Volume 5 Number 3 ■ FALL 2019

BDQ
BIG DATA QUARTERLY

BIG DATA 50

COMPANIES DRIVING INNOVATION

Top Three Considerations Before Moving to a Multi-Cloud Model

How the Financial Services Sector Is Leading the Pack With Cloud Analytics

The Artificial Intelligence Plateau

WWW.DBTA.COM

CALL FOR SPEAKERS
NOW OPEN!

DATA SUMMIT

UNLEASH THE POWER OF YOUR DATA

FEATURING THESE SPECIAL EVENTS

DATAOPS BOOT CAMP

AI & MACHINE LEARNING SUMMIT

DATA LAKE BOOT CAMP

DIAMOND KEYNOTE SPONSOR

ORACLE

PLATINUM SPONSOR

VERTICA

GOLD SPONSOR

Quest

BROUGHT TO YOU BY

database TRENDS AND APPLICATIONS

BDQ BIG DATA QUARTERLY

MAY 19–20, 2020

HYATT REGENCY BOSTON
BOSTON, MA

PRECONFERENCE WORKSHOPS
MONDAY, MAY 18

save the date

dbta.com/datasummit

#DataSummit

BDQ
BIG DATA QUARTERLY

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

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

CORPORATE HEADQUARTERS 143 Old Marlton Pike, Medford, NJ 08055

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

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

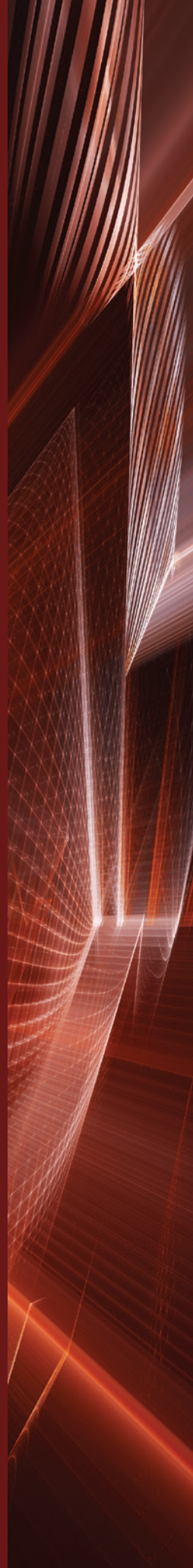
PRINTED IN THE UNITED STATES OF AMERICA

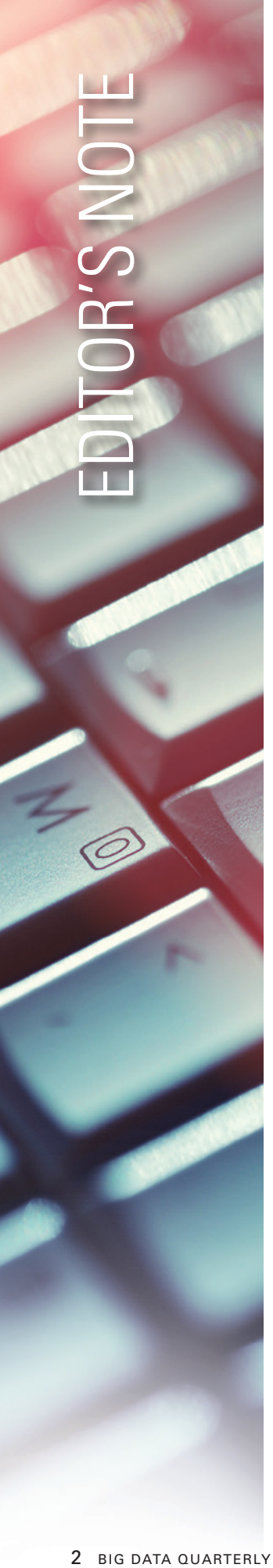
Big Data Quarterly is a resource for IT managers and professionals providing information on the enterprise and technology issues surrounding the 'big data' phenomenon and the need to better manage and extract value from large quantities of structured, unstructured and semi-structured data. *Big Data 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.

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

Information Today, Inc.

CONTENTS





EDITOR'S NOTE

The Race Is On

By Joyce Wells

As organizations race to achieve data-driven insights to improve decision making and enhance customer experience, data managers are embracing a wide range of new and established technologies to keep up with business demand.

A recent survey by Unisphere Research, a division of Information Today, Inc., underscored the importance of next-generation approaches in enabling organizations to better leverage information for business advantage. The survey, sponsored by Pythian, found that not only are data strategies becoming part of enterprise business planning but also that there is a strong shift to real-time delivery of data to advance intelligent enterprise strategies.

Analytics planning is being shaped by an assortment of competing requirements, including the need to reduce costs and deliver better information, the survey revealed. Adoption of machine learning has almost doubled over the past year, as well, to enable internal efficiencies and business growth initiatives. At the same time, data analytics is increasingly moving to the cloud, although concerns about data security linger.

Many of these topics are explored in this issue of *Big Data Quarterly*. In an article on how the financial services sector is benefiting from cloud analytics, Rachel Dines explains that Wall Street may appear to be conservative and have a low tolerance for risk, but when it comes to adoption of new technology to leverage data, the sector is actually at the forefront. And, in an article on the benefits and pitfalls of using more than one cloud platform, Pete Salamanca shares significant issues to consider before embarking on a multi-cloud journey. With the understanding that hybrid and multi-cloud scenarios are becoming increasingly popular, the idea of data orchestration is coming to the fore, notes Alluxio's Dipti Borkar in another article in this issue. "Just as Kubernetes is to compute and containers, data orchestration is to data and to active working sets of data," says Borkar.

Any examination of data management and analytics today must also include data governance. In this issue, Collibra's Myke Lyons shares his perspective on changing data privacy regulations, while columnist Anne Buff makes a case for both business and IT sharing involvement in data governance.

And to expand the discussion about the fast-changing data environment, this issue features the annual Big Data 50, a list of companies driving innovation in new and established data management areas.

There are many more articles in this issue that provide perspective into key industry trends, including DataOps, AI, and others. To stay on top of the latest data technologies, research, news, and thought leadership, visit www.dbta.com/bigdataquarterly.

And, be sure to save the date for Data Summit 2020, the annual conference that brings together IT practitioners and business stakeholders from all types of organizations. The event will take place May 19–20, with preconference workshops on May 18, at the Hyatt Regency Boston.

BIG DATA BRIEFING

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

Azure is collaborating with **Informatica** to help ease the process of migrating analytics workloads to the cloud. With this joint offering from Azure and Informatica, customers receive free code conversion for both the proof-of-value phase and when fully migrating to the cloud, as well as an SQL Data Warehouse subscription for the duration of the proof of value (up to 30 days). https://azure.microsoft.com and www.informatica.com

McAfee has announced the acquisition of **NanoSec**, provider of a multi-cloud, zero-trust application and security platform based in Cupertino, Calif., with an office in Bengalaru. NanoSec's security capabilities will be applied to applications and workloads deployed in containers and Kubernetes and will be integrated into McAfee MVISION Cloud and MVISION Server Protection offerings. www.mcafee.com

Datical, a provider of database release automation solutions, will provide support for Liquibase 3.7, empowering software professionals who use the open source project for managing database schema changes. Additionally, Datical is investing in strategic partnerships with database vendors, including Cassandra, Cockroach Labs, Couchbase, and SAP HANA, among others, to increase community contribution. www.datical.com

Oracle has announced the general availability in all commercial regions of Oracle Functions, **Oracle Cloud Infrastructure's functions-as-a-service** platform. Oracle Functions is built on the Apache 2.0 licensed Fn Project, which can be used anywhere, from a developer laptop to a cloud compute platform, and customers have the option to operate their own functions service in-house or use the cloud-scale Oracle Functions platform to avoid the costs associated with managing infrastructure. www.oracle.com

EdgeConneX is partnering with **Rackspace** to deliver a combination of enterprise cloud solutions with the global EdgeConneX data center platform. This partnership delivers edge-based cloud solutions to enterprises worldwide and facilitates cloud migrations with options ranging from dedicated bare metal environments to variable hybrid cloud and multi-cloud solutions from best-of-breed providers. www.edgeconnex.com

Percona, a provider of open source database software and services, is releasing the Percona Cloud Native Autonomous Database Initiative, a series of products that expand support for cloud-native applications and make it easier for organizations to manage their hybrid multi-cloud environments. www.percona.com

Hewlett Packard Enterprise has acquired the business assets of **MapR**, a data platform for AI and analytics applications powered by scale-out, multi-cloud, and multi-protocol file system technology. www.hpe.com

Cockroach Labs, provider of the distributed SQL database CockroachDB, is receiving $55 million in series C funding, enabling the company to further grow CockroachDB, and bringing total funding to $108.5 million. www.cockroachlabs.com

IBM has announced that its software portfolio is now cloud-native and has been optimized to run on Red Hat OpenShift. Enterprises can build mission-critical applications once and run them on leading public clouds, including AWS, Microsoft Azure, Google Cloud Platform, Alibaba, and IBM Cloud, as well as on private clouds. www.ibm.com

Hitachi Vantara is releasing **Pentaho 8.3**, introducing a series of features designed to support DataOps initiatives. This latest version delivers improved data agility from customers' edge-to-cloud environments while facilitating privacy, security, and overall data governance. www.hitachivantara.com/go/pentaho.html

Google Cloud will begin supporting VMware workloads. With **Google Cloud VMware Solution by CloudSimple**, Google customers will be able to run VMware vSphere-based workloads in GCP. https://cloud.google.com

Unravel Data, a provider of full-stack visibility and AI-powered recommendations, is releasing a new cloud migration assessment to help organizations move data workloads to Azure, AWS, or Google Cloud faster and with lower cost. www.unraveldata.com

Dotscience, a DevOps for machine learning (ML) provider, has emerged from stealth with its platform for collaborative, end-to-end ML data and model management. Dotscience provides a tool that manages the complete AI lifecycle by allowing data scientists and ML engineers to work in familiar ways. www.dotscience.com

Perspective on Data Governance

Myke Lyons, Collibra
Chief Information Security Officer

A data governance- and privacy-focused company founded in the EU that now headquarters its management team in New York City, Collibra has hundreds of customers across Europe and the U.S. As a result, the company keeps a close eye on changing data privacy legislation that may impact its own data governance platform and its customers' data management practices.

Following a $100 million Series-E funding round, led by CapitalG, Alphabet's growth equity investment fund, Myke Lyons joined the data governance and catalog software unicorn in April as its first chief information security officer (CISO). With a background primarily in security architecture, Lyons came to Collibra from ServiceNow, where he was head of security strategy. He recently shared his perspective on changing data privacy regulations, and the direction in which he sees data governance and compliance efforts moving today.

How would you characterize the current state of data privacy regulations in the U.S. now?
It is very much in its infancy. There is a massive gap between those that are creating the policy and those who are collecting and generating the data. There is an opportunity for organizations to improve their understanding and access to data. And by improve their access, I really mean improve how they are going to leverage the data for the betterment of their customers, business, and investors, and put some more clarity around that process.

Is it easy for organizations to understand the different data privacy regulations or are they too disparate?
They are not very consistent, frankly. I think organizations are really struggling with understanding what the goals are with the regulations, and then trying to put them into practice and attempting to work within those confines to help customers. However, the other side is that these regulations also present an opportunity for organizations to start to extract better value from their data and not just continue down the path of collecting more and more.

How so?
The regulations are going to allow organizations to create more innovative products. There are a lot of organizations that have matured their processes around sales and marketing, and around IT and HR, but they are not really improving their data processes, and this is something that I have a lot of interest in helping customers do at Collibra.

One of the newest data privacy regulations, the New York Privacy Act, also known as NYPA, was introduced in May but failed to advance in the state's most recent legislative session.
The roots of that particular regulation came out of GDPR [which took effect on May 25, 2018], similar to the California Consumer Privacy Act [(CCPA) which goes into effect Jan. 1, 2020]. The difference between GDPR, the California Consumer Privacy Act, and the N.Y. Privacy Act requirements was that, in the N.Y. Privacy Act [unlike CCPA], New York citizens would have had the right to sue any organization that has their data and removed a $25 million revenue starting point. The second key point is that it introduced the idea of a "data fiduciary," which is obviously very new and puts the onus on the collectors of data to be the appropriate stewards.

Would the fiduciary have been a single person?
I don't know if it came down to a single named person. They tried to introduce something similar to this in the U.K. a number of years back called the "data custodian" but the U.K. law lost its teeth. It is not a new concept and it is one that privacy advocates have been supporting and something that Kevin Thomas, the New York state senator who introduced the act, was keen on.

That was a key difference from other legislation?
The removal of the $25 million revenue threshold was a big one, as well. If the bodega down the street is collecting your data, it could in fact be sued. And it would have had some impact on startups and their ability to grow.

What is the larger impact of these current and proposed regulations surrounding data privacy?
They are going to help organizations understand concepts such as "privacy by design" as opposed to "privacy bolted on." These will introduce all different types of data products, different opportunities for revenue, and really provide the opportunity for organizations to get more connected to their customers from a service perspective. This is going to allow them to build more trust with customers as their transparency increases. But the first step is that they have to get a handle on the data, know where the data is, and know what they have.

Are there technologies or approaches that will be leveraged by companies to get a better handle on their data?
Definitely. In the governance space, the foundation for any of these efforts is locating the data and where those repositories, data lakes, and data warehouses are, and pulling information in so they can understand where it resides, as well as its lineage—and this is an area of focus for Collibra. And then the next phase is to be able to run machine learning or artificial intelligence to help categorize the data, because manual efforts are not going to get them where they need to be. There has to be a focus on automation and constant upkeep of the data, whether it is the metadata itself or the location of the data.

Looking ahead, say 5 years from now, what is your expectation for how the regulatory and compliance landscape will evolve?
There are some interesting times ahead of us. I can envision a slightly less toothy version of a data privacy regulation coming from the U.S. federal government in the next couple of years. That said, previously, legislation took such a long period of time to actually get passed, and nowadays, it is becoming faster because the speed at which businesses are changing the way they operate is so quick. The U.S. government has launched an initiative whereby every government agency needs to have a chief data officer.

Organizations are struggling with understanding what the goals are with the regulations, and then trying to put them into practice and attempting to work within those confines to help customers.

What is the significance of that mandate?
This is a first step, and once that is accomplished, there will be a number of chief data officers at these agencies, and there will also obviously be chief data officers within private businesses, so I can see there being a data privacy regulation from the federal government. But I can imagine that if the federal government does not act quickly that there could be 40 or 45 states launching their own data privacy regulations. In many cases, these are bipartisan efforts because there is cross-party agreement that data is being used in ways that not everyone is comfortable with.

Interview conducted, condensed, and edited by Joyce Wells.

Seven Trends SHAPING 'Big Data' INTO 'All Data'

By Joe McKendrick

SEVEN TRENDS SHAPING 'BIG DATA' INTO 'ALL DATA'

Has the meaning of big data changed? Many agree that data no longer has to be "big" to meet today's evolving requirements. Perhaps a better way to describe big data would be "all data," suggests Saptarshi Mukherjee, global head of product and solutions marketing, data analytics, at Google Cloud. "Today, we're seeing less of a distinction between 'big data' analytics and data analytics, because, inherently, businesses are exposed to massive data growth, which is coming from a variety of systems and applications," said Mukherjee. "Data analytics needs to address the traits of large amounts of data at all times." ▶

SEVEN TRENDS SHAPING 'BIG DATA' INTO 'ALL DATA'

In particular, open source and cloud tools and platforms have brought data-driven sensibilities into organizations that previously did not have such expertise, making big data more accessible. "Hadoop helped make it easy to collect data quickly," said Madhukar Kumar, VP of product and developer marketing at Redis Labs. "It was bundled with MapReduce to enable a way to crunch data—and the result was a new data ecosystem that grew the initial focus of the big data conversation. Apache Spark ran much faster by keeping all of the data in memory and helped alleviate some of the timeliness problem."

Cloud is opening up a whole new way of approaching tried-and-true systems, such as data warehouses, and blending them with new layers, such as logical data warehouses. "You are taking the traditional analysis engine and running federation over the data lake data and bringing in big data processing technologies, such as Apache Spark, and processing data in the warehouse," said Mukherjee. "This is enabling enterprises to break down data silos and take advantage of distributed computing to analyze any data at scale. Public cloud is accelerating this." When organizations go to the cloud, they are operating in an unconstrained environment where they can commission compute and storage capacity, making data warehouses far more powerful, Mukherjee noted.

In addition, the evolving world of big data has taken on a new meaning beyond simply formats and storage capacity. "It's not the size of the data that matters; it's how you use it. Big data was a marketing term built to describe data with volume, velocity, and variety," said Thomas LaRock, head geek for SolarWinds. "It doesn't have to be large. It could be tiny amounts of data coming from millions of IoT devices."

What are some of the major trends shaping data management as it takes on a larger role in enterprise decisions and operations? The following are some of the key developments seen this year.

REAL TIME

Real-time data and analytics—amplified by the Internet of Things—is a force altering the big data landscape. The need to have real-time analytics to improve business decisions is fueling demand for technologies such as in-memory computing, and it is a trend being seen at many of the world's largest companies, said Abe Kleinfeld, CEO of GridGain Systems.

"As use of automation, machine learning, and AI continues to increase, and as companies amass more and more types of data, the need to quickly analyze data will only become more pronounced," said Kleinfeld, who added that in-memory computing adoption will follow the upward trajectory of machine learning and AI.

Real-time analytics on streaming data is becoming accepted, said Mukherjee. "Enterprises are exposed to massive growth in streaming data. Streaming data is being generated from connected devices and connected applications. Streaming analytics is another technological development that, granted, has been around for many years, but is now ready for mainstream adoption. Specifically, you can analyze this data at a particular time and act on whatever insight you find at that moment, which requires a lot of compute power."

Real-time streaming data and analytics isn't just a technology endeavor—it also requires a transformation in business thinking. "Adoption of streaming analytics not only requires a technological shift but leads to changes in business operations," said Mukherjee. As more companies move to the cloud and have access to distributed compute capacity and fully managed services, there will be more organizations adopting streaming analytics techniques. But while the technology has become more accessible, there's a bigger change-management problem that organizations face when they implement real-time streaming-data analytics solutions. For example, a retailer can analyze real-time customer-demand patterns but its supply chain needs business processes to be in a position to act on those insights, said Mukherjee. Similarly, a bank may be able to analyze real-time transaction data and identify potential risks, but it also needs to have a process to mitigate that danger. "If your business processes are not set up to take those insights and act on them in real time, your organization will not benefit. There needs to be a data culture shift and an understanding at all levels of the organization to successfully implement this technology."

Many companies "still lack sophistication when it comes to understanding what they need to do to achieve their performance and scale goals and drive their real-time business processes," said Kleinfeld. "However, they are beginning to realize that companies that move slowly will be left behind and they need to get real-time business figured out if they want to stay competitive."

DATABASE AS A SERVICE

With the cloud comes rapid growth in demand for database-as-a-service (DBaaS) offerings. "Companies in every industry are looking to consume database as a service to capture all the traditional benefits of cloud computing, as well as to offload operational overhead to database experts so engineering resources can stay focused on driving value elsewhere," said Asya Kamsky, principal developer advocate at MongoDB. With DBaaS, they can more readily "automate database administration tasks such as database configuration, infrastructure provisioning, patching, scaling events, and doing backups so they can focus on delivering new applications and features to their customers far more quickly."

DBaaS also allows organizations to be strategic about where their data lives. For example, companies can take advantage of Microsoft Azure, Google Cloud Platform, and Amazon Web Services, and, if an application has a huge userbase in Australia, data for that application can be stored

Real-time streaming data and analytics isn't just a technology endeavor—it also requires a transformation in business thinking.

nearby in order to offer a low-latency user experience, Kamsky said. "If data privacy regulations demand that certain users' data stays in Germany, for example, it's very easy to manage that. Those are two very significant business challenges facing global organizations today."

DBaaS has additional benefits. For the business, DBaaS enables the offloading of database administration tasks, thereby "freeing up technical resources to focus on more important efforts that will actually drive value for a business," said Kamsky. "DBaaS also provides peace of mind that experts in database management are handling security configurations and optimizations, meaning business leaders are getting the most out of their data management strategies. As a result, their engineering teams will be more productive and will be able to bring new applications and product features to market at a much faster rate."

However, since DBaaS is a cloud implementation, many of the safeguards that need to be applied to cloud migrations also need to be used, Kamsky added. There is some data that organizations will always prefer to host in their own private data centers. In addition, Kamsky said, for massive, established applications, it can be a complex process to lift them to the cloud.

DATA WAREHOUSING AS A SERVICE

As DBaaS gains traction, a similar paradigm, data warehousing-as-a-service (DWaaS), is also under consideration as an option for making data available to the enterprise. "This technology is evolving rapidly with new innovations, including DataOps methodologies," said Itamar Ankorion, SVP and managing director, enterprise data integration, at Qlik. "Companies must be aware that these modern DWaaS platforms require innovation in how data is ingested and organized for analytical processing. By adopting a completely modern data architecture based on DataOps principles and technologies of continuous data integration and data warehouse automation, companies benefit from efficient data management."

While DWaaS is still in the relatively early stages, there is "fast growth in initial or small deployments," said Ankorion. "We expect to see continued high growth and adoption rates within the next year and mass market adoption in the next 5 years." The challenge, he said, is that DWaaS needs complementary technologies to facilitate efficient and continuous data ingestion as well as data warehouse automation. Selecting the right technology partners or perhaps looking deeper into more modern data frameworks such as DataOps will help to achieve success, Ankorion explained.

ENTER AI

Of course, no discussion of the power of data analytics is complete without considering the implications of AI and machine learning.

"We are at a stage where a lot of progress has been made on the analytics side in terms of building machine learning models, but businesses are now looking at applying those models to fast-moving transactions data in real time," said Kumar. "On that front, AI is one of the most significant new technologies that can take deep learning models and serve them against fast-moving time series or streaming data with less than a millisecond of response time. This enables making recommendations—and intelligently acting upon data in real time—a reality. In addition, new hardware technology such as Intel Optane now enables businesses to run extremely large datasets—petabytes—with relatively low costs, but with similar performance of data management running in-memory."

AI as a data management and analysis enabler has recently started to gain more traction, said Redis' Kumar. Some organizations have begun experimenting with running their AI models directly inside of the transactions that are put through a time series data model. This in turn opens up new use cases such as fraud mitigation in real time.

AI is also dramatically improving enterprise big data management capabilities. "It used to be that businesses would consider using different databases for different use cases, but as the need for a dynamic layer between applications and traditional databases increases, there are some who are looking to manipulate different data models closer to the application layer," said Kumar. "This opens up new use cases that combine data models in real time. Examples include searching graph data, combining streams data with JSON, or running AI models on time series solutions."

CONTAINERS, KUBERNETES, AND HYBRID CLOUDS

The trend of data environments becoming increasingly flexible and capable of being moved to where the business demands them is coming from containerization via Kubernetes, which has "blurred the line between the data center and public cloud," said Anupam Singh, general manager of data warehouse for Cloudera. "This allows data management to burst dynamically both inside the data center and the public cloud, without the knowledge of the end user or developer."

How are cloud and container technologies enhancing enterprise data management capabilities? "With the ability ▶

Open source and cloud tools and platforms have brought data-driven sensibilities into organizations that previously did not have such expertise.

to transparently scale computing across the data center and public cloud, without changing applications, developers can be much more ambitious and productive," said Singh. Consider a large financial institution with 22 new use cases across marketing, HR, sales, support, and other areas, said Singh. "In a world of static data center technology, it would have to lift and shift to the public cloud, which would create friction for the deployment of new use cases." With hybrid cloud data management, the IT department can satisfy demand without having to lift and shift petabytes of data.

At the business level, managers "will not have to choose between the data center and the cloud," Singh added. "Their investments in the data center will continue to accrue benefits while the public cloud will be integrated more seamlessly into their data management strategy."

The challenge for data managers is providing direction and oversight for the explosion in activity and innovation that cloud services and containers are enabling. "As soon as the business side can allocate an arbitrary amount of computing for queries, IT teams will be faced with workloads that are runaway, with lax security, and in silos," Singh said. This could result in serious adverse consequences for cloud computing itself. "In this world, the DBA has to be able to predict usage patterns, or data management will be chaotic. Another problem with transient cluster computing is that admins, architects, and developers lose context, history, and security when the dynamically allocated compute disappears. The technology teams will have to learn how to audit, secure, and tune in the world of dynamic computing."

MASTER DATA MANAGEMENT

When it comes to scaling to the enterprise, there's a need for consistency and quality. Master data management (MDM) has taken center stage as the means to keep data aligned with the enterprise. "Data quality has been understood as managing consistency within a dataset and fixing individual bits of data, but more and more, it's being understood as managing consistency across datasets and fixing data as it relates to other datasets," said Jake Freivald, VP, Information Builders. Even though people don't typically master the big data itself, there's a renewed interest in knowing how to connect sets of less-governed big data to highly governed master data. He predicted that while still nascent, MDM adoption will grow as data scientists look for higher-quality data for their analytics. "They are still suffering from the 'garbage in, garbage out' phenomenon, and integrating MDM with big data helps avoid it."

MDM brings more flexibility than previous efforts to centralize data management, such as data warehouses, Freivald added. MDM "doesn't require all of the transactional data to be fully modeled, but provides much more governance and consistency than using big data on its own. Because of that, using MDM is allowing people to have the same level of flexibility—the data is still stored in relatively free-form ways—but with sufficient governance to allow them to connect the dots reliably in more analytical situations."

The business also benefits from MDM, Freivald continued, due to "better data inputs for AI and machine learning, which will lead to better outputs and greater adoption for these potentially transformative technologies." There will be better analytics in general, as well as an "improved ability to house big data-related content in data catalogs, which will increase awareness and usage of big datasets to solve a wider variety of business problems more reliably."

'ALTERNATIVE' DATA

A big part of big data is, of course, data coming from outside the enterprise. While this "alternative" data has been on the scene for many years, there is a new wave of significant adoption, said Gary Read, CEO and founder of Import.io. This may include hedge funds that need extra guidance for investment decisions, as well as retailers that require rapid competitive pricing analysis or data on consumer reviews. The result is that companies are increasingly relying on substantial amounts of alternative data from sources, such as the web, in conjunction with the traditional data sources for business decisions, said Read.

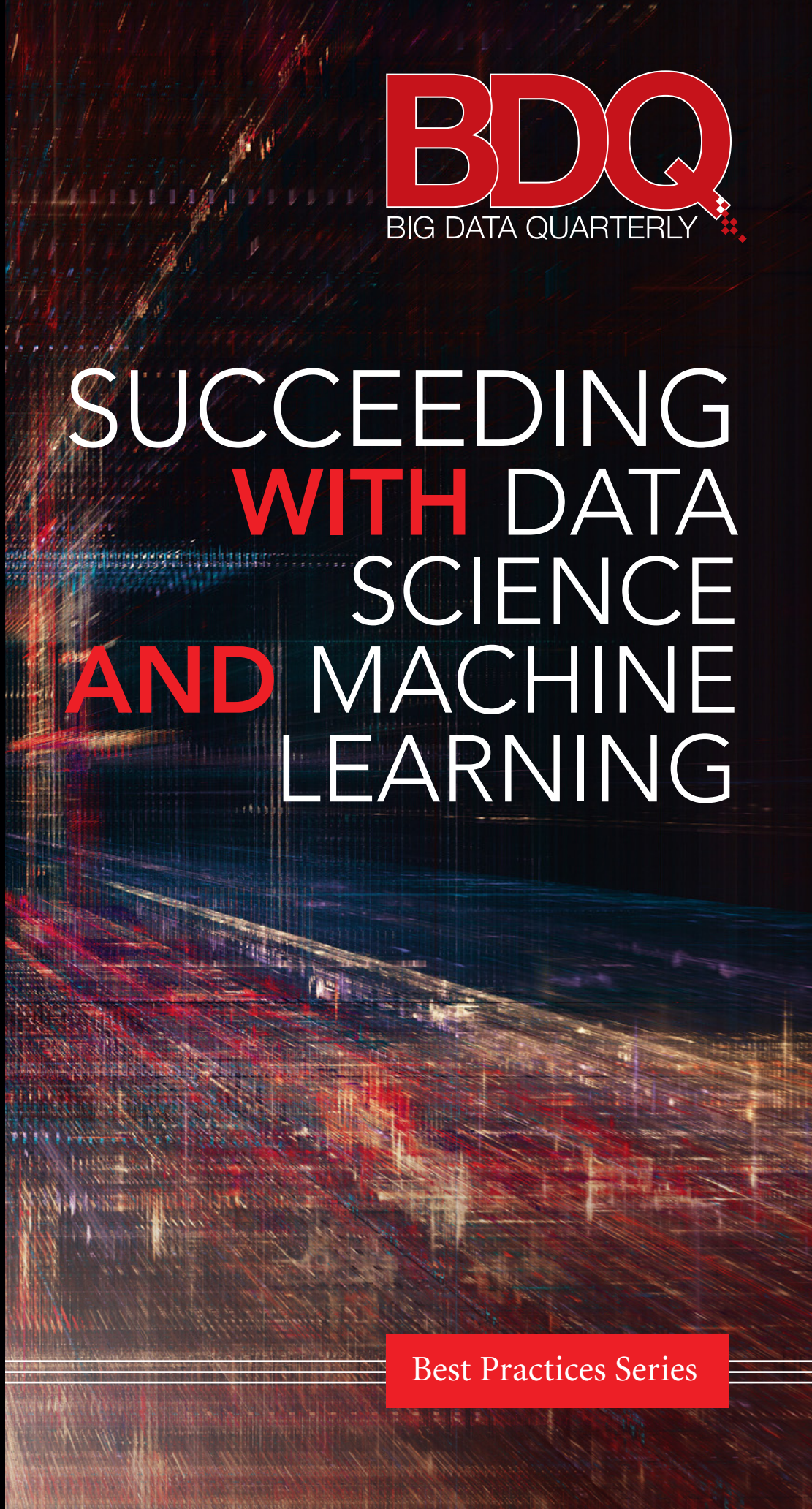
"Currently, the majority of companies are spending internal resources to manage alternative data," Read said. "However, looking forward, data extraction and integration technology will be better equipped to help businesses collect and harvest the value of alternative data, and demand for such solutions will continue to increase."

Technologies that enable the sourcing and collection of alternative data "provide huge efficiencies to the enterprise," Read noted. "Not only are internal resources no longer wasted on writing code and maintaining scripts to manage such data, but there are now an increasing number of services and platforms with the capability to extract data automatically so that the data can be quickly integrated with other data sources to drive decisions."

IRI Voracity

PAGE 14

MACHINE LEARNING IN ANALYTICS AND ANONYMIZATION



BDQ

BIG DATA QUARTERLY

SUCCEEDING WITH DATA SCIENCE AND MACHINE LEARNING

Best Practices Series

BDQ
BIG DATA QUARTERLY

EXPANSIVE DATA FUELS DATA SCIENCE AND MACHINE LEARNING

Best Practices Series

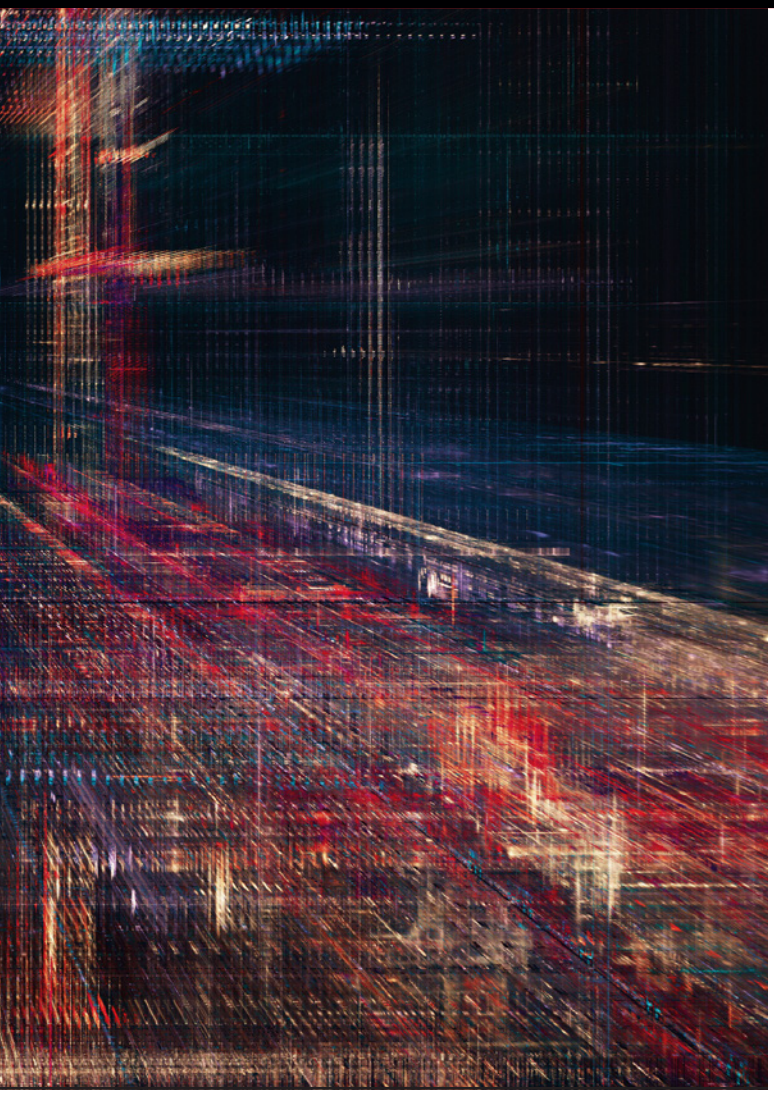
With growing attention devoted to AI, machine learning, and IoT, what we've come to know as big data has become an even broader version of itself. In recent years, big data was seen as an unstoppable force of nature that would either overwhelm enterprises or propel them to new heights. This next generation of big data—we'll call it *expansive data*, pulsing through systems in real time, powering processes unseen to human eyes, and adapting and learning as it goes along—is going to reshape enterprises in ways not even anticipated. This requires attention to new types of tools, platforms, and approaches to deliver value to today's data-hungry businesses.

This next generation of big data—we'll call it *expansive data*, pulsing through systems in real time, powering processes unseen to human eyes, and adapting and learning as it goes along—is going to reshape enterprises in ways not even anticipated.

Expansive data will represent ever-growing volumes of information, potentially increasing within enterprises at a rate of up to 36% a year, according to Dresner Advisory Services. Platforms supporting this growth include Amazon Web Services S3, Spark SQL, Hive, and Hadoop. Additional tools popular in enterprises are Apache Spark and Tensorflow.

Expansive data places even greater demands on enterprise infrastructures, processes, and the managers and administrators responsible for making it all work. That's because organizations are leaning more heavily than ever before on their data assets and analytics capabilities, and initiatives such as AI and machine learning, to help them compete.

Edge computing is also a defining factor in expansive data. There is likely to be greater activity at the edges—expansive data means more processing may be distributed across IoT networks. Data can be ingested, processed, and even stored within edge devices and systems, and, if it is deemed critical on an enterprise scale,



Thanks to the ubiquity of cloud-based services, from infrastructure to platform to applications, the power and capacity to support even bigger data environments are readily available.

moved to centralized data centers or clouds. Edge computing continues to extend its capabilities and encompasses a broad assortment of devices and systems that may require real-time interactions and responsiveness, including kiosks, autonomous cars and trucks, and sensors embedded across IoT.

With expansive data surging across all points of the enterprise, infrastructures could be quickly overwhelmed with ingestion, processing, and storage demands. Expansive data could also be expensive data without proper preparation. Fortunately, none of this is happening in a vacuum, and other developments may be helping organizations manage the challenge. Thanks to the ubiquity of cloud-based services, from infrastructure to platform to applications, the power and capacity to support even bigger data environments are readily available.

A new generation of database tools and platforms—led and enabled by machine-learning initiatives—is supporting the continuous, relentless data growth. Hadoop, the big data framework that made massive-scale data analytics a reality for every company that needs it, is beginning to show its age. While Hadoop was once seen as the single cure-all for big data challenges 10 years ago, today's expansive data calls for a variety of tools, platforms, and frameworks to help enterprises better manage their data. Nonetheless, the Hadoop Distributed File System can either support or be a part of data lake architectures, opening up a new mission for these environments.

According to a 2018 survey conducted by Unisphere Research, a division of Information Today, Inc., 44% of enterprises had Hadoop in production, which represents a downward shift from 2016, in which 55% reported using the framework ("2018 Next-Generation Data Deployment Strategies Report"). In addition, the survey found general satisfaction levels with Hadoop are mixed: Only 14% consider themselves to be "extremely satisfied" with Hadoop, while 64% are either dissatisfied or lukewarm toward the framework. While Hadoop provided one-of-a-kind functionality in its early days—such as parallel processing and management of a variety of data types—other technologies and solutions also now share these capabilities without the skill levels that Hadoop demands.

Predictably, the growth of expansive data is likely to track closely to that of IoT itself.

Accordingly, next-generation data technology initiatives represent new approaches to data management. The Unisphere Research survey found notable growth in the adoption of data lakes—places to store diverse datasets without having to build a model first. Their adoption continues to rise as data managers seek to develop ways to rapidly capture and store data from a multitude of sources in various formats. Overall, 38% of organizations are employing data lakes as part of their data architecture, up from 20% in a survey conducted 2 years prior. Another 15% said they were considering adoption. Data lakes are growing to impressive levels as well—close to one-third, 32%, support more than 100TB of data, the survey found.

With the relentless rise of IoT, AI, machine learning, and cloud-based services, enterprises are now challenged with accommodating and delivering value from the expansive data that surges through their systems. Data warehouses and Hadoop represented solutions for the pre-IoT, pre-AI enterprises. Today's opportunities and challenges call for the next generation of platforms and tools to bring it all together.

—Joe McKendrick

IRI Voracity
An Insatiable Appetite for Data

Machine Learning in Analytics and Anonymization

THE INCREASING NUMBER OF applications for machine learning testify to its ability to improve the speed and accuracy of informational assessment from ever larger sources of data. Users of the IRI Voracity data management platform can leverage two machine learning modules: one for predictive analytics, and another for protecting sensitive data. Many more are possible.

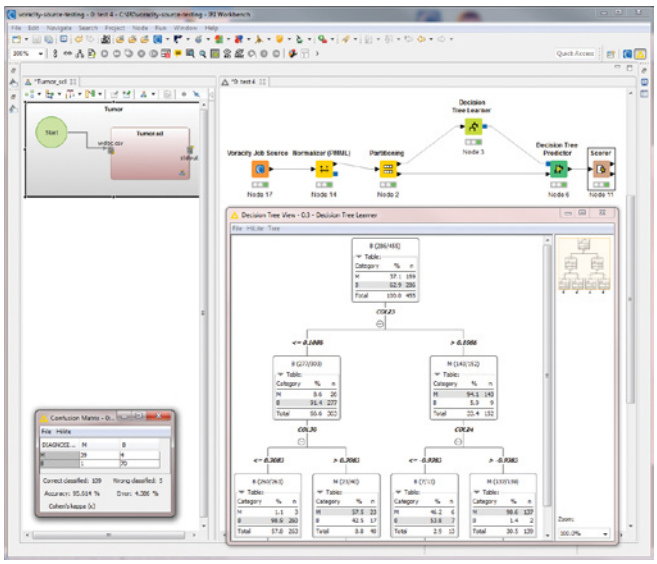
PREDICTING MALIGNANCIES

A common use of machine learning involves training a computer to evaluate data sets and create prediction models from trends in that data. Machine learning builds off traditional statistics and rapidly creates larger and more advanced models.

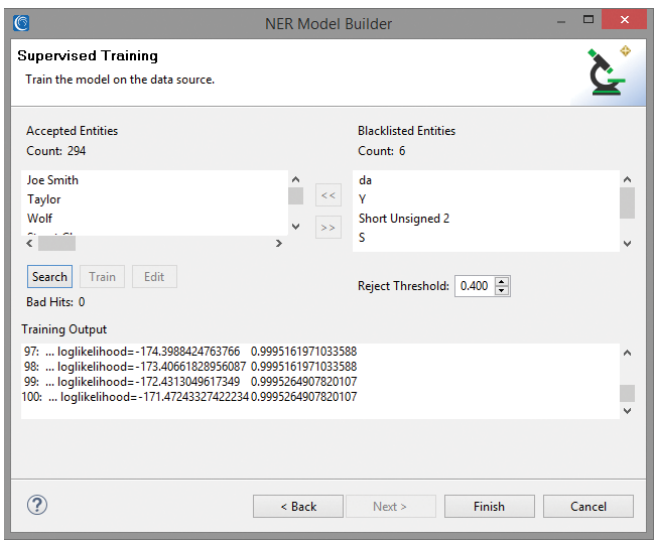
Many machine learning-related modules are included in KNIME, a popular open source data science platform that runs with Voracity in Eclipse. In this KNIME workflow, a Voracity data wrangling node feeds tumor measurement data into a KNIME decision tree node to improve breast cancer prediction accuracy:

Here, Voracity prepared raw data containing 20 different measurements of breast tumors, including their overall size, shape, and features of the cells' nuclei. Within seconds, the wrangled results flow into a decision tree to help determine if a tumor is likely to be malignant or benign.

The "Decision Tree Learner" node goes through different variables and creates multiple binary trees. Each tree determines if a given factor is likely to be a cause for a malignant tumor before it tries the next variable. Once the tree is built, the predictive model using those variables is tested for accuracy. In this case, it was about 95%, so the model should continue to be a reliable predictor future for data sets, too.



KNIME decision tree analysis of tumor data wrangled in IRI Voracity node



Using Machine Learning in IRI Workbench to train IRI DarkShield NER models

FINDING AND MASKING PII

Personally identifiable Information (PII) in documents like Word and PDF can be difficult to discover, delete, or de-identify en masse. This is particularly true when items like names or addresses — which do not match patterns or lookup values — can only be found in their Natural Language (Processing) context.

To address this challenge, IRI DarkShield technology in Voracity supports the use, and training, of Named Entity Recognition (NER) models to find those items. NER models are built from custom training data to improve search results.

The graphical front-end for DarkShield (and the rest of Voracity), called IRI Workbench, includes a wizard for either creating a NER model from existing annotated training data, or for training a model with actual documents that DarkShield parses. The latter employs a semi-supervised, iterative machine learning and annotation process.

Model training in this way improves the accuracy of PII search results when performed on a representative subset of documents. 15,000 sentences are considered a good minimum for teaching the machine to find named entities.

For more information email voracity@iri.com.

www.iri.com/voracity

INSIGHTS

Top Three Considerations Before Moving to a Multi-Cloud Model

By Pete Salamanca

Cloud has become the de facto standard for modern IT. Companies of all sizes and from virtually every industry have tapped the cloud to take advantage of the many benefits it offers—including cost reduction, scalability, and quick, easy provisioning. As of 2017, 86% of all data center workloads have been moved to the cloud, according to Statista. By 2021, the percentage is expected to grow to 94%.

Multi-Cloud as the Middle Ground

Cloud migration is picking up speed this year, even for those companies using long-relied-on Oracle applications and databases, according to a recent survey of 300 IT professionals from Apps Associates. Almost nine in 10 IT decision-making professionals said that their senior management mandated a cloud adoption strategy this year, and it was a focus for the company. Nearly six in 10 reported they already had a formal, documented plan in place for adoption, while an additional 36% had less formalized plans or are making decisions in phases. The truth of the matter is, the clear majority of these decision-making professionals have already made the move or are actively planning for cloud migrations.

Unfortunately, there continues to be confusion around cloud migration best practices. The crowded market landscape and frenzy over the "cloud wars" in the media often present contradictory points of view, making it difficult for these IT decision makers to determine the most efficient pathway for their specific company. In fact, more than half say that while they know cloud is the way forward, they aren't quite sure how to get there.

There are multiple cloud options to choose from—private, public, and multi-cloud—all of which are critical to business competitiveness. As the cloud landscape has grown and become more established, the process of selecting a cloud offering has gotten more closely tied to business goals and company preferences. Some businesses remain concerned about hosting their applications alongside thousands of others and elect to migrate to a private cloud to ensure a greater sense of control.

Many of these companies have spent years building their applications on-premise. These IT professionals typically appreciate the opportunity to be physically close to their key systems and applications and are not yet ready to move everything to the public cloud.

However, other enterprises have watched closely as cloud technology in public cloud offerings such as Amazon Web Services (AWS) has quickly advanced and matured. These decision makers often opt to select public cloud, because they believe the security precautions are as strong as what they might be able to provide in their own private cloud.

But whether public or private, over the past 5 years, it's become glaringly clear that any company not moving at least some of its critical applications from on-premise environments will be left struggling to compete, especially in areas such as efficiency and flexibility. The desire to leave some key applications on-premise is what's driving many companies to a multi-cloud model.

Multi-cloud can be defined simply. It's when a company elects to locate its environment in more than one cloud platform. Multi-cloud offers a unique advantage for companies that are ready to put some of their applications up with the large providers but still want to have some applications segregated from the masses.

Top Three Considerations Before Launching a Cloud Migration Journey

Multi-cloud offers many benefits—in terms of documented security, compliance, and savings outcomes—but every migration presents challenges. Understanding best practices around how to successfully launch a multi-cloud migration journey is crucial to avoiding the common pitfalls along the way.

Here are three elements IT decision makers should consider before starting the cloud migration journey:

1. Deeply Analyze the Environment

Create a holistic understanding of the company's IT stack by fully mapping out the existing IT environment. This will include analyzing the technology, applications, and databases and determining what each serves and what is currently in use or left nonactive. Often, IT decision makers may discover an opportunity to consolidate and improve the environment before even migrating. These consolidation efforts could be centered around reducing the number of employees managing information or removing unnecessary and unused environments. ▶

Multi-cloud offers a unique advantage for companies that are ready to put some of their applications up with the large providers but still want to have some applications segregated from the masses.

Pete Salamanca is VP of cloud services at Apps Associates (www.appsassociates.com), where his team specializes in migrations to both the public and private cloud, and offers expert support in Oracle application and databases, MS SQL Server, VMware, OVM, and all AWS offerings.

It's very typical in this phase for decision makers to find that many of their existing environments aren't being fully utilized, which is due to the rapidly evolving technological space and adoption of new platforms. Having a deep understanding of the environment helps create the foundation for a move to the cloud and enables IT to present information back to executives in their company, since not all may be sold on which cloud to move to.

2. Choose the Right Partner

IT decision makers are actively looking for partners to help them with the migration process.

When selecting a vendor to help with cloud migration, first ensure the partner has ample experience around migrating applications that are critical to the company's bottom line. Start by asking the right questions of potential vendors such as, "What is the average length of time for migrating a production Oracle workload to the cloud and how many iterations do you go through?" and, "How many similar workloads have you moved to the cloud?" This will help identify and measure the experience-based knowledge and technical understanding a vendor has.

Next, be sure to check customer references, a vital aspect to fully understanding vendors' migration successes and client relationships. With the cloud hype, many partners are just now jumping on the bandwagon to offer migration services they are not yet experts in. (For example, migrating Oracle applications is often an area not all partners will have in-depth expertise in.) Avoid the hype cycle by checking customer satisfaction ratings and ensure the company has some level of third-party validation. Finally, make sure the vendor can be available to act as an extension of the IT team, opening a space to collaborate while troubleshooting issues along the way and managing your environment beyond the initial migration. Creating a strong relationship at the start will result in lasting benefits.

3. Press the Permission Reset Button

Often when IT decision makers take a holistic view of the existing environment, they find the opportunity to evaluate certain privileges. As companies grow, evolve, and see faces come and go, privileges for who has access to certain business information and processes are often overlooked. The cloud migration journey is a convenient time to discuss who should have access to what and offers a chance to reset the permissions company-wide. When an environment is built within the cloud, there is the ability to use the "least privilege" rule upon setup, and allow and document access to only the individuals and teams that need that access.

Over the past 5 years, it's become glaringly clear that any company not moving at least some of its critical applications from on-premise environments will be left struggling to compete, especially in key areas such as efficiency and flexibility.

One Size Does Not Fit All: The Benefits of Multi-Cloud

By carefully considering these issues with a trusted advisor, IT leaders can identify which workloads need to be more agile and which need tighter governance and control. This process will help identify the right cloud choice. For business environments that require agility, scalability, and the ability to be deployed to regions outside the U.S. within minutes, the right choice is the public cloud. For business environments where IT leaders feel the need for a physical location or have a strong desire to keep a close eye on a custom application, the right next step is often a private cloud.

A one-size-fits-all approach to the cloud is no longer the norm. Multi-cloud strategies are highly attractive to IT decision makers looking to avoid vendor lock-in. This option is a good fit for those who are searching for the freedom to mix-and-match where applications live and want to benefit from cost savings. Secure connections between clouds allow the multi-cloud model to take even more shape with the ability to move workloads seamlessly between the two.

According to Gartner, by 2020, 75% of organizations will have deployed a multi-cloud or hybrid cloud model. Multi-cloud scenarios offer a good middle-ground option, allowing companies to evolve existing IT assets and effectively grow along with innovation and technological developments.

melissa®

How Data Quality and AI Support Big Data

Machine reasoning powers fraud detection, regulatory compliance, and CX

With the data deluge data managers face today, smarter options for managing, enhancing, and connecting data points need to find their way into enterprises' big data toolkits. Artificial intelligence (AI) capabilities are playing a new role in this effort. Initially developed to support advanced scientific research, tools like machine reasoning are powering up key operations that include identity verification, customer onboarding, and product-focused data quality. For the range of stakeholders in the enterprise data chain—compliance officers, database managers, marketers, and more—this supports a complex set of operations to help stem fraud and meet compliance regulations while also identifying opportunities and improving the customer experience. Data quality is at the core of these capabilities.

SMART DATA DISCOVERY

Customer data illustrates the big data challenge. Ideally, customer data should be aggregated into a single customer view (SCV), or a single source of truth, based on internal and external data, and unified across channels, locations, and business silos. Going beyond simple verification to provide a 360° SCV is useful for operations such as identity verification and fraud detection, as well as for sales/marketing or customization that meets specific needs. These use cases not only represent opportunities but also business risks, and call for higher quality data and smarter software. AI-enabled platforms and applications can automate data quality and improve interoperability, providing a shorter path to a 360° SCV.

UNDERSTANDING SEMANTIC TECHNOLOGY AND WHERE AI ADDS VALUE

As businesses digitize, they should also transform customer data to make it findable, accessible, interoperable, and reusable (FAIR). This can be achieved by using AI-enabled semantics and machine reasoning to normalize, harmonize, and richly connect information. Semantic technology is an extension of the current web that has been defined over the past decade by the World Wide Web Consortium (W3C) in collaboration with Stanford University, MIT, and others, and stands out in its design which facilitates universal data interoperability. In addition, semantic technology delivers a form of AI that associates words with meanings and recognizes relationships between them. This allows the business to tap into greater understanding of its customers by making powerful, real-time connections among the data in their records.

THE ROLE OF MACHINE REASONING

Machine reasoning applies knowledge about a subject of interest captured in the form of ontologies— formal descriptions of classes, entities, and relationships that are relevant to computing in a specific area of interest. This promotes context and reasoning, or the ability to make inferences. For example, machine reasoning helps to enable properly validated identities as well as broader data quality and integrity. AI-enabled machine reasoning can also quickly open up opportunities for rich pattern recognition and responses. By taking all available data into account, the process reduces false positives—problematic situations that slow down business and have the potential to annoy customers.

Semantically-enabled software platforms also complement machine learning, a more commonly known form of AI. Machine learning applies supervised and unsupervised algorithms to analyze training datasets. Its goal is to identify features that are, or seem to be, related to outcomes. This powers the generation of inductive hypotheses and facilitates pattern identification used for decision support. One of the primary values of semantic technology is that it can update and deliver new information as soon as additional data is entered into the system, allowing the enterprise to quickly redefine concepts and relationships.

SEMANTIC LAYER IMPROVES CURRENT AND LONG-TERM DATA VALUE

The combination of machine reasoning, machine learning, and traditional data quality tools makes it possible to deliver a flexible, interoperable semantic integration layer on top of enterprise systems. This supports improved data quality and interoperability, the ability to use these technologies to identify and align data to different standards, and finally, to apply advanced AI-based pattern recognition applications out of the box.

IT'S ALL ABOUT THE DATA

Semantically-enabled machine reasoning is an efficient form of AI that can help with essential enterprise requirements such as data quality and completeness, and can scale to provide automated pattern recognition for decision support in mission-critical applications. In this very powerful new era, errors are reduced, data insights are more sophisticated and quickly gleaned, and staff is freed to focus on excellent service, new product development, and overall business growth. However, it is critical to note that semantic technology is only effective with accurate data.

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

BIG DATA BY

MODERNIZING TO DELIVER DATA-DRIVEN CAPABILITIES

IT executives and line-of-business experts understand the importance of data for their success and are adopting modern technologies to enable the delivery of timely data and insights to enhance decision making. In line with their data-driven goals, organizations are leveraging hybrid and multi-cloud strategies. However, they are also finding that cloud approaches add their own complexity.

THERE ARE NUMEROUS BUSINESS OBJECTIVES THAT ARE DRIVING DATA STRATEGIES, BUT THE MOST OFTEN MENTIONED ARE IMPROVING THE DECISION MAKING OF END USERS, AND UNCOVERING CUSTOMER PREFERENCES AND PATTERNS.

Top 5 Business Uses Driving the Data Strategy

1. To inform decision making
2. To understand customers and trends
3. To improve internal operations
4. To provide smarter services and products
5. To support a better customer experience

ADOPTION OF MACHINE LEARNING HAS ALMOST DOUBLED OVER THE PAST YEAR, WITH USE CASES INCLUDING BOTH INTERNAL EFFICIENCIES AND BUSINESS GROWTH INITIATIVES. HOWEVER, CHALLENGES WITH ACCESS TO THE RIGHT DATA AND A LACK OF OPERATIONAL AUTOMATION HAVE BEEN SEEN.

Top 5 Business Drivers for Machine Learning Projects

1. Operational improvements
2. Security and risk
3. Customer retention
4. Revenue growth
5. Accelerate innovation

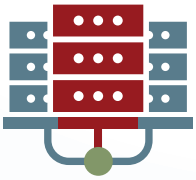
Source: "Profiling the Data-Driven Business, 2019," produced by Unisphere Research, and sponsored by Pythian

IMPLEMENTING A SUCCESSFUL HYBRID CLOUD STRATEGY IS CHALLENGING FOR ORGANIZATIONS.

Key problems include:

1. A lack of consistency between cloud platforms
2. Need for additional training and skill sets
3. Difficulty in porting data and applications

86% of enterprises are considering or have already "repatriated" one or more workloads from the public cloud back to the data center, showing that many are still in the early phase of cloud adoption.



Source: IDC white paper: "Benefits of the Consistent Hybrid Cloud: A Total Cost of Ownership Analysis of the Dell Technologies Cloud," sponsored by Dell EMC

THE NUMBERS

There are not enough on-premise resources to keep up with the growth of data management requirements. This means that cloud services are increasingly being tapped as a vital resource in the data manager's toolkit.

For their latest database projects, respondents to a survey were close to evenly split between deploying on cloud or on-premise, with a bit more emphasis on on-prem.

44% deployed their projects in the public cloud or as part of a hybrid architecture.

52% indicated their most recent database project involved an on-premise implementation.

A combination of scalability, cost, and maintenance benefits are driving public cloud deployments.

Greater scalability is the **#1 advantage.**

Lower cost is seen as the **#2 advantage.**

Reduced need for infrastructure maintenance **comes in at #3.**

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

Organizations are strongly positive about the cloud, but also concerned about the slow pace of analytics adoption in the cloud.

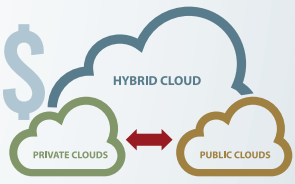
83% agree that public cloud is the best place to run analytics.

91% say analytics should be moving to the cloud faster.

69% want to run all of their analytics in the cloud by 2023.

Source: "The State of Analytics in the Cloud," conducted by Vanson Bourne on behalf of Teradata

Hybrid cloud, using a combination of public and private, is the focus for many enterprises, and along with that, they are keenly focused on cost optimization.



- **Organizations leverage almost 5 clouds on average:** Respondents are already running applications in a combination of 3.4 public and private clouds, and experimenting with 1.5 more for a total of 4.9 clouds

- **The #1 priority in 2019 is cloud cost optimization.** Optimizing existing cloud use is the top initiative in 2019 for the third year in a row, increasing to 64% from 58% in 2018.

Source: "RightScale 2019 State of the Cloud Report" from Flexera

A NEW GENERATION OF TOOLS is making it possible to leverage the wealth of data flowing into organizations from a previously unimaginable range of data sources. Machine learning, AI, Spark, and object storage are just some of the next-generation approaches gaining traction, according to recent surveys conducted by Unisphere Research, a division of Information Today, Inc.

But, it is also increasingly clear that there is no single way to approach data-driven innovation today. Open source-based technologies have gained strong adoption in organizations alongside proprietary offerings; data lakes are increasingly being implemented; data warehouses continue in widespread use; and hybrid deployments spanning cloud and on-premise are commonly accepted.

Organizations are seeking to use data-driven innovation for better reporting and analytics, real-time decision making, enhanced customer experience and personalization, and reduced costs. But with data coming in from more places than ever, being stored in more systems, and being leveraged for a wider array of use cases, there is greater recognition that security and governance must be addressed intelligently.

Evaluating new and disruptive technologies, and then identifying how and where they can be useful, can be challenging.

To contribute to the discussion each year, *Big Data Quarterly* presents the "Big Data 50," a list of forward-thinking companies that are working to expand what's possible in terms of capturing, storing, protecting, and deriving value from data.

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

In addition, following the list, under the Trailblazers header, executives share their perspectives on their companies' unique approaches to driving innovation.

Accenture
www.accenture.com

A global professional services company, providing a broad range of services and solutions in strategy, consulting, digital, technology, and operations, Accenture combines experience and specialized skills across more than 40 industries and all business functions.

Actian
www.actian.com

A hybrid data management, analytics, and integration company, Actian delivers data as a competitive advantage to thousands of customers worldwide to help ensure that business-critical systems can transact and integrate at their very best on-premise, in the cloud, or both.

Aerospike
www.aerospike.com

Powered by a Hybrid Memory Architecture and autonomic cluster management, Aerospike is deployed in the financial services, banking, teleco, technology, retail, ecommerce, ad tech, martech, and gaming industries and is well-suited for applications that require extreme uptime, performance, and scalability.

Alluxio
www.alluxio.io

The developer of open source data orchestration software for the cloud, Alluxio aims to move data closer to big data and machine learning compute frameworks across clusters, regions, clouds, and countries, providing memory-speed data access to files and objects.

Amazon Web Services
https://aws.amazon.com

A subsidiary of Amazon.com, Amazon Web Services is a comprehensive cloud platform, spanning 69 availability zones within 22 geographic regions around the world, with announced plans for nine more availability zones and three more regions in Cape Town, Jakarta, and Milan.

Arcadia Data
www.arcadiadata.com

Founded with the mission to connect business users to big data for powerful insights, Arcadia's enterprise provides AI-driven, in-data-lake analytics and BI software that runs natively within modern data platforms and can analyze large volumes of data without moving it.

Attunity, a Division of Qlik
www.attunity.com

A leader in modern data integration, enabling enterprises to employ DataOps for analytics for better business outcomes, Attunity provides a platform that accelerates the discovery and availability of analytics-ready data by automating real-time data streaming, refinement, cataloging, and publishing.

Cambridge Semantics
www.cambridgesemantics.com

A modern data management and enterprise analytics software company, Cambridge Semantics helps enable seamless access, integration, and analysis of all enterprise data via a graph-driven data fabric architecture to allow IT departments and business users to accelerate data delivery.

Cloudera
www.cloudera.com

Delivering an enterprise data cloud for any data, anywhere, Cloudera emphasizes equivalent functionality on- and off-premise with support for hybrid and multi-cloud; application of multiple analytic functions; emphasis on security and compliance; and support for open compute architectures and open data stores.

CloverDX
www.cloverdx.com

Engineered to solve complex data scenarios, CloverDX provides a data integration platform for designing, automating, and operating data jobs at scale, whether migrating data from a legacy system to the cloud, or ingesting information from multiple sources into a BI system.

Collibra
www.collibra.com

Offering a cross-organizational data governance platform, Collibra aims to break down traditional data silos and open up organizational data so all users can find the data they need, collaborate on it, and easily understand its meaning.

Couchbase
www.couchbase.com

With a database architected on top of an open source foundation, Couchbase aims to combine the best of NoSQL with the power and familiarity of SQL, all in a single platform spanning from the cloud to the edge.

Databricks
https://databricks.com

Founded by the original creators of Apache Spark, Databricks provides a Unified Analytics Platform for data science teams to collaborate with data engineering and lines of business to build data products, enabling faster time-to-value through analytic workflows.

DataKitchen
www.datakitchen.io

Incorporating agile software development, DevOps, and manufacturing-based statistical process control into analytics and data management, DataKitchen provides a DataOps platform for data-driven enterprises, enabling them to support data analytics that can be adapted to meet evolving requirements.

DataStax
www.datastax.com

Delivering an always-on, active-everywhere hybrid cloud database built on Apache Cassandra, DataStax helps organizations seamlessly build and deploy modern applications in hybrid cloud scenarios, and also offers managed services provided by experts at handling enterprise applications at cloud scale.

Denodo
www.denodo.com

A provider of data virtualization software, Denodo helps enable high-performance data integration, data abstraction, and real-time data services across enterprise, cloud, big data, and unstructured data sources so that customers can have faster and easier access to unified business information.

Dremio
www.dremio.com

Known as the "data-as-a-service" platform company, Dremio—which was created by veterans of open source and big data technologies, and the co-creators of Apache Arrow—provides a data analytics approach that helps companies get more value from their data.

erwin
https://erwin.com

Known for years for its data modeling software, erwin has evolved through acquisitions and R&D to provide a data governance software platform with integrated capabilities for enterprise modeling, data cataloging, and data literacy that facilitates collaboration between IT and the business.

Franz
https://franz.com

An early innovator in AI and a supplier of graph and document database technology for knowledge graphs, Franz's technology is used by leading organizations worldwide in customer service, healthcare, life sciences, publishing, and more.

Google Cloud
https://cloud.google.com

Google Cloud Platform is a suite of cloud computing services that are offered on the same infrastructure that allows Google to return search results in milliseconds, serve billions of hours of YouTube videos per month, and provide storage for 1 billion Gmail users.

GridGain
www.gridgain.com

Powering the digital enterprise, GridGain offers an in-memory computing platform built on Apache Ignite that provides in-memory speed and massive scalability for data-intensive applications and can be deployed on-premise, on a public or private cloud, or on a hybrid environment.

Hewlett Packard Enterprise(HPE)
www.hpe.com

A global technology leader focused on developing solutions that allow customers to capture, analyze, and act upon data from edge to cloud, HPE recently acquired MapR to accelerate its Intelligent Data Platform capabilities and help customers optimize solutions for mission-critical big data workflows.

HVR
www.hvr-software.com

Providing real-time cloud data replication to support enterprise modernization efforts, the HVR platform optimizes the extraction of data from on-premise and cloud source databases to target data platforms, including large data volumes in complex environments, with real-time data updates, access, and analysis.

IBM (International Business Machines)
www.ibm.com

One of the largest IT companies in the world, with innovative big data solutions and services spanning AI and machine learning, cloud, blockchain, Hadoop, and Spark, IBM earlier this year acquired its longtime partner, Red Hat, for $34 billion.

InfluxData
www.influxdata.com

The creator of InfluxDB, the open source time series database, InfluxData offers technology that is purpose-built to handle the massive volumes of time-stamped data produced by IoT devices, applications, networks, containers, and computers.

Informatica
www.informatica.com

A provider of enterprise cloud data management software, Informatica accelerates data-driven digital transformation—including journeying to the cloud, reimagining data governance and compliance, delivering intelligent analytics insights, and unleashing Customer 360 engagement—so that organizations can realize new growth opportunities.

MariaDB
www.mariadb.com

Deployed in minutes for transactional, analytical, or hybrid use cases, MariaDB delivers operational agility without sacrificing key enterprise features, including ACID compliance and full SQL, using pluggable, purpose-built storage engines to support workloads that previously required a variety of specialized databases.

MarkLogic
www.marklogic.com

The MarkLogic Data Hub is a highly differentiated data platform that eliminates friction at every step of the data integration process, enabling organizations to achieve a 360-degree view faster than ever.

Melissa
www.melissa.com

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

Microsoft
www.microsoft.com

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

MongoDB
www.mongodb.com

Founded in 2007 by Dwight Merriman, Eliot Horowitz, and Kevin Ryan—the team behind DoubleClick—MongoDB provides a general-purpose, document-based, distributed database built for modern application developers and for the cloud era that supports ad hoc queries, indexing, and real-time aggregation.

Ontotext
www.ontotext.com

Ontotext, whose driving force is to put enterprises in control of their knowledge assets, provides the Ontotext Platform, a cognitive content analytics technology, and Ontotext GraphDB, a graph database for metadata management and cognitive analytics.

Oracle
www.oracle.com

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

Pure Storage
www.purestorage.com

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

Quest Software
www.quest.com

Quest provides software solutions that help simplify challenges in the rapidly changing world of enterprise IT caused by high data volumes, cloud expansion, hybrid data centers, security threats, and stringent regulatory requirements.

Redis
https://redis.io

Redis Labs is the home of Redis, an in-memory database, and commercial provider of Redis Enterprise that delivers performance, reliability, and flexibility for high-speed transactions, recommendation engines, data ingestion, fraud mitigation, real-time indexing, session management, and caching.

Reltio
www.reltio.com

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

SAP
www.sap.com

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

SAS Institute
www.sas.com

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

SnapLogic
www.snaplogic.com

SnapLogic's AI-powered workflows and self-service integration capabilities help make it faster and easier for organizations to manage all their application integration, data integration, and data engineering projects on a single, scalable platform.

SQream
https://sqream.com

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

Software AG
www.softwareag.com

An independent integration, IoT, analytics, process software, and services company, Software AG drives enterprise innovation by helping to connect and integrate everything—from applications and devices to data and clouds—and is trusted by more than 70% of the world's top 1,000 enterprises.

Striim
www.striim.com

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

Syniti
www.syniti.com

Syniti, formerly BackOffice Associates, was founded to solve complex data challenges, bringing synergy between data and business, delivering confidence, and enabling progress along clients' business transformation journeys through a combination of data expertise, services, and intelligent software leveraging AI and machine learning.

Tamr
www.tamr.com

The company's patented software system uses machine learning supplemented by human expertise and rules to unify and prepare data across silos, enabling previously unavailable business-changing insights and transforming how companies get value from their data.

Teradata
www.teradata.com

Teradata transforms how businesses work and people live through the power of data by leveraging all of the data, all of the time, to allow users to analyze anything, deploy anywhere, and deliver analytics that matter.

TigerGraph
www.tigergraph.com

Based on native and parallel graph technology, TigerGraph unleashes the power of interconnected data and applications, supporting applications such as fraud detection, Customer 360, MDM, IoT, AI, and machine learning to help organizations make sense of ever-changing big data.

Trifacta
www.trifacta.com

Drawing on decades of experience in academic research in machine learning and data visualization, Trifacta helps make the process of preparing data faster and more intuitive so users can support a variety of analytics and operational use cases.

VMware
www.vmware.com

Streamlining the journey for organizations to become digital businesses that deliver better experiences to their customers and empower employees to do their best work, VMware's software spans compute, cloud, networking and security, and the digital workspace.

VoltDB
www.voltdb.com

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

Waterline Data
www.waterlinedata.com

Waterline Data automates data discovery, compliance, and the ability to take action on data by using a combination of AI, machine learning, ratings and reviews, and tribal knowledge to deliver an AI-driven data catalog.

Yellowbrick Data
https://yellowbrick.com

Built for enterprises and the hybrid cloud, Yellowbrick Data provides the Yellowbrick Data Warehouse which deploys powerful analytics anywhere with compelling economics, empowering companies to make faster decisions with their data.

Aerospike

Srini Srinivasan, Chief Product Officer and Founder

IN WORKING WITH CUSTOMERS across industries around the globe, I've found they each have critical real-time business moments that can make or break their customer experience.

Data ingestion rates are increasing so rapidly that it has become impossible to make the best possible decision using a traditional system architecture combining a cache with an operational database. Aerospike is a high-performance database that runs as fast as a cache and supports real-time applications at extremely high scale.

Unlike legacy NoSQL databases, Aerospike's unique Hybrid Memory Architecture unlocks the full potential of modern hardware and entirely eliminates the friction that prevents companies from delivering unimaginable value from vast amounts of data at the edge, in the core, and on the cloud.

Aerospike powers hyperscale data solutions that give customers what they call an unbreakable competitive advantage. They save millions with a smaller, simpler-to-manage footprint. As a result, they can free up resources to make investments in important areas like IoT, machine learning, and artificial intelligence.

Aerospike enterprises are some of the boldest and most innovative companies in the world. Aerospike's vision is to make it easy and affordable for companies of all sizes to build next-generation hyperscale data solutions similar to those built internally by the largest internet-scale companies like Google and Facebook.

We are proud to have helped companies like Adobe, Airtel, FlipKart, Kayak, Nielsen, Nokia, PayPal, and Snap meet their customer needs in the moments that matter.

FOR MORE INFORMATION, visit Aerospike.com.

Aerospike
www.aerospike.com

Alluxio

Steven Mih, CEO

TODAY WE SEE MORE ENTERPRISE architectures shifting to hybrid and multi-cloud environments. And while this shift allows for more flexibility and agility, it also means having to separate compute from storage, creating new challenges in how data needs to be managed and orchestrated across frameworks, clouds, and storage systems.

To address these data challenges, enterprises are adopting a new platform: a data orchestration platform. A unified data orchestration platform simplifies your data's cloud journey and fundamentally enables separation of storage and compute. It brings speed and agility to analytic and AI workloads, reduces costs by eliminating data duplication, and enables users to move to newer storage solutions like object stores.

Proven at global web scale in production for modern data services, Alluxio is the developer of open source data orchestration software for the cloud. Founded at UC Berkeley's AMPLab by the creators of the Tachyon open source project, it is a compute agnostic, storage agnostic and cloud agnostic solution for analytic and machine learning applications.

Alluxio moves data closer to big data and machine learning compute frameworks in any cloud across clusters, regions, clouds and countries, providing memory-speed data access to files and objects. Intelligent data tiering and data management deliver consistent high performance to customers across all industries.

LEARN MORE at https://www.alluxio.io/data-orchestration/

Alluxio
www.alluxio.io

Denodo Technologies

Ravi Shankar, Senior VP and CMO

THE BIGGEST CHALLENGES WITH big data are not around storage, but around leveraging the data for an increasingly demanding set of use cases. These include data science, which aims to discover insightful business patterns through highly descriptive data models, as well as AI and machine learning, which have enormous potential with today's advanced processing capabilities.

Big data implementations are often unable to adequately support such use cases, and ironically, this is due to the ease with which data can be stored; with schema-on-read capabilities, different departments can quickly and easily store data in a variety of different formats, creating silos.

Data virtualization serves as the critical big data fabric that knits together the disparate domains of a big data implementation, in real time. Rather than replicating data to a new, centralized repository, data virtualization provides real-time views into the data, no matter the format and no matter where it is physically stored. It acts as a single, unified access layer across the entire big data implementation. And because it can deliver views of the data across all of the standard interfaces, it can support myriad applications.

The award-winning Denodo Platform offers the most advanced data virtualization capabilities available for leveraging big data in the service of data science, AI, machine learning. It enables data scientists to work in R and Python in addition to SQL, and it employs machine learning to ensure that data scientists have access to the most actionable intelligence. Denodo Platform 7.0 takes this further with the integration of Zeppelin Notebook, facilitating the sharing of data models.

Denodo Technologies
www.denodo.com

erwin, Inc.

Adam Famularo, CEO

DATA VISIBILITY, GOVERNANCE & VALUE

Knowing what data you have, where it lives and where it came from is complicated.

The lack of visibility and control around "data at rest" combined with "data in motion," as well as difficulties with legacy architectures, means organizations spend more time trying to find the data they need rather than using it to produce meaningful business outcomes.

As adoption for data governance grows, best-in-breed enterprises are looking for ways to use their data for competitive advantage. These organizations are evolving their data governance practices to data intelligence—connecting all the pieces of their data management and data governance lifecycles to create actionable insights.

With erwin's approach to data governance, organizations can discover, understand, govern and socialize mission-critical information. And because many of the processes are automated, both errors and reliance on technical resources are reduced while the speed and quality of the data pipeline increases.

erwin is the only software provider with a complete, metadata-driven approach to data governance. The erwin EDGE platform integrates enterprise modeling and data intelligence suites to create an "enterprise data governance experience" so customers can:

- Understand their business, technology and data architectures and the relationships between them
- Create and automate a curated enterprise data catalog, complete with physical assets, data models, data movement, data quality and on-demand lineage
- Increase data literacy with agile, well-governed data preparation and integrated business glossaries and data dictionaries that provide business context

With erwin, you can stop wasting time on data discovery and start using it to produce real value.

erwin, Inc.
www.erwin.com

Franz Inc.

Jan Aasman, CEO

AI KNOWLEDGE GRAPH SOLUTIONS: FRANZ INC.

Gartner recently identified Knowledge Graphs as a key new technology in both their Hype Cycle for Artificial Intelligence and Hype Cycle for Emerging Technologies. Using AI to create "Enterprise Knowledge" and link it across the Enterprise to create a "Knowledge Graph" is a key differentiator for companies in an increasingly competitive landscape. Semantic Graph databases, such as AllegroGraph, provide the core technology environment to enrich and contextualize the understanding of data. The ability to rapidly integrate new knowledge is the crux of the Knowledge Graph and depends entirely on Semantic Graph technologies.

ALLEGROGRAPH

AllegroGraph is a multi-model (Graph and Document) database technology that enables businesses to extract sophisticated decision insights and predictive analytics from highly complex, distributed data. AllegroGraph employs graph technologies that process data with contextual and conceptual intelligence and significantly enhances the document database model with its native support for JSON and JSON-LD. Knowledge Graphs can leverage JSON-LD to swiftly integrate with web-based applications. Organizations can therefore link specific information in their internal Knowledge Graphs (e.g., pertaining to customers or products) to web applications for timely action such as recommendations.

KNOWLEDGE GRAPH DEVELOPMENT

Franz provides a variety of services as part of its Knowledge Graph solution, from architectural consulting and technical seminars to training. If you really want to develop your corporate Knowledge Graph and address complex AI problems, you need a data system that goes beyond just data. You have to create a system that can link to anything outside your own predefined parameters—and that can learn from previous experiences. That is where a Semantic Graph database, like AllegroGraph, comes into the picture.

Franz Inc.
https://franz.com/

HVR

Anthony Brooks-Williams, CEO

HVR IS THE LEADING independent provider of a real-time cloud data replication solution that supports enterprise modernization efforts. The HVR platform is a reliable, secure, and scalable way to quickly and efficiently integrate large data volumes in complex environments, enabling real-time data access and analysis.

By 2022, 90% of corporate strategies will explicitly mention "information" as a critical enterprise asset, and "analytics" as an essential competency, according to a 2019 Gartner report. As more companies look to cloud-based technologies as an efficient, cost-effective method of maximizing data use, they must address challenges associated with increasing complexity in data management as well as concerns about efficiency and security.

HVR's real-time data replication technology is a scalable solution that allows data-driven organizations to harness the power of their data. With its log-based change data capture functionality, HVR enables efficient high-volume data streaming between multiple on-premise systems such as Oracle, SQL Server, SAP HANA, and Db2 z/OS, and modern cloud-based systems such as AWS, Azure, Kafka, and Google Cloud. Broad and heterogeneous platform support gives HVR customers freedom in their selection of data technologies in order to best address their unique needs.

HVR is deployed in a distributed architecture, which is flexible and scalable, and can support a wide variety of use cases.

Additional benefits include:

- Accuracy: Compare, HVR's data validation feature, ensures data correctness
- Security: In-flight data is compressed and encrypted as data from hundreds of on-premise and cloud databases is replicated without creating firewall vulnerabilities
- Low-latency: HVR's proprietary compression and log-based CDC capabilities allow for fast and non-intrusive data replication

HVR is committed to providing organizations with the ability to make informed business decisions using the freshest data possible. With a team that boasts over 200 combined years of data replication and integration experience, customers who choose HVR get a solution, not a tool.

FOR MORE INFORMATION, visit www.hvr-software.com.

HVR Software
www.hvr-software.com

melissa

A DATA QUALITY TOOLBOX KEEPS BIG DATA NICE AND TIDY

Big data can be one big mess. It's a challenge to handle the vast number of data sources that can be brought together, coupled with the sheer speed and volume of incoming data. You need data collection processes to standardize, format and validate data for big data analytics.

That's where Melissa comes in. We provide the Data Quality Toolset you need to augment your data capture, automate your cleansing tasks, and alleviate manual data prep tasks.

ESSENTIAL TOOLS FOR YOUR DATA QUALITY TOOLBOX

Melissa offers a full spectrum of data quality solutions available as on-prem and Web APIs and components for leading DI platforms, including Talend, Pentaho, and Microsoft SQL Server Integration Services (SSIS). These tools include:

- Data Profiling
- Data Parsing
- Data Cleansing
- Address Verification
- Identity Verification
- Identity Matching
- Fuzzy Matching & Survivorship
- Location Intelligence

PUTTING IT ALL TOGETHER—UNISON

Our newest product, Unison, combines many of Melissa's data cleansing solutions in one locally-hosted platform that requires no programming. It's one, centralized hub for creating simple-to-complex data quality tasks with the ability to connect to multiple RDBMS platforms, schedule jobs, visualize reports/analytics, collaborate on projects, and much more.

It's the ideal solution to cleanse sensitive customer information securely and safely—data never leaves your organization.

With the addition of Unison and new identity, document verification and biometrics capabilities, Melissa remains committed to delivering leading-edge solutions that meet every organization's need for clean, accurate data and a trusted single customer view—from enterprise to mid-market and SMB. For over 35 years, Melissa has helped 10,000 customers across multiple sectors and around the world turn data into insight. With offices in the UK, Germany, India and now Singapore, we're ready for the next 35 years, and more.

melissa
www.melissa.com

Syniti

Giacomo Lorenzin, Managing Director, Direct Solutions

THE INFINITE POTENTIAL OF DATA REPLICATION

Consider the concept of data replication. In its simplest form, it is the process of copying data. Now let's look at **real-time data replication**. This is not merely the act of making a copy. It is instantaneous, selective copying of only changed data to reduce impact on key systems and business processes—a critical capability in today's data-driven business cycles.

Production systems should function flawlessly, while behind the scenes data is moving where it is needed. A low-touch, versatile solution like Syniti Data Replication can support a consistent and trusted copy of your business data ready to be used by other applications without impacting the responsiveness of your business-critical systems. Those source and target systems could be relational databases, message queues, or big data platforms pretty much anywhere in the cloud or on premise!

Your mission-critical data environment probably consists of multiple different systems that share data to serve users and customers. Those systems could range from large legacy relational databases through nimble application-specific databases to data storage and data analytics in the cloud. Large nightly batch copies are no longer adequate for today's demanding consumers. Syniti Data Replication can keep the necessary data in sync so an amazing user experience is not dependent on time of day.

Fresh, relevant data assures well-informed and successful decision-making. Gone are the days when outdated data could result in error-prone reports and potentially disastrous decisions.

TAKE ADVANTAGE of the Infinite Potential of Data Replication to get the job done. We can help. www.syniti.com

Syniti
www.syniti.com

Yellowbrick Data

YELLOWBRICK DATA provides an optimized, integrated data warehouse solution capable of scaling from terabytes to petabytes to deliver the very best data analytics experience with flexible deployment options—on-premises, managed service, PaaS or anywhere in-between.

As companies strive to create value and bolster competitiveness using real-time business analytics, data warehouses are becoming overloaded with data stored for future analysis. Most of today's data warehouses cannot support the ingest of data in real-time or interactive applications and are not built to retain the massive amounts of historical data. They are also plagued with rising support costs and escalating use expenses.

Yellowbrick Data runs analytics 10x to 100x faster to achieve superior analytic insights at a fraction of the cost of other data warehouses. It allows enterprises to shrink their data warehouse footprint by as much as 97%, while saving millions in operational and management costs. It also lowers TCO by 4x or more and delivers ultra-fast, predictable performance for mixed cloud/on-premises workloads.

Yellowbrick Data's high-speed data warehouse is an optimized, integrated data warehouse solution that can easily scale from terabytes to petabytes. Featuring non-volatile memory (NVMe) flash drives, multi-core processors, high-performance networks, and a direct-to-cache design, Yellowbrick Data's massively parallel processing (MPP) hardware enables its sophisticated SQL query engine and integrated storage manager to deliver exceptionally low data access latencies. And it does so without requiring extensive tuning, carefully constructed schemas, or constant monitoring by highly skilled, in-house experts.

Enterprises across a variety of industries, including insurance, banking, telecommunications, retail, healthcare, and entertainment, rely on Yellowbrick Data to power their real-time data warehouses.

Yellowbrick Data
https://yellowbrick.com

AEROSPIKE
ALLUXIO
denodo
erwin
FRANZ INC.
HVR
melissa
Syniti
Yellowbrick

INSIGHTS

Survival of the Fittest: How the Financial Services Sector is Leading the Pack With Cloud Analytics

By Rachel Dines

The financial services sector has been one of the earliest proponents of the notion that every company is a tech company.

"Survival of the fittest" is one of the most basic principles of evolution. However, the concept also applies to the world of business. As companies march forward on their digital transformation journey—leveraging emerging technologies and digitizing standard processes for maximum advantage—market competition has become fiercer than ever before, and within crowded industries, players vie for consumer attention.

One such industry leading the pack in business evolution is financial services.

While on the surface Wall Street may appear conservative and risk-averse, when it comes to IT, the financial services industry has continually led the adoption of new technologies—sometimes out of a drive for innovation, and other times out of necessity.

So, how do financial services providers differentiate themselves? Often, they offer more or less the same core services and, as a result, must look for every added opportunity to gain a competitive edge. This need to differentiate creates enormous pressure to innovate and adopt new service offerings. As companies clamor for an outlet to explore new service options, the cloud has emerged as a welcome solution.

The financial services sector has been one of the earliest proponents of the notion that every company is a tech company. As a result, the industry has always remained at the forefront of adopting the latest and greatest tech. In a similar vein, it has also been among the first to utilize the latest cloud services.

According to Amazon Web Services (AWS) spending trends analyzed by CloudHealth by VMware, in 2018, 69% of the total cloud spend from financial institutions was on compute infrastructure. As of January 2019, that percentage dropped to 55%.

The reason? Financial services companies are shifting away from treating the cloud as just another data center with traditional virtual machines, storage, and networks. Instead, they are taking advantage of its latest offerings such as containers, serverless architectures, and analytics to help fuel innovation. This shift—from Cloud 1.0 to Cloud 2.0—enables organizations to truly unlock the benefits that the cloud promises: agility, elasticity, and scalability. While all these new cloud services had significant growth in the financial services industry, none saw more than analytics.

Cloud and Compliance Go Head-to-Head

Financial services institutions—which are already heavily reliant on analytics—are realizing they can save time, money, and resources by utilizing the big data, analytics, and machine learning services offered by cloud providers. Today, services such as Amazon Redshift, Elasticsearch, Elastic MapReduce, and Kinesis are extremely popular among financial services institutions, with adoption growing rapidly between 2018 and 2019. This wasn't always the case, however.

Historically, the stringent regulations enforced in the finance sector and other highly regulated industries, such as healthcare and insurance, have proved challenging for cloud providers. Without proper security built in, cloud computing in the finance sector was often limited to supporting auxiliary business processes for customer service, admin systems, and human resources. While these types of processes can certainly benefit from cloud analytics, financial services organizations were in need of a more complete shift to the cloud to fully reap the benefits.

Today, mature cloud providers have enhanced built-in security features that meet and often exceed regulatory requirements. With capabilities such as real-time threa detection and automated security and compliance reporting, cloud users are equipped with tools to easily comply with the long laundry list of financial regulations.

Pioneering a New Way Forward

Today, the financial services industry serves as a prime example of how to unleash the true value of data analytics in the cloud. Cloud transformation is rarely straightforward. Organizations embarking on ▶

Rachel Dines is head of product marketing at CloudHealth by VMware (www.cloudhealthtech.com), a cloud management platform designed to drive increasing business value at every stage of the cloud journey.

Financial services companies are shifting away from treating the cloud as just another data center with traditional virtual machines, storage, and networks, and instead are taking advantage of its latest offerings to help fuel innovation.

this journey can learn a lot from the financial services industry. The following are key steps to take to ensure that you get the most out of the latest technologies.

1. Establish a data-driven culture.

Businesses across all industries are abuzz about building a "data-driven culture," yet many struggle to define what that actually means for their business and how to make it a reality. According to Forrester, 74% of businesses say they want to be data-driven, but only 29% say they're good at connecting analytics to action. It's clear there's a massive cultural deficiency surrounding data.

Simply put, being data-driven means efficiently and regularly leveraging data to make better business decisions. While deploying the right tools is a great place to start, this is only the beginning. It's about how you take action on that insight. Business and technology leaders need to create a culture that empowers everyone within an organization to become their own data master. This means they need to arm all personnel with the resources and training to effectively collect, collate, and act on the data gleaned from day-to-day business functions. Easy access to cloud-based data analytics services significantly lowers the barrier to entry.

2. Identify the right tools/services.

Cloud providers such as AWS, Microsoft Azure, and Google Cloud Platform are all providing their customers with a plethora of analytics tools—which is both exciting and challenging as customers struggle to determine how and which tools to use. If companies want to achieve their desired results, they need to identify which services properly align with their business goals and strategies.

Today, big data analytics ranges anywhere from prescriptive analytics to diagnostic analytics, and each type of analytics is popular within specific industries and markets. What works best for the manufacturing industry may not be the best for the automotive industry, and so on. When exploring options, organizations should align themselves with an analytical service that is purpose-built for what they're looking to achieve with their data.

3. Be an early adopter of emerging tech.

According to the analysis of AWS spending trends, in the past year, cloud platforms that enable machine learning to integrate with applications had the most growth in the financial services industry. As companies started to recognize the value in machine learning, cloud providers reacted by allowing these solutions to be developed and hosted within the cloud, even taking on much of the training workloads involved with preparing a machine learning algorithm for practical use.

For example, financial services companies are looking to leverage machine learning technology to assist with predicting market trends and assessing risk management. By using tools such as AWS SageMaker, companies can deploy, train, and build machine learning models directly within their cloud environment, making it easier to run predictive analytical processes faster, less expensively, and in a more accessible fashion.

The financial services industry is staying on top of the latest tech trends, following the best practices of other enterprise technology innovators, and paving the way for others to follow. By doing so, they're merging their capabilities with the tech industry and expanding beyond traditional service offerings.

Today, some of the world's most traditional industries are taking advantage of the latest and greatest tech on the market. By establishing the right culture, identifying the right tools, and adopting emerging technologies early, the financial services market has become well-positioned to outpace and inspire innovation across all sectors.

TRENDING NOW

The Rise of Data Orchestration

Q&A With Alluxio's Dipti Borkar

Dipti Borkar,
VP, product management and marketing

'[E]nterprises are coming to the conclusion that data is going to be siloed.'

ALLUXIO WAS FOUNDED in 2015, with its 1.0 release following in 2016. It originated as the Tachyon project at the UC Berkley's AMP Lab by then Ph.D. student Haoyuan Li, who is now Alluxio's CTO.

As the 2.0 release was rolled out in July, Dipti Borkar, VP, product management and marketing at Alluxio, reflected on the data engineering problems that have emerged as a result of the increasingly decoupled architecture for modern workloads. Just as compute and containers need Kubernetes for container orchestration, Alluxio contends, data also needs orchestration—a tier that brings data locality, accessibility, and elasticity to computing processes across data silos, zones, regions, and clouds.

What is going on in the data management and analytics space in general now?

Historically, databases, data management systems, and data warehouses were all kind of tightly aggregated and vertically integrated to work in one location, one server, or one virtual machine. What we are seeing now is a lot more of the separation of the processing itself using different techniques, and then the storage may live somewhere else in a completely different location.

Why?

In the past, everything was on-premise, but now with cloud getting a lot more adoption for data applications, users have started to move data to the cloud and are getting more comfortable with that notion. In addition, the compute framework is not in the same location as the storage system, which might be the Hadoop HDFS, or something else on-premise. This idea of a single data lake, which was supposed to be the HDFS, is becoming harder and harder to achieve. Every business unit has its own data, and there are external data sources. And so I think that a lot of enterprises are coming to the conclusion that data is going to be siloed. Those are a few things that we are seeing from a data management and analytics perspective.

How does Alluxio address this?

Alluxio is actually a technology built to embrace data silos no matter where they live, as opposed to data being in a single data lake, where everything works where it is located. There are a lot more disaggregated stack scenarios, and there is a lot more adoption of cloud.

What does it enable?

Because of this separation and disaggregation of compute and storage, increasingly, you will need a layer that moves data closer to the compute—the data that is needed, the data that is actively being used—rather than all the data, which is going to be impossible to move around. That is where data orchestration comes in. It helps bring data closer to compute and makes it more accessible to the compute. The idea is to cache the most active data so you get data locality as well as make the same data accessible to many different APIs.

The file might be the same, it might be a Parquet file or an ORC [optimized row column] file, but different frameworks may want to access it in different ways. Maybe they want to use the HDFS API, or an S3 API, or a file system API on the same data. And that is the other aspect of data orchestration—making the data more accessible to the computer frameworks on the top.

What else is changing?

The other thing that we are seeing is the rise of the object store. AWS S3 has become incredibly popular and is driving more object storage usage. However, these stores were not built for analytics and interactive applications, so again, you need a layer on top to accelerate them and bring that data to wherever that compute is. This enables you to get that data locality and speed up metadata operations as well as make the access strongly consistent—and that is something Alluxio helps with, as well.

What does this enable?

Eventually, all of this is in the context of making the data access self-service. Typically, we see that there are platform teams that are responsible for serving this data, as well as the frameworks, and they create a service within the enterprise itself for the different business units. ▶

Technologies that are storage-agnostic and cloud-agnostic make self-service data access a lot easier and more efficient because they are not managing multiple copies of different datasets. The data is being synced back with wherever it lives and is being pulled out as needed, or on an on-demand basis. This will be the future. The most valuable data could be the most recent for one company or for a retailer, it could be data about specific stores. Wherever it is, the query will pull that data. It is a compute-driven approach, in which, based on the compute, you bring the data closer to the frameworks. That is the overall way in which Alluxio fits in.

This sounds similar in some ways to containerization.

There are two ways to think of it. First, as a piece of software, it works with containers. We have a Docker container and we work within Kubernetes as a daemon set. You can scale the cluster up and down and most importantly co-locate the Alluxio workloads, the Alluxio containers, with the compute that has the data within Alluxio.

Just as Kubernetes is to compute and containers, data orchestration is to data and to active working sets of data.

And the second way?

The second aspect is a little more of an analogy because just as Kubernetes and container orchestration is for compute where it is basically making compute more elastic and enabling the deployment of containers anywhere they are needed, data itself needs orchestration in this disaggregated data world. When the data is needed, the working set for that compute, that framework, must be moved closer and that requires orchestration. That is what we are calling data orchestration. Just as Kubernetes is to compute and containers, data orchestration is to data and to active working sets of data.

Is data orchestration spreading?

Alluxio itself is an open source project, and because it is open source we have a very large community. We have 1,000-plus contributors to the project, but more importantly, we are seeing worldwide usage. There are large communities in Asia and the U.S.

And the use cases?

We live in the big data ecosystem, and that is where Alluxio fits. We are seeing different use cases that are popular today. The first is accelerating cloud analytics and the second is hybrid analytics—hybrid cloud analytics using on-premise data. We are seeing most users start with a simple use case, and then it grows out from that point. We are also seeing different projects starting to emerge to solve this data orchestration type of problem on the NAS [network-attached storage] side of things. There are also other projects coming up in other spaces to solve similar problems. It is a new problem that is emerging, so whether it is our community itself or other projects and products, we are seeing a lot more movement in the market.

Interview conducted, edited, and condensed by Joyce Wells.

Accelerating Data Science With RAPIDS

It is well known that data scientists spend about 90% of their time performing data logistics-related tasks. After that part is done, to some degree, the data scientist finally gets to ask questions pertaining to the problem that needs to be solved. The thing is, 90% is a huge percentage—especially when spread across a team of data scientists. Anything that a data scientist can do to reduce it is a good use of time and a benefit to the organization as a whole.

With the advent of big data, frameworks such as Hadoop MapReduce and Apache Spark have been heavily relied upon to implement data manipulation and ETL pipelines. They have been well-regarded choices for data processing at scale due to their distributed nature. However, there are pros and cons to most tool choices. While these tools are distributed, the downside is that they are arguably some of the most complex tool stacks to manage. For this reason, the market has had adoption issues, especially with finding adequately experienced people to build and manage these systems. Couple that with the complexities and learning curves of writing the software, and it becomes a rather precariously steep hill to climb.

The utopian vision of software engineers and data scientists is to be able to obtain performance benefits and scale-out capabilities with the least impact possible—perhaps by changing as little as a single line of code. Of course, having to change a couple of lines is OK, but minimal change is really the key. While this may seem to be a pipe dream to some, it is a reality for many. For Python users working in the realm of software libraries such as XGBoost, Scikit-learn, or Pandas, there is the unique opportunity to not worry about setting up, configuring, or managing a cluster for Hadoop or Spark. They can make one minor library import change within their Python code and they can benefit from GPU (graphics processing unit) acceleration with RAPIDS.

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

The RAPIDS Framework

RAPIDS is a data science framework offering support for executing an end-to-end data science pipeline entirely on the GPU. Instead of creating new APIs to work with the GPU, RAPIDS has taken the approach of adopting and leveraging other popular APIs and libraries, and doing the hard work under the covers for the developer. RAPIDS hides the "how to" interaction with the GPU. The data scientists continue using the same APIs that they were already using and can just swap out the implementation of the library from the original CPU-based version to the new GPU-accelerated version.

A key underpinning of RAPIDS is its reliance on DASK to handle distributing Python jobs. Arguably, DASK is simpler to implement and manage than other distributed schedulers/managers. For nearly 2 years, it has had support for running within Kubernetes and has even broader integrations within the high-performance compute arena with SLURM (Simple Linux Utility for Resource Management) and LFS (Linux from Scratch). This is important to consider because, with this broader reach, we are more likely to see long-term success.

Benefits

Ultimately, the benefit of RAPIDS is to streamline the entire data science pipeline by allowing the developer to do more work, in less time, within the constrained resources provided, and with the tools they are accustomed to using. This GPU-accelerated approach takes the scaling concepts further than where the Hadoop ecosystem has gone by enabling scale-out over hardware which supports delivering substantially greater performance improvement. This approach provides for better compute density within the data center.

This leads to better utilization of hardware and puts a lot of CPU back into the processing pool to then be reallocated. This is a really good thing for those writing the checks for infrastructure utilization, whether on-premise or in the cloud. It provides a cost benefit to businesses that have previously only experienced CPU scale-out, allowing the business to continue scaling out in a much more affordable way. It is very common to see workloads that required hundreds of CPU servers be tackled with just a couple of handfuls of GPUs in fractions of the time they previously required.

With RAPIDS, it is very reasonable to see performance improvements ranging from one to three orders of magnitudes. When you start adding up the time spent on what are effectively mundane activities, which can be as much as 90%, anything that can cut out wasted time is worth considering.

Data Science Education Gets Visual

In 2010, there were a total of 131 confirmed, full-time business analytics university degree programs, including 47 undergraduate-level programs. When I wrote an article on this topic in 2017, that number had more than tripled at 564 programs, including more than 50 undergraduate programs, nearly 100 certificate programs, and close to 400 master's programs. As of April 2019, the numbers were even larger. The numbers had grown to 59 verified undergraduate programs, 101 verified certificate programs, a whopping 420 master's programs, and 23 doctoral specializations for a total of 603 verified programs globally.

As business analytics education, including specific instruction in data visualization, becomes more solidified in higher education, the question is not: "Are we teaching business analytics?" but instead becomes: "What are we teaching in business analytics?" To make education most valuable, it should align with what the market is looking for in potential job candidates.

The Rise of Data Visualization as a Desirable Skill

In an IEEE paper published earlier this spring, my fellow researchers and I studied the trends of industry job postings for business analysts, focusing primarily on the field of data visualization practitioners. Not only did we find that since 2010 there has been a 1,500% increase in job postings listing data visualization as a desirable skill, but also that the demand is expected to continue growing in 2019 and beyond, making data visualization one of the hottest areas of job growth and academic opportunity.

Paired with data visualization, employers specifically sought candidates with experience in SQL (51%), Tableau (41%), Microsoft Excel (34%), Data Analysis (31%), and Python (30%). Beyond technical abilities, they want software prowess. Of note was their interest in Tableau (41%), Excel (34%), and SAS (22%). Other non-technical skills included such quantitative skills as communication skills (47%), research (37%), writing (32%), teamwork/collaboration (31%), problem solving (30%), and general project management skills. While it may seem obvious that most employers want employees to communicate effectively, these postings specifically mention this and highlight communication and qualitative abilities requisite of jobs in analytics and data communication. This reflects an increased focus on visual data storytelling as a new skill that is an aspect of data visualization.

A Focus on Best Practices and Other Skills

This statistical information on industry demand gives us a sharp view into what we should be teaching—and learning—in this growing cohort of analytics degree programs and a yardstick by which to evaluate them. An appropriate class for professional students should focus not only on the underlying principles and best practices of data visualization, but should also include training on other skills, such as software specialization.

While exact curricula and approaches are up to individual schools, educators interested in designing courses appropriate for professional students should provide education on data visualization best practices in terms of graphics as well as visual cognition so that students know what the different types of charts and graphs mean, how they are used to represent data, and how to appropriately communicate insights.

No Longer Limited to Research

Visual design principles should also be taught as a foundation for how to apply elements such as color, shape, and visual hierarchy in data visualization charts, dashboards, and stories. These concepts can and should also be incorporated into a more traditional visualization class to help the students apply advanced programming concepts within an industry setting.

No longer limited to a research area, data visualization is an integral skill of business analysis.

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

The Artificial Intelligence Plateau

AI—A Dangerous Tool or a Fool's Errand?

The concern has become fear and may soon develop into a panic. The worry that has permeated the annals of science fiction for 100 years and recently become a reality for some is due mostly to the alternative use of, amazingly, video game computer chips.

VMware CEO Patrick Gelsinger has called AI a "30-year overnight success" story. The concerns are wide-ranging as new applications for the technology emerge. What will people do for work? Does this mean that a "universal basic income" will become a necessity? How many occupations will be replaced by a ubiquitous and pervasive world dominated by AI?

The Fear That Machines Will Replace Us All

Throughout scientific history, a relatively constant theme has been that mathematicians deal in the abstract. In the 1600s, Sir Francis Bacon, the "father of empiricism," helped transform pure philosophy into technology when he developed what became known as "the scientific method."

It was with this turning point of Renaissance genius that scientific thinking started producing technological innovations and products. Machines were created, and those machines could far exceed the productivity of the singular human counterparts who operated them.

Machines have created valuable products, and, sometimes, these machines have replaced the very people who produced the same products because they did it less expensively and with higher quality. Throughout the next centuries, the first Industrial Revolution, the 19th and 20th century thinkers and think tanks, politicians and purveyors of science fiction, doomsayers and religious leaders have feared (or profited from the fear) that machines would quite simply "replace us all."

Now we are in the 21st century and the time of extreme high-speed computing. For example, NVIDIA chips, which are known as GPUs (graphics processing units) and were originally created for video games, can process thousands of times the instructions of chipsets designed less than a decade ago. These chips and others have the ability to handle datasets of multidimensions and nearly infinite size with incredible speed. The interpolations and extrapolations are done so fast on such vast amounts of data so as to appear to be done at "the speed of thought" or faster—and certainly more accurately. After all, the computers are not distracted, don't have ulterior motives, aren't affected by fatigue or stress, and certainly don't complain when they are tasked with, well, multitasking. So, if they are not only faster than humans but also cheaper, does this mean that they are better?

The Reality of AI

We should consider the various applications carefully before we resign the human race to a universal basic income with a lifetime sentence of banality. Some examples should be informative. The applications developed from AI's precursors, machine learning and deep learning, are indeed remarkable. The science of neural networks seeks to develop machine systemizations that mimic animal (and even human) thought through the creation of non-biological neural messaging. Automated vehicles, automated cleaning systems, and advanced sports metrics are three areas of technology that have seen incredible advancements in recent years. Yet none of these areas have tangibly exceeded the ability of the human brain to perceive and infer.

Automated Cars and Trucks

Automated cars and trucks can be seen wandering the streets of the six sister cities of Silicon Valley (www.dbta.com/Editorial/News-Flashes/The-Six-Sister-Cities-of-Silicon-Valley-125972.aspx) through the dark of night and even during the day, collecting data-building sets of information that the GPUs can crunch in nanoseconds to instantly calculate the best-possible option to take in every conceivable situation on the roads. Automated vehicles seldom make mistakes, but they also never infer, are never creative, and get stuck when the best answer is not the product of a calculation. ▶

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

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

Robotic Cleaning Systems That Learn

Robotic cleaning systems are fun to watch and tremendously useful. Sometimes, they know how to learn. After your automated cleaning system gets stuck a few times inside a particular chair, it can collect enough data to stop going through that chair's legs, and it may even remember that same lesson when you move the chair. However, the whole system needs to be retaught when you buy a new carpet, or your dog decides that the robot is now a competitor for a particular corner of the house. Automated cleaning systems do not infer, do not exercise creativity, and simply stop when they run out of power.

When creativity is combined with brilliance, genius emerges—and genius is the only power that changes the world.

Sports and Life

Sports is the perfect allegory to life. Modern baseball has become the refuge of those who could never qualify for a sports team. The computer calculations can tell the general manager who to hire as a manager and which players to put at the correct positions on the field when a particular batter steps up to the plate at a critical point in the game. That is, until the innovative batter decides to apply a different approach to the crucial at bat. So, when the infield has shifted to the spot where that particular batter hits the ball 95% of the time, the batter can use his mind, his creativity, and his inspiration to hit the ball to the opposite side of the field. Score one for the human and zero for Sabermetrics. When creativity is combined with brilliance, genius emerges—and genius is the only power that changes the world. Genius is most often manifested in art, including the art of sports.

No Substitute for the Human Mind

Returning to the original thesis of this article for a moment, we should remember that throughout history, mathematics has preceded science by 50–100 years. Mathematicians work on their abstractions, and then sometimes, decades later, a technologist uses that mathematics in some component of a product. Today, the reverse is true. The technology is far ahead of the algorithms necessary to give a computer the ability to infer or be creative. Until automated cars have the ability to infer, they won't be able to master dotted lines on the highway that were drawn incorrectly, automated cleaners won't be able to outmaneuver the curious beagle, and sports calculations certainly won't understand that all at bats are situational. There is no substitute for genius, and despite the awesome power of the GPU and the majesty of the new manifestations of AI, there is no substitute for the human mind.

The Conundrum of Data Governance

It might be the most frequently asked question of a data governance consultant: "Who should own data governance, the business or IT?" And man, that's a loaded question! When you dig deeper into the root of the question, most people really want to know one of two things—"Who should ultimately own data decision making for our company?" or, "Where will data governance be most successful?" Let's take a closer look at those two questions.

Authority for Decision Making

Who should own data decision making for our company? Businesses are wedded to hierarchical reporting structures that distinctly outline decision-making authority and borders of control. This means that when the question of where data governance should reside in a hierarchical organizational structure is asked, one is typically implying the question of where the authority for data decision making should reside. This question, of course, makes everyone squeamish because neither the business nor the IT team wants to give up the authority for data decision making or, perhaps more importantly, they don't want the other side to have the control. Ironically, neither side wants the responsibility or accountability to go along with the perceived power. And, absent the responsibility and accountability, you end up with ample blame and criticism.

The funny thing is, when implemented correctly, data governance actually brings no authority, responsibility, or accountability for data decision making, at least not directly. Instead, data governance establishes a decision-making framework where authority rests at the appropriate location and level of the organization based on context. The context may be based on a multitude of things such as system or application expertise (accounting system, CRM, customer portal), subject area knowledge (customer, product, patient, student), business unit experience (sales, marketing, operations), or cross-functional process involvement (invoicing, inventory, credit approval). Data governance defines roles and responsibilities to identify who has the authority and accountability for which data. When there is conflict due to shared responsibility in a data decision such as defining key business terms or outlining a business process, data governance facilitates and provides the objective means for the involved individuals to come to a decision. And when decisions are made for data standards and business rules, data governance develops and implements policies and procedures to enforce them.

Once we establish that owning data governance does not mean exclusive authority, control, and power over all enterprise data decisions, the number of interested owners drops significantly. Weird, right? So now, no one wants to own it, but everyone agrees we need it. It's the conundrum of data governance. Enter the second version of the question.

Business Versus IT

Where will data governance be most successful? The answer here is that, to be successful, it doesn't matter which side of the house data governance is on. What does matter in determining where data governance should live in order to be successful is the corporate culture. In other words, where does it most comfortably fit based on how the company behaves and operates today? Where will the implementation of the data governance program be the easiest and smoothest? If the corporate culture is compliance-driven, then data governance may best live on the business side in operations or legal, where a top-down approach is expected. Conversely, if the business operates in silos, a single business department will likely not exist, and data governance may have a very happy home on the IT side.

Consider the design of a car as an analogy, and the question of which side is best for the steering wheel and driver. In order to physically drive the car, it doesn't matter. Regardless of which side the steering wheel and driver are on, the car still starts and stops the same way. The engine still runs the same. The steering wheel still moves right and left. But, once we know which side of the road the car is supposed to travel on, the position of the driver's seat changes how easy it is to drive.

Functioning as a Bridge

Data governance has a bridge role, requiring involvement from both sides of the house—business and IT. Regardless of where data governance sits within the organization, both sides will have data decision-making authority, accountability, and responsibility. The question should not be: "Who should own data governance?" Rather, the question should be: "On which side should the data governance driver's seat be for the most successful journey?"

Anne Buff is the director of data governance for Envolve Healthcare (www.envolvehealth.com). She has been a specialist in the world of data for more than 20 years and is an industry leader in data governance.

Be Obsessed With Analytics

. . . And wisdom will be your cup of tea

HAVE YOU HEARD that data is the new oil? If not, then the chances are great that you will read this in an article somewhere—tomorrow or the day after. And, if data is the new oil, then the way to make money from it is by exploiting it—in as many ways possible. (Borrowing from businessman and author Stephen Covey, the habit of being obsessed with analytics is part of my continuing series on the seven "habits" that successful IoT projects have in common.)

How do you exploit data, you may ask? Well, just as in the oil industry, exploitation comes down to how good your refinery capabilities are. Just as oil has to be refined to get valuable products out of it, such as gasoline and jet fuel, data has to be refined into insights. And, just as in the old days of the oil business, the rush is on.

For this reason, many companies are snapping up data scientists. Those who can't get a data scientist to join them try to hire consultancy firms to help. Consultancy firms that specialize in data science are exploding in number—and this is causing some problems.

Not Enough Data Scientists

For a start, there are not enough data scientists to quench the current thirst and appetite, driving salaries through the roof. If they get poached away from you, your domain/company knowledge could be gone tomorrow. I heard someone remark that today's data scientist is what the programmer was in the 1980s. If that is true, the worst is still to come. (It is interesting to see that data science consultancy firms are establishing strongholds around university cities with a strong pedigree for analytics. Cluj in Romania is one such place, where tens of thousands of fresh graduates address the need for new hires.)

What do you do, however, if you don't have the means to start centers of excellence in exotic places to get those data scientists? There are solutions coming that can help fight fire with fire.

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

The Advance of Self-Service Analytics

One of the most exciting trends in the market is the advance of self-service analytics, bringing the cup of wisdom to the many. One of the examples I encountered recently was the acquisition of a self-service analytical tool, TrendMiner, by my own company. It has capabilities that enable self-service analysis of time series data. That means knowledge workers, such as process engineers, can visualize any time series data they have access to and then ask the tool—through predefined analytical models and techniques such as machine learning—for correlations and similarities. The whole idea is to find indicators and early warnings that can be used to prevent major anomalies.

The future and the way forward is to make the data science work easy. Similar to how we simplified programming by going from binary to third- and fourth-generation languages, we will see simplification of data science tools.

Just last week I was called by a colleague who had his first experience of using it for a day, as well as a knowledge worker in a pharmaceutical company, who said that within a few hours, they had found correlations between events they never thought were related.

The Future of Data Science

This strengthens my belief that the future and the way forward is to make the data science work easy. Similar to how we simplified programming by going from binary to third- and fourth-generation languages, we will see simplification of data science tools.

That is what democratization of data is all about. If data science is about asking questions, then it will not be about who can program the smartest answer, but who can ask the smartest question. An obsession with analytics pays off.

CALL FOR SPEAKERS NOW OPEN
CLOSES NOVEMBER 18

AI & MACHINE LEARNING SUMMIT

MAY 19-20, 2020

HYATT REGENCY BOSTON | BOSTON, MA

A featured event at DATA SUMMIT
UNLEASH THE POWER OF YOUR DATA

REGISTRATION OPENS SOON

Artificial intelligence (AI), along with related technologies, has come into its own. AI, including machine learning (ML) and cognitive computing, is revolutionizing and transforming business operations. From healthcare and financial services to manufacturing and education, every industry is profiting from implementing AI and ML initiatives. Organizations can better engage with customers and employees, transform business processes, and make better decisions with insights from deep learning and enhanced analytics. AI's complex analysis functionality leads to the ability to create new products and optimize existing ones. In a data-driven environment, maximizing new AI and ML technologies provides important business predictions. To equip you with the knowledge to succeed, we are bringing together the leading industry experts for a 2-day immersion into the leading AI and ML use cases, strategies, and technologies that every organization should know about. **Mark your calendar now!**

bit.ly/AIMachineSummit2020 **#AIMachineSummit**

Address the ELEPHANT IN THE ROOM

Bad address data costs you money, customers and insight.

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

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

See the Elephant in Your Business -

Name it and Tame it!

melissa

www.Melissa.com | 1-800-MELISSA

Free Trials, Free Data Quality Audit & Professional Services.