business processes. The data warehouse is no longer playing catch-up. Today we can spin up prototype designs in minutes and get our infrastructure up and running in days. We use cloud platforms like Snowflake and Microsoft Azure Synapse to run queries in seconds and only pay for the amount of compute and processing power we need. The choice of database is no longer a ten-year decision, given how much easier it is to migrate using metadata-driven tools.

## WHY USE A DATA LAKE?

With the pace of change in data technology, it's hard to predict how we will use the data we ingest now in the future. Data lakes are useful because you don't have to prep or clean the data before storing it so that you can retain as much potential value as possible for future use.

Data scientists can harness AI and ML, which represent limitless opportunities, but also potentially large and expensive workloads. For these reasons, organizations are increasingly turning to data lakes to help store large amounts of unstructured data that can be accessed by a wide variety of services.

This can range from raw to semistructured data and varying grades of curated data sets. A data lake gives a variety of consumers access to the appropriate data for their needs, but it can be stored relatively cheaply, without the need to use ETL or ELT to ingest it into a data warehouse.

## DATA IS INEVITABLE, BUT INFORMATION IS NOT

While the exponential growth of data makes more insight available, it also means the infrastructure that stores and analyzes it becomes necessarily more complex. This infrastructure needs to adapt as new demands emerge (constantly) and as data sources evolve (periodically). It's a fallacy to think we can create the ultimate data infrastructure that won't need to be changed.

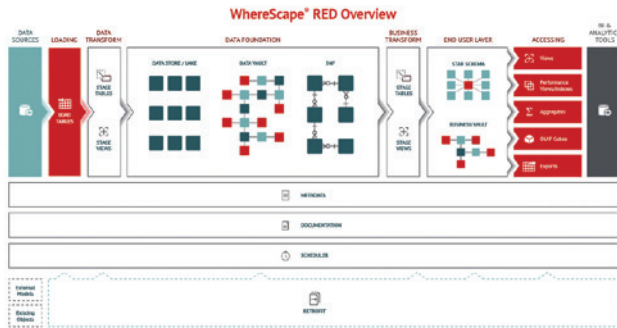
However, it is possible to design and build for change, and pairing a data lake and a data warehouse together gives the most agile framework for getting rapid insight from conventional and big data ingestion now and in the future. So how do we add structure to this rich but complex data fabric? While many companies will offer huge teams of expensive data wizards to do the job, in reality, we cannot hope to harness its potential effectively without automating time-consuming, repeatable processes.

With the automation technology available today, these complex processes can be abstracted into an orchestration layer, from which IT teams can maintain control without having to do the menial tasks by hand.

## DATA WAREHOUSE AUTOMATION

Software such as WhereScape puts a simplifying model on top of your existing data ecosystem to make it far easier, faster, and cheaper to build a complex and powerful data warehouse. Staff can design structures in a drag-and-drop GUI, build prototypes with actual company data, then once all requirements have been agreed upon, the tool physicalizes the model by generating thousands of lines of code, essentially doing many weeks of hand-coding work in seconds. This enables teams to produce usable infrastructure in days, not months.



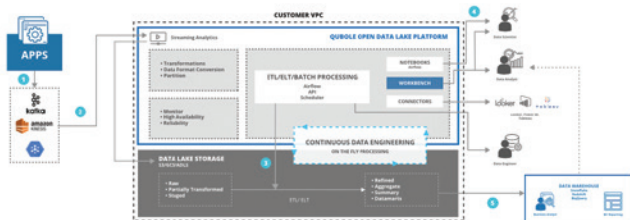


WhereScape is metadata-driven, which means that every action taken is recorded in metadata and stored in a repository. Therefore, documentation can be produced at the touch of a button, with full lineage, which enables track-back and track-forward functionality.

## DATA LAKE AUTOMATION

A data lake platform such as Qubole provides end-to-end services that reduce the time, effort, and cost required to run data pipelines, streaming analytics, and machine learning (ML) workloads on any cloud.

For ad hoc and streaming analytics, the platform's workbench allows the data team to author, save, collaborate and share reports and queries. Qubole enables the Data Analytics team to develop and deliver ad hoc SQL analytics through optimized ANSI/ISO-SQL (Presto, Hive, and SparkSQL) and third-party tools such as Tableau, Looker, and Git native integrations. The Data Analytics team can build streaming data pipelines, combine them with multiple streaming, and batch datasets to gain real-time insights.



Qubole's ML-specific capabilities such as offline editing, multi-language interpreters, and version control deliver faster results. Qubole supports data scientists with comprehensive support for Jupyter Notebooks, Spark, TensorFlow, and Scala, as well as integrations with ML tools like RStudio, SageMaker, and others. It enables the data science team to access all the data and tools they need to build predictive analytical models in collaboration with each other and the other data teams across the enterprise.

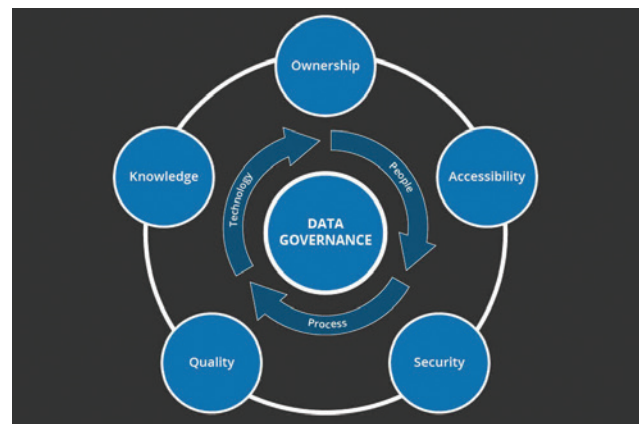
The platform automates pipeline creation, scale, and monitoring, allowing the data team to easily create, schedule, and manage workloads for continuous data engineering. Use

the processing engine and language of choice like Apache Spark, Hive, Presto with SQL, Python, R, and Scala.

Qubole reduces complexity at the data management level and eliminates infrastructure complexity besides providing a near-zero administration experience for the Data Admin teams by automating the mundane daily tasks needed to provide users with access to your data. It puts saving costs first for customers which can be reinvested in use cases, running more workloads, deriving more results.

Moreover, the platform runtime automates the cloud infrastructure provisioning, deployment and further optimizes it with workload-aware autoscaling, cluster lifecycle management, intelligent spot management, and heterogeneous cluster management to give the best TCO industry-wide.

In addition, Qubole has built-in security and governance controls which are enterprise-grade, like providing ACID compliance for granular read writes for GDPR and CCPA compliance, RBAC, and IAM integrations with cloud providers.



Qubole even provides built-in integrations with solutions like BigQuery and AWS Data Lake Formation as well as the ability to integrate other processes with your data through the Qubole platform APIs and SDKs.

## IN CONCLUSION

The danger of not knowing precisely what each structure does, or viewing them as an either/or choice, is that we can miss out on the potential agility and business value of using a data lake and a data warehouse in conjunction. As our data fabric becomes more complex, a diverse portfolio of analytical tools becomes more useful, and a symbiosis of lake and warehouse makes increasingly more sense.

Qubole [www.qubole.com](http://www.qubole.com) WhereScape [www.wherescape.com](http://www.wherescape.com)

To read Gartner's report, "Data Hubs, Data Lakes and Data Warehouses: How They Are Different and Why They Are Better Together," by Ted Friedman and Nick Heudecker (2020), go to [wherescape.com/DBTA](http://wherescape.com/DBTA).

# Three Essential Attributes to Modernize Your Data Real-Estate: A Data Warehouse, Data Lake, and Real-time Streaming



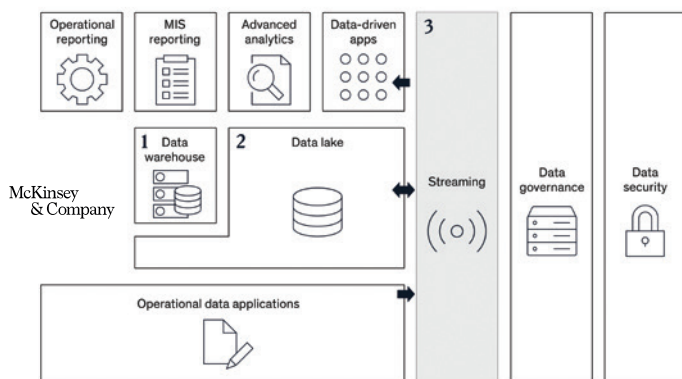
Clive Bearman, Director of Product Marketing for Data Integration, Qlik

**I'VE GOT A CONFESSION.** I'm a data guy. There, I said it. In fact, I'm such a data guy that it borders on obsession. I'm always scanning the feeds for articles about data. Well, you can imagine my delight when I stumbled upon a recent [blog post](#) by McKinsey & Company's [Henning Soller](#) and [Asin Tavakoli](#) that highlighted three requirements to transform your company into a data-driven powerhouse. I must admit, however, that I often read with skepticism, and it's not long before my inner monologue starts dismissing the authors' position with a large "Pah! What do they know?"

This time something was very different. The McKinsey crew was onto something that I could agree with. Some readers might say that it's just confirmation bias, but I'd have to respectfully disagree. The reason is that the McKinsey authors were describing market forces I knew very well and advocating an approach we advise here at [Qlik](#).

The three attributes for digital transformation recommended by the McKinsey authors start with a cloud [data warehouse](#), add an open [data lake](#) and finish with a [real-time data streaming](#). If I didn't know better, I'd think they were paid to promote Qlik's thinking. For the record, they aren't. So, now you can see why I was so stunned. There it was in black and white, with a reference architecture to boot.

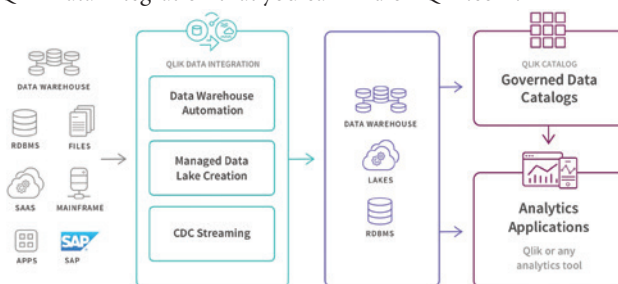
Overview of the typical components of a reference data architecture



McKinsey's reference data architecture is based on three pillars that sit on a foundational [data-ingestion](#) layer:

1. The data warehouse pillar supports predictable, highly critical reporting, such as regulatory compliance and financial reporting. In my last blog post, I commented how many of these great options today exist in the cloud.
2. The data lake pillar is ideal for less stringent reporting needs, as well as advanced analytics use cases that require large-scale data processing.
3. Real-time streaming. This pillar enables real-time use cases as well as rule-based analytics. The transactional databases that serve the pillars are connected either directly (streaming) or through the data lake (exhibit).

Now, let's compare the former diagram to the architecture of Qlik Data Integration that you can find on Qlik.com.



The similarities are striking, while the differences are mainly in layout. McKinsey's diagram operates in both the horizontal and vertical; the Qlik diagram reads linearly from right to left. Both contain the same conceptual components. In an event, the architectural benefits are clear.

Companies that base their data architectures on these core principles can be more agile, scalable and resilient. They can accommodate additional use cases for the data and prove more cost effective. In fact, I'll point you to a recent Nucleus ROI report that documented how one company discovered a 400% productivity improvement for its cloud data warehouse. You can download that report [here](#).

## CONCLUSION

Business leaders around the world recognize the value of becoming a data-driven organization, and many companies have begun to implement advanced analytics and artificial intelligence use cases. However, they often struggle to morph their existing legacy systems to support new requirements and analytics at scale. This was highlighted in a recent McKinsey and Company article, which stated that in order to become a data-driven organization you must modernize the IT estate with a cloud data warehouse, an open data lake and a real-time streaming platform. Not surprisingly, these three tenets form the backbone of Qlik's [Data Integration platform](#).

## ABOUT QLIK

Qlik's vision is a data-literate world, where everyone can use data and analytics to improve decision-making and solve their most challenging problems. A private SaaS company, Qlik provides an end-to-end, real-time data integration and analytics cloud platform to close the gaps between data, insights and action. By transforming data into Active Intelligence, businesses can drive better decisions, improve revenue and profitability, and optimize customer relationships. Qlik does business in more than 100 countries and serves over 50,000 customers around the world. For more information, visit [www.qlik.com](http://www.qlik.com).

Qlik [www.qlik.com](http://www.qlik.com)



# The Rise of Open Data Lake Analytics

## LATELY WE'VE SEEN A MASSIVE

rise in popularity and usage of the cloud data warehouse. And it's not a surprise when we see some of the benefits around performance and data ingestion. However, as businesses are becoming more data-driven with a need to make more informed decisions faster, cloud data warehouses pose some challenges.

Specifically, accessing and analyzing data in the cloud data warehouse has become time-consuming and costly, not to mention the challenges that come with vendor lock-in. This approach assumes that data must be ingested and integrated into one database to provide those real-time insights, but data warehousing systems are closed source, use proprietary formats for data storage, and tend to be very expensive.

Companies today have a mix of structured, unstructured, static, and streaming data that's usually in many different formats and stored in many different systems. Eventually this data will end up in the data lake because it is less costly and it's easy to store massive amounts of data in it. The challenge in this scenario is how to join all that data and analyze it.

## THE SOLUTION? OPEN DATA LAKE ANALYTICS

The value of data comes from running analytics and making decisions based on those results. And increasingly, these decisions are based on evaluating and processing not just parts of the data, but all the data in the data lake and across the data lake and other databases.

An Open Data Lake Architecture consists of a loosely coupled, disaggregated stack that enables

querying across many data sources, without having to move any of the data. It is built on open source, a federated, extensible distributed SQL query engine. The key tenets of Open Data Lake Analytics are open source, open formats, open interfaces and open cloud.

1. Open source: PrestoDB is governed by the Linux Foundation's Presto Foundation and is completely open source under the Apache 2.0 license.
2. Open formats: PrestoDB doesn't use proprietary formats and can read data from the same schemas and tables using the same data formats Apache ORC, Apache Parquet, and more.
3. Open interfaces: By adhering to the ANSI SQL standard, PrestoDB allows for seamless integration with existing SQL systems.
4. Open cloud: PrestoDB is cloud-agnostic and it runs as a query engine without storage, natively aligns with containers, and can be run on any cloud.

## WHY MOVE TO AN OPEN DATA LAKE ANALYTICS ARCHITECTURE? THERE ARE SEVERAL BENEFITS:

- No vendor lock-in: An open data lake architecture gives you the advantage of using open source technologies to get analytics on your data lake without having to use proprietary software that requires vendor lock-in.
- Flexibility to apply multiple data processing techniques on the same data without copying it in many different places: There are many optimized open formats to store data in a structured yet

highly compressed form. Open query engines like PrestoDB that support these formats give users the power to choose which engine to use for different use cases on the same set of data. This is extremely powerful—using open formats gives companies the flexibility to pick the right engine for the right job without the need for an expensive migration.

- Best technologies from the best engineers: Many open source projects like PrestoDB are built at internet giants and used by companies like Uber and Twitter. These companies continue to innovate on these projects and end users get the testing & innovation built for Facebook-scale. In addition, companies get the goodness of what comes with open source software—flexibility plus the power of a community that can provide help, fixes, and quick development.

If you're ready to move to the PrestoDB-based Open Data Lake Analytics approach, Ahana can help. [Ahana Cloud for Presto](#) is a managed service for Presto in the cloud that is easy to use, comes fully integrated with a built-in catalog and data sources, and is cloud native. In less than an hour you'll be running PrestoDB for your data lake analytics.

Want to learn more about PrestoDB? Download our free whitepaper, <https://ahana.io/whitepaper-what-is-presto>, to learn more about the technology including architecture, use cases, and how to get started.

[ahana](#)  
www.ahana.io



# A Modern, Cloud-Native Approach to Accelerating Data Insights

## EVERY BUSINESS IS A DATA BUSINESS.

Yet most organizations still struggle to capture and transform data from diverse and disparate sources to stay competitive. Today's data teams struggle with what we call the "Three Vs" of modern data:

- **Volume:** This year, Matillion customers loaded 5.4 trillion rows of data per month into cloud data warehouses. And that number continues to go up.
- **Variety:** Enterprise organizations used an average of 1,080 data sources in enterprise analytics (IDG, 2019).
- **Velocity:** Business is moving faster than ever. Organizations need to act on information as close to real-time as possible, which means that information has to be available and ready for analytics.

## DATA TEAMS MUST IMPROVE EFFICIENCY TO KEEP UP

For data teams to continue bringing data in from various sources and making it analytics-ready at the pace that the business demands (i.e., as soon as possible), the most effective strategy is to find efficiencies that speed up work: minimize the time it takes to do "custodial" data tasks like coding pipelines, do more work in parallel, and productionize workflows so that multiple team members can step in at any time.

The right tools can help you improve data team efficiency and analytics productivity to not only reduce the workload of your teams but also help people across the organization get to insights faster. Here are four ways modern data teams can begin to move at modern speeds.

### 1. Reuse and borrow to create repeatable processes

If you run a data analysis that yields useful knowledge for end users, they will want you to run that analysis again. And again. If you can build repeatable

pipelines and processes into saved jobs, that will be immensely helpful to your future self. In coding, one of the first things developers do is look at libraries that they can reuse. Data tools should offer that same functionality.

## SLACK SAVES TIME WITH REPEATABLE PROCESSES

[Slack](#), a Matillion customer, used repeatable patterns to reduce the number of discrete workflows they had from 10 down to just one. By streamlining and productionizing efforts, they were able to reduce the time needed to generate new reports from six hours down to just 30 minutes. Those aren't incremental improvements—they're game-changers for the company.

### 2. Unleash the Cloud

If you're not working with data in the cloud, you're missing out on a major opportunity to modernize how you work with data and move faster. The cloud is faster, more scalable, and more affordable than traditional data architectures. Add in cloud data warehouses and cloud-native data integration tools that are built to take full advantage of the speed and scale of the cloud, and data team productivity can skyrocket.

## REDUCE TIME TO INSIGHT SIGNIFICANTLY IN THE CLOUD

Several Matillion customers have seen huge speed gains in the cloud. DocuSign [reduced its ETL runtime by 72 percent in the cloud](#). The San Francisco Giants, working with Matillion partner Data Clymer, [reduced time to new insights by 50 percent](#).

### 3. Leverage the Lakehouse architecture

Different teams use data differently. Data scientists are likely to pull the data

they need from a data lake, while data analysts and engineers work within a data warehouse. They are working in two different environments but duplicating data and processes, which creates extra work.

This more traditional architecture shows a split, where one group goes off to a data lake to do data science and the other is working within a data warehouse environment. Why aren't their needs met using a central data team and data location?

## ENTER THE LAKEHOUSE

The more modern approach is utilizing [the Lakehouse](#). With a Lakehouse, you load data once, apply transformation and clean up the data once. Then you make sure all data teams have access to that nice, clean data, whether they're doing modeling, reporting, or any other activity. By consolidating data in a Lakehouse, you're consolidating work and making it possible to speed up analytics for faster time to insight.

### 4. Choose tools that foster collaboration

Any tool or platform that enables collaboration is essential for working efficiently within data teams. Collaboration can mean different things. It could be working independently on parts of projects that will be combined later (for example, using Git). Or it can mean collaborating in real-time within a shared workspace. Ideally, you want a tool that supports both collaboration types.

Ready to move faster? See how Matillion can help your data team improve efficiency. [Request a demo at \[www.matillion.com/demo/\]\(http://www.matillion.com/demo/\)](#).

Matillion  
matillion.com



# Convergence of Data Hubs, Lakes, and Warehouses is Needed to Unify BI/AI Teams and Projects



## IT'S A FOREGONE CONCLUSION

that the locus of the great engine of modern analytics will be the cloud. Less clear is the source of the data that will be consumed by this engine as an operational workload. Use cases will vary, of course, with some determined by external demands, such as sovereignty of data. Others will be determined by internal demands, such as security, latency and bandwidth. In still other cases, some combination of user preferences and tools will determine where data is located.

Historically, structured data has resided in on-prem databases or enterprise data warehouses. Semi-structured data has been pooled in data lakes, though today it is finding its way into cloud object stores, such as S3, ADLS, and Google Cloud Storage. Importantly, though, all these repositories house consolidated datasets. Innumerable pick-up files remain at large in the wild, such as spreadsheets and web service-enabled data repositories that must be accessed through RESTful APIs (in JSON formats) or XML (in SOAP format).

The question is, can data warehouses and data lakes, as we have known them, feed the cloud analytic engine of the future properly? Traditional data warehouses are complex and require significant IT support, ranging from IT ops (for capacity planning and initial deployment) to database administrators (for system tuning, configuration and end-user support). Further, getting data in and out of these systems often requires an expensive collection of complex tools—such as those from traditional data integration vendors—that constitute a component we'll call a "data hub." Using these tools demands yet another layer of specialized IT skills and resources.

Data lakes were supposed to simplify and improve this situation. After all, they supported schemaless semi-structured data, used open-source, open-standard tools and, for more than a decade, were largely based on a single standard paradigm: Hadoop. And while you didn't necessarily need DBAs or Integration Specialists to work with a data lake, in order to extract real value from the depths of a Data Lake, you needed specialized skills and knowledge.

For all these reasons, yesterday's data hubs, lakes and warehouses are anachronisms in a world of cloud-based analytics. Most data lake and data warehouse vendors have essentially forklifted their existing on-premise architectures and assumptions into the cloud. What the cloud needs, though, is a new type of data hub/lake/warehouse, one that has been freshly conceived to meet the demands—and support the opportunities—enabled by the cloud.

Modern analytics must grapple with critical outstanding questions. Where to unify disparate data? What tools are best for fusing the data? Who can directly access, enrich and consume that data? What kind of support is needed to effectively deliver tangible business value? A persistent though flexible data repository, with built-in data integration capabilities and that acts in a self-service (rather than IT-mediated) mode, is going to be key. It must be a fully-managed, rapidly-scalable service, capable of bridging disparate and diverse data sources and analytics tools. It must be directly usable by business analysts, data engineers and scientists, power-users and, yes, even IT.

These next-generation cloud data warehouse architectures are emerging. For example, Actian's Avalanche Cloud

Data Warehouse has been built to take advantage of the economics and elasticity of the cloud. The compute and storage are separated to support a more granular, burst-scalable, pay-for-what-you-use experience. The containerization of Actian Avalanche and its interoperability with Google Anthos eliminates the need for separate APIs and integrations by cloud and on-premise platforms. Importantly, enterprise-grade data integration for SaaS and on-premise applications is built in as a seamless managed service.

But a new architecture is only part of a modern analytics solution. The other part lies in empowering anyone with even a modicum of analytical experience to pull data directly into this next generation cloud data warehouse. Application developers need to be able to easily use analytics functions such as SQL or functions they write in popular programming languages like Python, and business users should be empowered to use the visualization, BI, or advanced analytics tool of their choosing—from the same simple-to-use environment and without the need to rely on IT.

Yes, the cloud will be the locus of the great engine of modern analytics but think of it more like the power plant of a Tesla rather than a next-generation V8 combustion engine. You need to fuel it in a different way. You need to rethink the entire paradigm of data capture, storage and analytics—not just from a technical perspective but also from a user-engagement perspective. Once you begin thinking anew, you can start to make the most of the opportunities afforded by that great analytic engine in the cloud.

Actian

[www.actian.com](http://www.actian.com)

# Pythian

## Accelerating Your Analytics Journey

### DATA IS THE MOST VALUABLE ASSET

a business can harness to drive innovation, performance and profitability. This hunger for actionable information fuels tremendous growth in analytics spending.

But, for the leaders responsible for an enterprise-wide analytics practice, the obstacles to success can be numerous. Not only is the technology still evolving, so are the processes and skills it takes to succeed.

### FIRST, SIZE UP THE ROAD AHEAD

Whether you're at the beginning of your analytics maturity journey or well on the way, knowing where you're headed is critical.

These are the top five business drivers fueling analytics today:

1. Grow revenues
2. Increase operational efficiency
- 3 Strengthen customer experiences
4. Drive innovation
5. Improve compliance

Aligning your analytics to at least one of these goals establishes a strong foundation. It gives stakeholders a way to frame expectations, implement the right technologies and measure forward progress.

### CREATE YOUR ROADMAP

Once you've established your destination, it's time to hit the road. The analytics journey typically progresses through these four stages of maturity:

1. **See.** Deliver insights on where the business is today and was historically.
2. **Predict.** Project future scenarios and calculate what to do next.
3. **Do.** Automate and orchestrate processes by using data to improve other applications.
4. **Create.** Identify new opportunities by sharing analytics data with product development or external stakeholders.

Every journey progresses differently; many do not move in a linear fashion. We've seen organizations launch data platforms exclusively to develop predictive models, skipping the See stage entirely. Others immediately leverage the Create stage to fuel new applications.

Most common is a fragmented, multi-year journey where departments operate at different stages. This makes a strong data platform and a clear strategy more important than ever. Both provide mechanisms to support users who have very different data needs.

### 1. SEE

Most organizations embark with the simple desire to "see" their business by comparing current conditions to past performance.

Here, forward-thinking organizations embrace self-service analytics. Instead of static reports or limited dashboards, self-service analytics empower the business to use the tools of their choice and access any or all data for exploration. Its benefits include faster access to insights and flexible outputs.

Key to this environment is a platform that can leverage data from an almost unlimited number of data sources, at any volume and velocity.

### 2. PREDICT

Once users gain visibility to analytics on present and past performance, predicting the future becomes the next logical stage.

Machine learning technology typically underpins this phase, making significant quantities of good data and a well-designed data platform essential. The ideal platform will have a structure equipped for well-governed, aggregated and clean data as well as ungoverned (or lightly governed), non-aggregated data.

Two other elements to prioritize at this stage are choosing the best

model and automating as many of the supporting steps as possible.

### 3. DO

The third phase of maturity is all about making analytics actionable. The outputs themselves serve as raw material that feeds into another application to drive business outcomes.

A good example is feeding insights into the recommendation engine of an e-commerce platform, which then powers near real-time delivery of personalized marketing to consumers.

This stage transforms the analytics program—and the underlying data platform—into a mission-critical business system. That means optimizing for near real-time responsiveness, high uptime and clear data quality.

### 4. CREATE

While companies realize analytics yield great business value, they often overlook the revenue potential and product development opportunities available. This is the fourth stage.

Here, the product development team leverages analytics to drive new service offerings, enhance features within an existing product or to position the analytics themselves as a product.

These revenue opportunities often come as a byproduct of other work. Instituting feedback loops helps recognize and channel ideas to the right business teams.

### ARE WE THERE YET?

Ready to fast-track your analytics journey with more detailed guidance? Download our full analytics maturity whitepaper at [https://resources.pythian.com/hubfs/691534/White\\_Papers/Pythian-Whitepaper-Accelerating-Your-Analytics-Journey.pdf](https://resources.pythian.com/hubfs/691534/White_Papers/Pythian-Whitepaper-Accelerating-Your-Analytics-Journey.pdf).

Pythian

[www.pythian.com](http://www.pythian.com)



# Open, No-Copy Data Lake Architecture

**A NEW GENERATION** of cloud data lake services is emerging that brings together the best properties of data warehouses and data lakes to satisfy the requirements of modern-day analytics use cases. Organizations no longer need to choose between low-cost cloud data lake storage and the performance and easy data access of a data warehouse.

- Characteristics of a next-generation cloud data lake include:
- Separation of compute and data (not just compute and storage)
- Elimination of copies of data
- Improved usability via transactions, record-level mutations and time travel

## SEPARATION OF COMPUTE AND DATA

The concept of separating compute and storage has been around for years. However, the next-generation cloud data lake architecture goes beyond that by enabling the separation of compute and data with data being its own tier.

The separation of compute and data allows for the preservation of data in standardized open file and table formats, providing the flexibility to use the best technology (e.g., Apache Spark, Dremio or Apache Kafka) for the analytics use case at hand while avoiding lock-in with a particular vendor. Organizations can continue to leverage low-cost, infinitely scalable data lake storage services such as Amazon S3 or Azure Data Lake Storage (ADLS) in concert with a rich data management layer that structures the data to obtain the usability attributes of a data warehouse.

## NO-COPY ARCHITECTURE

Previous iterations of data architecture involved creating and

managing many disconnected copies of data. [IDC](#) found that 60 percent of storage is dedicated to managing copies of data at a cost of \$55 billion per year collectively. The primary reasons for this are:

- Copying the data into a data warehouse
- Creating aggregated or sorted tables in the data warehouse for performance reasons
- Making personalized copies of data in the data warehouse for different users or teams
- Creating BI extracts or imports
- Creating cubes to segment data
- Downloading data to a local machine to get better performance with tools and libraries

Disconnected copies of data are the key obstacle to making data available to data consumers—they lead to astronomical costs and slow down time to value. They also make security and governance difficult because of the resulting data inconsistencies and the inability to easily determine and control who has access to a dataset.

The next-generation cloud data lake architecture eliminates the copies and confusion. With solutions such as Dremio, there is no need to load or import the data to query it because it can be queried directly from S3 and ADLS. Performance-related copies are also eliminated by features such as [data reflections](#), which automatically maintain physically optimized representations of datasets and use compression techniques to deliver fast performance without creating and maintaining separate aggregation tables, cubes and BI extracts, which also limit the scope of data available to a user. The need to export data

into local files is also eliminated via high-throughput parallelized data access interfaces such as Apache Arrow Flight, which is more than [20x faster than pyodbc](#). These fast data transfer rates mean data scientists can use client applications such as Jupyter to access data and train models without having to persist the data.

## TRANSACTIONAL DATA TIER

In past generations of data lakes, it has been difficult to replicate some of the powerful capabilities of data warehouses such as data versioning, concurrent transactions and record-level mutations (updates, deletes, etc.) across large-scale datasets. These constraints have resulted from the table structure of the schema. However, technologies such as [Apache Iceberg](#) and Delta Lake have overcome these limitations, enabling these management capabilities for datasets at data lake scale. While both technologies provide a transactional data tier, Iceberg best supports the loosely coupled, open architecture theme because it supports transactions from any engine, allowing organizations to choose the best tool for the job. Traditional data warehouses provide similar capabilities but, with Iceberg, these same capabilities can be provided within an open and flexible data lake environment.

## LEARN MORE

To learn more about this next-generation cloud data lake architecture, download this [white paper](#) today.

Dremio  
www.dremio.com



# Using Logical Data Fabric to Unify Data Lakes, Data Warehouses, and Clouds for Analytics

**DATA LAKES ARE EMERGING** as the preferred modern resource for storing and managing data, but the traditional data warehouse is not going away anytime soon. Gartner maintains that whereas data lakes are best suited to data scientists who need to perform multidimensional analysis in order to discover the right questions to ask, data warehouses are better suited to analysts who already know which questions to ask. And cloud data lakes and data warehouses are evolving as the modern-day equivalents of their traditional on-premises counterparts. As a result, all these types of repositories will continue to coexist to meet their different use cases.

## THE SILOS PERSIST

The challenge is that data in the data warehouse will be siloed from data in the data lake, and if we add cloud repositories to the mix, data is now siloed across all three types of systems. This restricts data access and prevents stakeholders from gaining a holistic view across all three systems or to run queries across all three simultaneously. Data can be brought together from these systems for analysis, but this takes time, effort, and investment, as the data would need to be replicated into an additional repository, and would need to undergo heavy transformation in order to be understood across all three systems.

Ideally, regardless of the use case, analysts and data scientists should be able to access all of the available data, equally. But with limited data access, organizations are unable to gain real-time, complete views of all relevant data about customers, supply chains, business performance, and more, to make timely, informed decisions.

## WEAVING THE LOGICAL DATA FABRIC

One of the most powerful, effective ways to unify data across silos, without having to collect all of the data and house it in a new repository, is through a data architecture called “logical data fabric.” With logical data fabric, adherents aim to provide a more universal and holistic approach to integrating diverse components of physically distributed data environments. Logical data fabric uses data virtualization to integrate the siloed data, so that the users and applications do not have to separately access each data silo.

A data virtualization layer can provide a single view of logically integrated, multisourced data while supporting federated queries to the sources. Data virtualization layers offer transparent access so that users do not need to know how to access the data or even where the data is stored—an important tenet of logical data fabric.

## ANALYTICS ON STEROIDS

By unifying data warehouses and data lakes across on-premises and in the cloud, logical data fabric provides substantial benefits for analytics. It not only supports both structured and unstructured data, but also data-at-rest and data-in-motion side-by-side. Thus, logical data fabric provides real-time insights from the freshest data as they are collected in these repositories.

Because logical data fabric serves as the central source for data across the enterprise, it also provides a strong foundation for establishing a central catalog to document the location, type, and format of each dataset. Using a logical data fabric’s data catalog, analysts can conduct data discovery in one central place rather than having to use a variety of separate business intelligence (BI) tools. Logical data fabric enables key parts of its data catalog to be fully automated, while at the same time allowing users to edit descriptions or add notes.

Through its very structure as a central data access layer, logical data fabric can capture the lineage of any listed data set, indicating the original sources as well as any transformations it underwent along the way, and catalog this information as well. Similarly, a logical data fabric’s catalog can capture a dataset’s associations to other datasets.

## UNIFY ALL DATA ACROSS THE ENTERPRISE

Data lakes and cloud systems are not replacing the traditional data warehouse. Logical data fabric makes it possible for all three systems to work in concert, for a sum that is significantly greater than its parts.

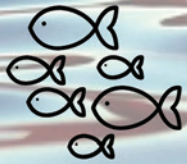
---

Denodo  
[www.denodo.com](http://www.denodo.com)



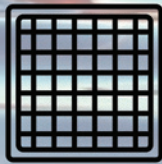
# What Makes a Data Lake Thrive?

Integrate



multi-source

De-muck



cleanse

Protect



mask PII

Fish



analyze

[www.iri.com/blog/business-intelligence/the-use-of-data-lakes](http://www.iri.com/blog/business-intelligence/the-use-of-data-lakes)

info@iri.com  
[www.iri.com/voracity](http://www.iri.com/voracity)  
[linkedin.com/company/iri-the-cosort-company](https://www.linkedin.com/company/iri-the-cosort-company)



**IRI Voracity**  
An Insatiable Appetite for Data

## YOUR CONNECTION TO THE INDUSTRY

*Database Trends and Applications* produces 10 original email newsletters, each with targeted content on specific industry topics. Subscribe today to receive concise reports on what's happening in the data world.

**database**  
TRENDS AND APPLICATIONS

**BDQ**  
BIG DATA QUARTERLY

[dbta.com/newsletters](http://dbta.com/newsletters)

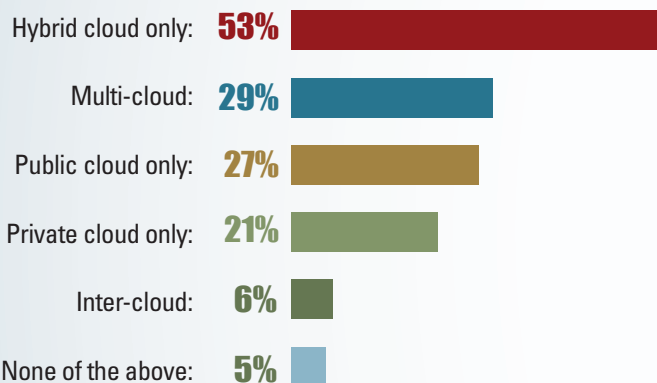
## HYBRID AND MULTI-CLOUD DEPLOYMENTS EXPAND

**O**rganizations are no longer choosing between on-prem and cloud deployments. Today, a hybrid scenario is a given, and, increasingly, the approach is also multi-cloud. Companies are choosing this route for a range of reasons, including agility, a desire to avoid vendor lock-in, the need to support new use cases, and a keen focus on customer experience. As cloud strategies are increasingly adopted, there is also an embrace of containers and serverless computing, as well as orchestration and automation.



### HYBRID CLOUD USE IS GROWING.

Which type of cloud do you anticipate your organization will be using primarily by 2025? (More than one answer permitted.)



Source: "DBTA Digital Transformation and Cloud Workloads Study" ([www.dbta.com](http://www.dbta.com))

ALONG WITH INCREASING USE OF PUBLIC CLOUD PLATFORMS, THE USE OF CONTAINER AND SERVERLESS FUNCTIONS IS EXPECTED TO RISE.

### Forrester predicts that:

- The global public cloud infrastructure market will grow **35% to \$120 billion in 2021**.
- Alibaba will take the **#3 revenue spot** globally, after AWS and Microsoft Azure.
- The use of container and serverless functions to build new apps and modernize old ones will increase from about **20%** of developers pre-pandemic using these approaches to **25%** using serverless and **30%** using containers regularly, creating a surge in global demand for both multi-cloud container development platforms and public-cloud container/serverless services.

Source: "Predictions 2021: Cloud Computing Powers Pandemic Recovery" (<https://go.forrester.com>)

### PUBLIC CLOUD ADOPTION CONTINUES TO GROW.



- More than 50%** of enterprise workloads and data are expected to be in a public cloud within 12 months.

Source: "Flexera 2020 State of the Cloud Report" ([www.flexera.com](http://www.flexera.com))

ORCHESTRATION AND AUTOMATION HAVE BECOME CRITICAL TO THE USE OF CONTAINERS, WHICH ENABLES ORGANIZATIONS TO ACCELERATE DELIVERY CYCLES. KUBERNETES IS THE MOST WIDELY USED CONTAINER TECHNOLOGY.



- Almost 90%** of containers are orchestrated by Kubernetes, Amazon ECS, Mesos, or Nomad.
- Kubernetes use has **more than doubled** since 2017, and it has become the leading container orchestration platform, used by half of the organizations that are running containers.

Organizations running containers use Kubernetes in self-managed clusters or through a cloud provider service such as:



- Google Kubernetes Engine**
- Azure Kubernetes Service**
- Amazon Elastic Kubernetes**

Source: Datadog 2020 Container Report: "11 Facts About Real-World Container Use" ([www.datadoghq.com](http://www.datadoghq.com))



*The democratization of analytics capabilities has led to a much more rapid pace of analytic discovery within companies—but it also creates new challenges.*

**Tim VanTassel** is a vice president at FICO ([www.fico.com](http://www.fico.com)), overseeing solutions, advisory, and special sales. His responsibilities span the global credit lifecycle, fraud, and marketing lines of business.

# Information and Wisdom: The Art of Building and Operationalizing an Analytic Model



**Tim VanTassel**  
FICO

## Combining Data With a Deeper Understanding of the Relationships Captured by Machine Learning AI

*By Tim VanTassel*

TODAY, THE IDEA OF PICKING UP THE PHONE TO CALL A TRAVEL agent to plan out a vacation seems archaic. It used to be that everyone used a travel agent, but nowadays it's an occupation that caters mainly to the wealthy. Yet there was something about it that was uniquely valuable, something that the legions of vacation-planning websites and travel-booking apps—powered by oceans of data and cutting-edge algorithms—haven't been able to fully replicate.

Most of us aren't lucky enough to plan vacations for ourselves every waking hour of the day. We might do it once or twice a year. Travel agents, by contrast, do this all the time. And because they make travel arrangements so frequently, they naturally develop heuristics (i.e., shortcuts) and gain skills for doing it well.

Experienced travel agents can plan out a trip across Europe very efficiently by stitching together data on ticket prices, transfers, local attractions, and hotel availabilities. They combine this with the wisdom they've accumulated from years of planning similar trips and then produce itineraries that have been expertly mapped out.

Vacation planning websites and travel booking apps have disintermediated the travel agent by delivering the information that they relied on directly into the hands of the masses. Now we all have as much information as travel agents once had (or perhaps even

more), but what we lack is their accumulated wisdom. As a result, we're left to assemble our itineraries as best we can, with all of the wasted hours and suboptimal routes that entails.

### The Age of the Citizen Data Scientist

In the business world, we've seen a similar transition happen in the realm of analytics. We are living in the age of the citizen data scientist—empowering a wide swath of employees to follow their curiosity and plumb the depths of their organizations' data lakes using sophisticated machine learning algorithms.

This is great! The democratization of analytics capabilities has led to a much more rapid pace of analytic discovery within companies. But it also creates new challenges.

Machine learning is excellent at uncovering new information by discovering new correlations within the data. However, it is not able to fully understand the relative value of those correlations and the trade-offs between the costs associated with operationalizing the model and the lift that are being achieved. Machine learning AI lacks the wisdom to understand the trade-offs inherent in translating analytic insights into real-world decisions. This understanding can also lack the rigor to ensure that the model is ethical in its use of the data. ▶



## A Practitioner's Guide to Analytic Model Development

When developing new analytic models, what's needed is the combination of information (produced efficiently by machine learning algorithms), explainability (also provided by technology), and wisdom (provided by experienced data scientists and decision management experts).

Many organizations specialize in helping companies do exactly that. In most situations, any good model development follows these steps:

### Step 1. Start with the problem.

You can't identify valuable datasets until you have clearly defined what problem they will be valuable in helping you solve. This brings us back to our travel agent paradigm—make sure that your citizen data scientist understands the business context of what is being modeled.

### Step 2. Define the behavior you are trying to predict.

For example, if a bank is trying to predict attrition, it will need to define exactly what types of customer behavior constitute attrition. This may include information such as the account closing, becoming inactive, the account balance dropping below a certain level, or account spending dropping below a specified threshold.

### Step 3. Evaluate all potentially relevant data sources.

Once you have defined what you are trying to predict, you will want to be fairly liberal in terms of the different types of data you consider. Machine learning is enormously helpful at this stage because it can automatically evaluate data, such as the following:

- Account data
- Transactional data—recency, frequency, transaction type, volume
- Call center data
- Collections data

### Step 4. Plan for implementation.

Assess the implementation environment for limitations on data availability and model type compatibility. Understand what it will take in terms of time and effort to fill any gaps that may exist. Additionally, any approvals or governance processes for model signoff should be well understood at this stage.

### Step 5. Wrangle the data.

Once you have the confidence that the data in question will deliver significant predictive value for a key business problem, you can then invest the necessary time and resources into wrangling that data and getting it ready for use in production

---

*Now we all have as much information as travel agents once had (or perhaps even more), but what we lack is their accumulated wisdom.*

*In the business world, we've seen a similar transition happen in the realm of analytics.*

---



systems. A key consideration at this step will be to understand and address any data biases that naturally reside in the data or that may be unintentionally manufactured through sampling.

### Step 6. Build models.

After you have wrangled the relevant data, you'll then need to apply the right analytic techniques (e.g., feature generation, variable reduction, random forest, etc.) to build models and see how predictive that data is for the different behaviors you are focused on. Go back to Step 1 and make sure that the model itself is using the underlying data in an appropriate, ethical, and efficient way. If you don't get significant predictive power for one of the data sources, don't include it because it will make the next step (Step 7) unnecessarily difficult.

### Step 7. Operationalize your model.

When the data is ready for use in production, the model can be incorporated into your organization's existing decision strategies and business rules, and then operationalized for use in making better decisions.

## Good Model Development

Many organizations are adept at helping companies implement new analytic models; what's needed is that combination of information, explainability, and wisdom—bringing expertise in building and operationalizing analytics models to bear. Any good model development will likely follow these seven steps.





# Insider Threat Detection With Accelerated Machine Learning

SECURING INFORMATION SYSTEMS AND DATA is a foundation for any organization. Detection of insider threats can be a considerable challenge for threat detection systems and security analysts. This is due to the difficulty of determining non-normal actions from internal system behavior data.

Each organization's internal network behavior is very complex. Access controls, data permissions, credentials, tokens, and application integrations are applied to the many business units, cadres, and teams based on their purpose to fulfill the organization's goals. In most cases, the users of each of these groups will be accessing different utilities within the organization multiple times a day. Any time that data or proprietary information is accessed, a log is generated to account for the action. The activities carried out by system users generate massive volumes of data over time.

Most engineers will consolidate logs into a single location with access to the cloud, making this even easier to accomplish. The logs will be stored in their "raw" format in addition to going through a transformation and enrichment workflow to structure logs from heterogeneous sources into datasets with specific schemas. From here, behavior of users and additional methods of threat categorization can be developed.

Rules-based threat categorizations take this data and look for certain behaviors and return results that meet a certain condition. Statistics-based threat categorizations can look for meaningful outliers of risk from a distribution generated by sampling from the overall population.

## Threat Categorization

Many researchers have devoted their efforts to building threat categorization models with machine learning. One traditional approach to the insider threat detection problem is supervised learning, which builds data classification models from training data captured in the logs. Unfortunately, the training process

for supervised learning methods tends to be time-consuming and expensive while dealing with highly imbalanced data.

Attempts to combat challenges associated with training time and data have led to new approaches leveraging deep learning and graph analytics on GPU hardware. In the last decade, new innovations in machine learning frameworks that were written in the Python coding language have leveraged the power of GPUs. These frameworks provide the ability to explore, transform, and apply algorithms on the data in the GPU but at a speed never before seen on a CPU.

This allows for a reduction in training time and expense as well as the opportunity to apply advanced analytical techniques on the data that is not subject to the same constraints of traditional machine learning methods. For example, a deep autoencoder is a deep learning algorithm that can represent nonlinear relationships in the data and does not necessarily require a label associated with each log to generate a feature that represents a baseline of user behavior on a network. The autoencoder is able to encode normal behavior, and, when decoded, there will be minimal error. If new feature vectors are generated containing non-normal behavior, the decoded output would result in an error that could trigger an alert for further investigation.

## Graph Analytics

Graph analytics is a method of analysis concerning the nature of networks and relationships (edges) between entities (vertices). In a dynamic network of many entities, graph analytics can tell us about the behavior and relationships between them. The issue that comes into play is the scale to which many of these networks can grow. If the network is derived from the logs captured from system user interaction, the number of edges could reach billions. Recent innovations in graph analytics on GPUs offer the capability to process networks of this size using standard graph algorithms such as PageRank and Louvain 1,000x faster than comparable CPU methods, resulting in quicker insights and applied responses.

The goal of insider threat detection is to distinguish non-normal behavior from a mountain of recorded actions. To identify potential threats, the organization must establish relevant algorithms and automate these processes in an efficient way. The application of GPUs is no longer limited to graphics and display, but has expanded to scientific computation, engineering simulation, and AI. Leveraging these algorithms and advanced hardware provides the potential to greatly boost the performance of computationally intensive programs for both machine and deep learning.



**Jim Scott** is head of developer relations, Data Science, at NVIDIA ([www.nvidia.com](http://www.nvidia.com)).

Over his career, he has held positions running operations, engineering, architecture, and QA teams in the big data, regulatory, digital advertising, retail analytics, IoT, financial services, manufacturing, healthcare, chemicals, and geographical information systems industries.



# Denial of Service Attacks Can Come Directly From Silicon Valley

THE WORLD CHANGED OVER THE LAST YEAR. Future historians will complete their theses focusing on different quarters or even specific months of 2020. But one of the most overused clichés in thinking about this period of time has been the idea that “the more things change, the more they remain the same.” Let’s consider sports in 2020. Major League Baseball had a 60-game season, the NBA finals were played in October, and cardboard cutouts took the place of fans in every sport. However, the Lakers won the NBA finals, the Dodgers won the World Series with the Yankees playing deep into the playoffs, and Tom Brady went to his 10th Super Bowl. The more things change ...

## Why Are We Surprised?

No one should be surprised at the significant catastrophic security breach centered around SolarWinds’ Orion Platform last September, which affected approximately 18,000 customers directly with an unlimited amount of potential secondary effects. The attack may have come from Russian hackers, but it’s possible that there were other nefarious parties involved heretofore not considered.

How does this type of breach occur? Occam’s razor, the postulate that the simplest reason is the most likely reason is worth considering. And the simplest explanation is that so many breached software systems are connected to the internet. The information superhighway is the neural network of global communications that unites all electronic things and respectively allows all electronic devices to be snooped, monitored, penetrated, and violated by all other electronic things. So, if the monitoring of a software system is necessary, and that monitoring is transmitted through the internet, then the 21st century’s

soft underbelly is exposed. The same is true for development of that monitoring software, but only if the development system is connected to the internet, or anyone working on it is connected to the internet, or if anyone working on the software can somehow carry that software out of the semi-secure location in which they work. Why are we surprised when a significant security breach happens of this magnitude? Continuing with the Occam’s razor theme, the answer to this question is equally simple. Please pardon us for sounding very 1975 (we are getting up there in years), but simply stated: The best cybersecurity is concrete backed up by air.

The magnitude of the SolarWinds Orion “hack” cannot be overstated because, regardless of how many high-priced consultants and veterans of the tech overlords of the six cities of Silicon Valley ([www.dbta.com/BigDataQuarterly/Articles/The-Six-Cities-of-Silicon-Valley-125014.aspx](http://www.dbta.com/BigDataQuarterly/Articles/The-Six-Cities-of-Silicon-Valley-125014.aspx)) attempt to create a softer perception of this disaster, every network and system that was touched directly or indirectly by the incursion will need to be rebuilt from the ground up. We say this because there will never be a comprehensive understanding of the degree and precision of penetration at one level or another or in one system or another. Recovering from this calamity will require an effort on the level of Y2K.

## Old-School Methods

More importantly, a legitimate plan to prevent a recurrence or something worse will require that the business-critical applications at every one of the affected entities be secured via the “old-school” methods. Readers, young and old, should contemplate how a similar situation in the late 1990s would have unfolded. For example, consider a young database administrator in 1999 saying, “Hey boss, I have an idea. We should take our most critical data and store it remotely on servers being offered by this guy in Seattle. He was just named *Time* magazine’s ‘Person of the Year,’ and he sells books. Plus, he has extra storage space, and he is leasing it cheaply. No guarantees at all, but it should work!” This story would not have had a happy ending, as it is likely that the next conversation that the intrepid DBA had would have been outside of work and include the line, “Well, I’m between jobs.” The point of this sarcastic allegory is that in a mere 2 decades, the very philosophy of protecting the most critical application functionality and data has completely veered into the bizarre. In 2021, we casually use remote cloud services, all run by massive companies that have their own agendas that do not always coincide with their customers’ critical business needs, using those



**Michael Corey** is co-founder of LicenseFortress ([www.licensefortress.com](http://www.licensefortress.com)). He was recognized in 2015 and 2017 as one of the top 100 people who influence the cloud. Corey is an Oracle Ace, VMware vExpert, a former Microsoft Data Platform MVP, and a past president of the IOUG. Check out his blog at <http://michaelcorey.com>.



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

services. Again, the best cybersecurity is concrete backed up by air.

Extending our sensitivities regarding critical applications to the electrical, communication, military, and medical grids of Western civilization's infrastructure, the apparent reason for using cloud resources is the perception of reduced cost that does not always live up to its contrived media coverage. That

same point can be applied to connecting to the internet of these components of critical data infrastructure. To quote a March 2018 article in *The Washington Examiner*, "The federal agency that oversees the nation's power grid was a prime target of nine Iranian hackers whom the Justice Department is indicting for 'malicious' cyber activity." Why is our power grid infrastructure even on the internet? Again, the best cybersecurity is concrete backed up by air.

A different but equally volatile situation is the tremendous power in the tech overlords' hands to eliminate companies—temporarily or permanently—that choose to place their critical applications and data in the cloud. Regardless of the circumstances behind recent events or anyone's personal opinions on the efficacy, responsibility, authority, or power of the cloud providers, hardware producers, and mobile application disseminators, the facts remain clear. They have the unprecedented power to shut down your company if you give them that power. Without a moment's notice, they can perform the greatest denial of service attacks conceivable. The actions may be genuine or malevolent, but we know they are all legal, apparently, at least for now. There are many valid reasons to use remote hosting or cloud services. Non-critical scalability, bursting, simplicity of access, speed of allocation, and unique microservices are among those justifications. However, the perception of

---

*Over the last few months, the world has seen that implementing impervious cybersecurity on the internet and protection from nefarious actors is nearly impossible.*

---

security and control over your critical data and cost reduction should not be among the reasons.

### **Critical Infrastructure Remains On-Premise**

To summarize, in case the reader has yet to infer our point, it is our opinion that all critical infrastructure and data should remain on-premise.

We believe that solid disaster recovery plans should be built and tested and that those DR systems must be sufficiently geographically dispersed to circumvent malicious actors of all stripes. The concept of the hybrid cloud is a perfectly elastic model to meet these critical requirements. Over the last few months, the world has seen that implementing impervious cybersecurity on the internet and protection from nefarious actors is nearly impossible. We have also seen that the most powerful tech overlords can immediately take complete control over and shut down a business that a moment before was their "customer." One more time, the old-school security approach works because the best cybersecurity is concrete backed up by air.

---

### **Reference Items**

- <https://www.channele2e.com/technology/security/solarwinds-orion-breach-hacking-incident-timeline-and-updated-details/#:~:text=Adjusted%20Attack%20Timeline%3A%20SolarWinds%20CEO,SolarWinds%20on%20September%204%2C%202019>
- <https://www.nytimes.com/2021/01/02/us/politics/russian-hacking-government.html>
- <https://www.washingtonexaminer.com/policy/energy/iranian-hackers-targeted-power-grid-watchdog-justice-department-says>

**BEST** SUMMER  
PRACTICES **2021**

## **DATA STRATEGIES FOR THE REAL-TIME ERA**

For sponsorship details, contact Stephen Faig, [stephen@dbta.com](mailto:stephen@dbta.com), or 908-795-3702.



# Move Over Unicorn, It's the Rock Star's Time

AS THE WORLD OF DATA ANALYTICS CONTINUES to evolve and reshape after a tumultuous 2020, the need for agility is rapidly driving a new era in data culture in which it is imperative to handle data immediately and at scale. While emphasis on self-service data and analytics has been top-of-mind for some time now, the shift to self-sufficiency is held back by culture, not technology. With the new year pushing more robotics process automation at all levels of the business—and for all data users—organizations are becoming more acutely aware that true enablement isn't just about tools and tech. It's about people.

Achieving enterprise value from all data requires a new animal, and we're not talking about the elusive unicorn—the data scientist. While the blend of business, big data, and math that the data scientist brings to the table will always have a home in data-driven organizations, the most impactful animal going forward will be the data professional—a business-oriented user who can confidently, efficiently, and actionably work with data.

Similar to musical rock stars, “data rock stars” crave independence, freedom of movement, and the opportunity to show off their skills. They are collaborative, inspirational, and driven to push the limits of their ability. Going forward, true data enablement in an organization will be organic and viral—and what's more rock 'n' roll than that?

## Pick an Instrument

Contrary to popular opinion, becoming a data rock star isn't only about mastering a specific skill. Rather, it begins with having a vision (or, a melody) you want to manifest, then picking the right instrument—in this case, an area of expertise, be it in data visualization, architecture, integration, or so on—and learning to make music, whether as a solo musician, band member, or part of a larger data symphony.

The ability to pick your instrument—your skill of choice—is a liberating one that propels the data rock star from jam sessions in the garage to touring in their venue of choice. Becoming an empowered data user means going beyond learning how to use data and analytics in any particular role and attaining transferrable job skills that supersede job function.



**Lindy Ryan** is the chief content officer at Radiant Advisors, a research and advisory firm that leverages experience and industry involvement to deliver pragmatic guidance in executing data and analytics strategies. She is an award-winning professor of visual analytics and the author of two textbooks on visual data culture and data visualization. Follow her on Twitter @Radiant\_Lindy.

---

*While the blend of business, big data, and math that the data scientist brings to the table will always have a home in data-driven organizations, the most impactful animal going forward will be a business-oriented user who can confidently work with data.*

---

## Learn to Play

If music theory is the study of the practices and possibilities of music, then data theory is the foundation that data users need to develop the skills and competencies to become true data rock stars. Just as music theory is a practical discipline that encompasses the methods and concepts musicians use in creating music, so, too, is data education of paramount import for the data rock star.

An uneducated data user is a dangerous data user. Before one can learn to play their instrument, budding data rock stars need to understand the principles, best practices, and limitations of working with data and analytics. Here, not only is attaining data literacy critical, but, as with many things, practice—or, more aptly, iteration—is key. Once the basics have been established, data rock stars will be equipped with the knowledge and theories needed to learn to play in harmony with the rest of the organization.

## Be a Conductor

Enabling business users to become data rock stars requires culture change, and when we talk about culture, there are myriad strategies to achieve pervasiveness across dozens, hundreds, even thousands, of people within the organization. Many call it “stages of maturity,” but what it truly distills to is this: How can you empower an individual and make them a luminary so others will want to emulate and follow in their footsteps—or, how do you transform data users into data rock stars?

Leaders should focus on enabling their people to become data rock stars, both for the organization and for their employees' personal career development. Create positive paths with engaging education and certifications, and facilitate engaging communities while removing negative obstacles of intimidation, embarrassment, and risk-taking that come while learning. Think low-barrier, high-reward.

Ultimately, as organizations continue to look for ways to achieve data at scale, it's not a matter of using tools and technologies to their full potential—it's about helping *people* reach their full potential. After all, even skill areas are tools. It's people who make the music.





# The \$100-Trillion Opportunity for IoT

THE FULL POTENTIAL FOR COMPANIES that digitally transform is predicted to be \$100 trillion by 2025. This astonishing value would be reached through the combination of digital technologies—mobile, cloud, AI, sensors, and analytics, among others. These are accelerating progress exponentially, said the World Economic Forum, but this growth can be realized only if there is collaboration between business, policymakers, and non-governmental organizations.

The crux is in the word “collaboration.” In the past, companies used to build big entry barriers to their businesses; Warren Buffett calls them “moats.” In this way, to be successful, a company has a competitive advantage, allowing it to maintain pricing power (and fat profit margins). That philosophy is changing with the emergence of the ecosystem-based economy, sparked by the likes of Google and Alibaba—which make their living by linking other organizations and people together. The born-digital ecosystem strategy has been to drive a wedge between the producers and their consumers by linking them on one platform—becoming the self-declared aggregator. This has been going on for a long time, but, similar to proverbial frogs in the boiling pot, non-digital companies were concerned, but not yet frightened, about the wedge being driven between them and their customers as they should have been—until now.

## Designed Ecosystems

COVID-19 was a wake-up call. Lockdowns cut off traditional sales channels overnight, and finally companies understood the real danger of no longer being in touch with their own customer base because the aggregators were getting all the feedback—the data, the reviews, and the related purchases.

The answer to this is explicitly designed digital ecosystems. Before going into the digital part, a quick word on the explicitly designed ecosystem: A good example is Nestle. After it designed its patented espresso cups, it gave licenses to just a few manufacturers to make and sell the coffee machines to go with them, allowing others to sell the machines and Nestle to sell the coffee (again and again)—a win/win.

If you add “digital” to the explicitly designed ecosystem, the potential for growth is exponential. But there is the danger of defining digital in the narrow sense—in other words, thinking you can just automate and integrate the interactions that are already going on, for example, by using APIs to transform your interactions electronically. The better way is to think of the ecosystem as an

advanced infrastructure from production to outlets and beyond. True digital transformation not only automates but turns your infrastructure into an “infostructure.”

What does this mean? In the past, companies focused on the production quality, and, as soon as the product left the shop, the job was done. Now, by using the latest technology—converting your infrastructure to an infostructure—you can learn what your customers do, what your customers’ customers think, and what makes your partners tick. All are deeply connected, so you benefit from the data, the insights, and the relationships.

## The Infostructure Journey

The first part of the infostructure journey is to accept that your traditional company will become a software company. By embedding sensors into your products, they will be made smart, IoT-enabled, and connected—allowing you to stay in touch with the product throughout its lifecycle. Now, envision the customer’s location as the furthest outpost of your organization—becoming the “edge,” so to speak. In other words, IoT will allow you to grow the customer experience into a connected customer experience.

But no experience stops at the edge—and the true value will come out of how your wider ecosystem partners will use the infostructure. Just imagine what the digital extension of Nestle’s ecosystem would mean. Driving home in my connected Mercedes, my arrival would not only switch on my Nespresso coffee machine but also, through my nest, the heating in the house and the outside lights.

For the latter to stand a chance, it needs to have a solid foundation in B2B integration. This means using not only the latest API integration technology, such as microservice gateways, but also application and service mesh technologies to solve complex interdependencies between those continuously evolving relationships.

## Drag, Drop, Done

To keep up with the pace of change, these technologies need to move across to the business side, allowing them control through visual drag-and-drop interfaces that create the digital fabric of the ecosystem.

Technology will not be the only hurdle in the new digitally driven ecosystem economy. The approach is inherently more stressful and chaotic than more traditional approaches, and it brings a lot of challenges.

But the prize is clear: exponential growth by tapping into a \$100 trillion opportunity. COVID was the wake-up call through which adoption of digitalization leapt ahead by 2 years in the space of 2 months. The boards of traditional industry—from electronics to cars and retailers to oil drillers—now get it. The race is on to become a software-first company in a rich, bespoke, digital ecosystem.



**Bart Schouw** is vice president of technology and digital alliances, Software AG ([www.softwareag.com](http://www.softwareag.com)).



# Rethinking the Case for Governance

AFTER A WILD AND TURBULENT 2020, the new year has ushered in a renewed commitment to establishing or improving corporate governance. Yet, positive energy aside, our traditional approach to endorsing governance of data, analytics, or AI remains fraught. As a result, governance initiatives springing from an earnest desire to do right (e.g., responsible AI), as well as the need to not do wrong (e.g., regulatory/compliance), struggle to enlist broad coalitions of the willing.

The fault lies with our collective understanding about what governance is and does. Too often, governance is still seen as a necessary evil—one that imposes bureaucratic overhead by mandating activities that slow progress, inhibit innovation, and prohibit teams from rapidly delivering on their commitments. Sadly, it is often unwittingly promoted as such even by ardent supporters.

## Putting Governance to Work

To put governance to work, we must change the mindset of our internal and external constituents. Yes, governance is a means to an end. But the endgame is not about compliance, perfunctory completion of checklists, elimination of risk, or passing the buck. Rather, the purpose of governance is to increase the cadence and quality of our collective and individual decisions.

This is true whether the deliverable is a simple BI dashboard or a complicated AI algorithm. From providing the proper interpretation of a business metric to deciding if an AI application can be deployed equitably and safely, governance serves as a universal translator and navigational beacon—one which ensures that everyone, regardless of role, is speaking the same language and working toward the same outcome.

Governance should, therefore, result in better-informed and more confident executives, managers, analysts, developers, and ultimately—consumers. This is perhaps counterintuitively accomplished by making

time for mindful discussion of common, clearly defined goals. More specifically, governance encourages and indeed mandates collective, and therefore more robust, consideration of questions such as the following:

- Who is the application designed for? Intended to benefit?
- What is the target population or environment?
- What are the expected operating conditions?
- What decisions or actions will result from this application?
- What other interactions may impact this decision/action?
- What are the limitations of the information/system?
- Can and should we act within those limitations? If so, when and how?
- Will this achieve the intended outcome?
- Is “it” doing what we thought it would do? Running amok? Veering into unexplored territory?
- How will/do we know?

Equally important, governance creates mechanisms to probe and refine the collective understanding over time as systems and processes are developed and deployed. As such, governance is continuous learning in practice. All governance activities—from informing go/no-go decisions to profiling data to instituting robust ModelOps—operate in the service of a singular goal. Namely, the continuous enrichment of the team’s knowledge and ability to make informed decisions. This increases the probability that desired outcomes are achieved while unintended consequences are expeditiously identified and addressed. Governance can also go one step further by soliciting consideration of not just whether the intended goal is achievable but whether it is the right goal at all.

It has been well established that effective governance escalates innovation, improves product and services offerings, enhances the user experience, and reduces risk. Success often comes from avoiding unforced errors before grievous harm occurs or too much time and money are wasted in pursuit of an unrealizable or unwise goal. But, in order to benefit, the organization must engage.

Garnering broad, willing engagement requires shifting the narrative. Done right, governance encourages curiosity, increases knowledge, and helps teams uncover what they need to know to build better, do better, be better. If the activities mandated by your governance process cannot be articulated in these terms, a rethink is in order.



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

## AD INDEX

Melissa .....	Cover 4
IRI .....	23

## BEST PRACTICE

Action .....	19
Ahana .....	17
Denodo .....	22
Dremio .....	21
Matillion .....	18
Pythian .....	20
Qlik .....	16
Qubole/WhereScape ....	14



# DATA SOURCEBOOK

Now, more than ever, the ability to pivot and adapt is a key characteristic of modern companies striving to position themselves strongly for the future. Download this year's *Data Sourcebook* to dive into the key issues impacting enterprise data management today, and gain insights from leaders in cloud, data architecture, machine learning, and data science and analytics.

**Download Your Copy Today!**  
**<https://bit.ly/BDSbook8>**

FROM THE PUBLISHERS OF

**BDQ**  
BIG DATA QUARTERLY

**database**  
TRENDS AND APPLICATIONS



# Data Quality, Easy as 1-2-3!

Bad data is bad for business. Its real cost is three costs: time, money, and your sanity. Leverage Melissa's 35+ years of expertise in data quality, address verification and identity solutions to stop bad data in its tracks.

**With our solutions, you can:**

**1.iDEnTiFy** - Target your ideal customer and prevent fraud

**2.VERiFy** - Employ real-time address, email, phone and name verification to keep bad data out of your systems

**3.MAtCH** - Consolidate duplicate records for a single customer view



**Repeat after me:** Data Quality, Easy as 1-2-3. Visit [Melissa.com](https://www.Melissa.com) for (1) free trials, (2) free developer credits and (3) code samples. It's easy, like counting up to three!