IDC | ANALYZE THE FUTURE

# RESHUFFLING THE DATA DECK:

# Benefits of a Modern Cloud Data Lake Platform

Sponsored by Qubole and Google Cloud

RESEARCH BY:

**Dan Vesset**
Group Vice President, Analytics and Information Management, IDC

September 2020 | IDC Doc. #US46799120

# Navigating this Spotlight

*Click on titles or page numbers below to navigate to each.*

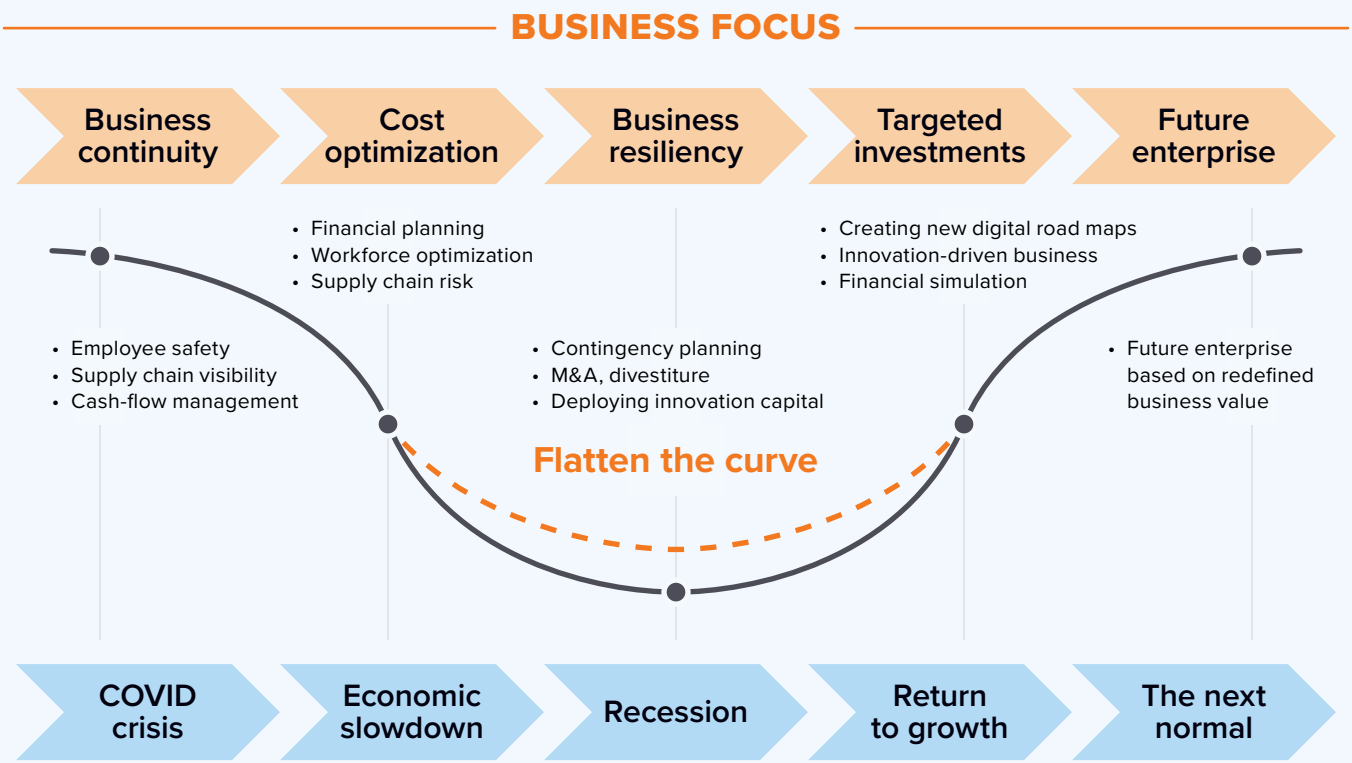
# Introduction

## From Crisis to Recovery

In less than three months of 2020, the term "COVID-19" redefined itself multiple times, from a novel virus to an outbreak, an epidemic, a pandemic, and then a world crisis. *Unprecedented*, *uncharted*, *unknown*, *unclear*, *unprepared*, and *understated* are just some of the adjectives that continue to be used about the crisis in news reports, situation analyses, and government announcements. These words compose a wake-up call to the leaders of organizations and administrations that a new set of digital capabilities is needed to underpin new policies, plans, and models for addressing the challenge and opportunities.

When learning from and advising organizations about the path forward, we at IDC use the Crisis to Recovery framework as shown in Figure 1.

FIGURE 1




Source: IDC, 2020

This framework is not meant as a prediction of a U-shaped curve but rather as a mental model for evaluating the stages of moving from the COVID-19-induced crisis to a "new normal" and the key business or organizational stages aligned to the curve. At each stage of the curve, executives' imperatives change and drive a need for evolving strategic, tactical, and operational analytics and decision making. These needs, in turn, highlight the requirements for a new set of data and analytics technology capabilities.

The technology alone will not "flatten the curve," but it will enable new and accelerated efforts to respond to the crisis, whether in pharmaceutical and public health research or in commercial efforts to analyze supply chains and customer experiences. In short, COVID-19 has exposed at the executive level a set of enterprise intelligence deficiencies and opportunities that a new generation of chief data officers must address.

Whether your organization has extensive data management experience and has "felt the pain" of previous missteps or is a relative newcomer, we hope the information contained in this IDC Spotlight helps guide your decision-making process about new data and analytics technology solutions and helps you successfully navigate the road from crisis to recovery.

# What If … ?

What if you need to provide your team of data scientists with ad hoc, self-service access to data resulting from 2 million transactions per day? What if you need to process 10TB of data per day? What if you could do this by running your cloud data lake without a dedicated maintenance engineer? What if you could run 10 times more jobs than you currently run and do so for less money? What if you could improve system reliability, resulting in an 80% reduction in end-user complaints? What if you could provide a single platform for all of your data scientists?

These are not only hypothetical questions but also results that a growing number of organizations are achieving as they transform their data platforms to meet the needs of modern analytics, machine learning (ML), and real-time streaming data processing requirements. Among these companies are firms such as a major digital commerce platform provider in Asia/Pacific that migrated its data lake platform to a new cloud data lake platform from Qubole running on Google Cloud.

What if your business depends on building an analytic data platform for your B2B client organizations? What if your data scientists' success defines your whole company's success? What if these data scientists require 99.999% reliability for the whole analytic pipeline? What if you need to operationalize the production of 1 billion predictions per day? What if you need to process 20TB or 200TB of data per day? This is what a major North American digital experience service provider has achieved with technology support from Qubole running on Google Cloud.

These two providers and other similar companies are well on their way to a future of enterprise intelligence that IDC defines as an organization's capacity to learn combined with the ability to synthesize the information it needs in order to learn and apply the resulting insights at scale.

## AT A GLANCE

**Enterprise intelligence is defining tomorrow's winners and has emerged as a top priority for business leaders.**

### Key Stat

In a recent IDC study, 87% of CXOs cited enterprise intelligence as a key priority for the next five years.

### What's Important

Technology leaders must address the growing number of and more mission-critical analytics and decision-making requirements. This requires modernization of data platforms to enable more automated solutions to address all data-at-rest and data-in-motion workloads and usage patterns.

The new generation of data analysis, decision support, and decision automation capabilities of these companies has become especially relevant and has taken on added urgency during the year of COVID-19. In an era of unprecedented uncertainty, executives are demanding new data and analytics capabilities to support ongoing changes to decision making to address the crisis, to become more resilient, and, eventually, to prepare for the recovery.

In an IDC study of 157 U.S. chief experience officers (CXOs) conducted in the first quarter of 2020, 87% of respondents cited enterprise intelligence as a key priority for the next five years. What these CXOs are asking for themselves and their enterprises is not simply more data or more algorithms—they are specifically citing the need to move from using data to analyze performance to using insights to affect performance in these turbulent times.

A growing number of organizations are appointing data and analytics leaders, such as chief data officers (CDOs) and/or analytics officers, to not only optimize investment in data and analytics technology but also drive efforts to create or improve a data-driven culture and to raise data literacy to ensure that outputs of analytics are available to everyone.

Yet many companies and their technology leaders are challenged by traditional thinking about what constitutes enterprise intelligence and the technology architecture and capabilities needed to achieve it. As Herbert Simon, a Nobel laureate in economics, wrote in 1971, "In an information-rich world, the wealth of information creates a poverty of attention." This statement is even more relevant today.

To filter out noise and deliver actionable information to internal users and external clients, organizations can't rely only on the legacy data warehousing paradigm. While data warehousing remains an important technology component of a comprehensive analytic data management platform, its capabilities must be extended with the latest cloud-native capabilities for supporting analysts and data scientists tasked with cross-functional analysis of data from multiple internal and external sources, data arriving in batches and streams, and data residing in the cloud and on premises.

# Use Cases and Usage Patterns

Gone are the days of building a data warehouse or a data lake because "we need to invest in Big Data." Today's market environment dictates a use case — and an ROI-driven approach to business analytics, including investment and deployment in an analytic data management platform.

However, the concept of a *use case* has itself become problematic when prioritizing technology investments. As a market research and advisory firm, IDC often receives inquiries about identifying use cases for analytics or machine learning. That is a difficult question to answer, and we feel it's often the wrong question to ask. A use case for business analytics exists anytime any decision by anyone is being made within your organization. Thus, use cases occur:

- **In any business function,** such as marketing, sales, finance, supply chain, and operations
- **At any level,** because decisions are made not only by executives and managers but also by knowledge workers and frontline workers and, in some cases, by machines
- **For different types of decisions,** such as strategic, operational, and tactical or portfolio, planning, and execution

What are data architects and data engineering teams to do? One of the effective approaches we have seen employed among our clients includes assessing internal user requirements based on usage patterns.

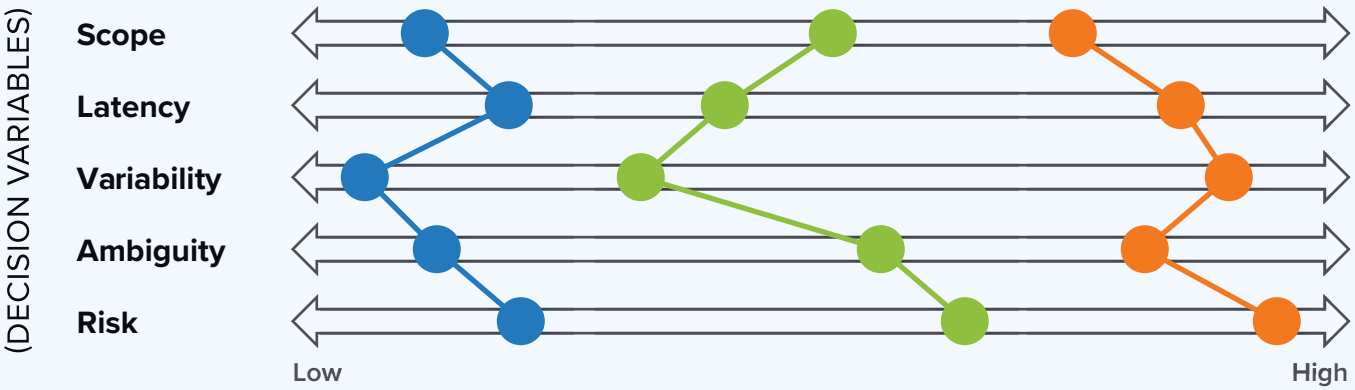
### As shown in Figure 2 (next page), each usage pattern is characterized by five decision variables:

- **Scope** defines the breadth of the impact of a given decision. Does it impact a single customer or many or all customers? Does it impact a single activity or one whole process or multiple processes?
- **Latency** defines the time window or time interval within which a decision needs to be made or an issue needs to be resolved. Some decisions need to be made in subseconds (e.g., real-time recommendations), while others may have weeks or months of lead time (e.g., acquiring another company or entering a new market).
- **Variability** defines to what extent the issue is predefined versus ad hoc. Is this a regularly occurring decision or a rarely occurring decision?

- **Ambiguity** defines how open-ended the issue at hand is. How open to interpretation is the data needed to make the decision?
- **Risk** defines the monetary value at risk of the decision. Decisions with narrower scope tend to have a lower level of risk; however, there is not a perfect correlation between risk and scope. For example, a planning process could be affecting a narrow part of the enterprise but have high risk associated with compliance. Similarly, a narrowly defined tactical decision could have high reputational risk.

FIGURE 2

## Usage Patterns




Source: IDC, 2020

Responses from business users to questions about these characteristics will allow the data engineering team to identify decision-making patterns, which need not be aligned with individual personas or departments or groups. Instead, similar patterns could emerge in a scenario for logistics optimization and website visitor experience optimization. Both functional examples may require ingestion of streaming data, location awareness, time series data management, and an ability to parse log files. Even if two requests from different parts of the enterprise may seem different from the business perspective, they could exhibit similar usage patterns from the perspective of the underlying data platform.

## The three common categories of usage patterns are:

- **Data exploration and investigation**
  This decision-making pattern is about helping users understand and explain what happened in the enterprise over a given time and why it happened. The analysis is performed by business analysts or data scientists. This usage pattern typically involves capabilities for ad hoc analysis, integration of data from multiple sources, and storytelling.

- **Decision automation**
  Decision automation represents tactical decision making in the flow of operations. The automation, whether conditional (rules-based) or algorithmic (machine learning-based), can involve straight-through processing without any human involvement during the whole end-to-end process, or it can include the augmentation of people with lower-level task or activity automation.

- **Enterprise performance management**
  Enterprise performance management supports the ongoing measurement of the activities of the enterprise and of the external factors affecting it. It provides managers and executives with better situational awareness of the current condition of the enterprise and the ability to plan and forecast.

Note that these usage patterns are intentionally not aligned with any one data engineering or data science capability.

## A technology platform should unify capabilities for:

- **Real-time streaming data processing**
- **Advanced analytics and artificial intelligence (AI)/machine learning (ML)**
- **Data visualization and ad hoc analysis**

The platform must enable descriptive, diagnostic, predictive, and prescriptive analytics. Not every platform provides these capabilities in a unified fashion, nor is technology the only issue facing organizations on their path to improving their enterprise intelligence.

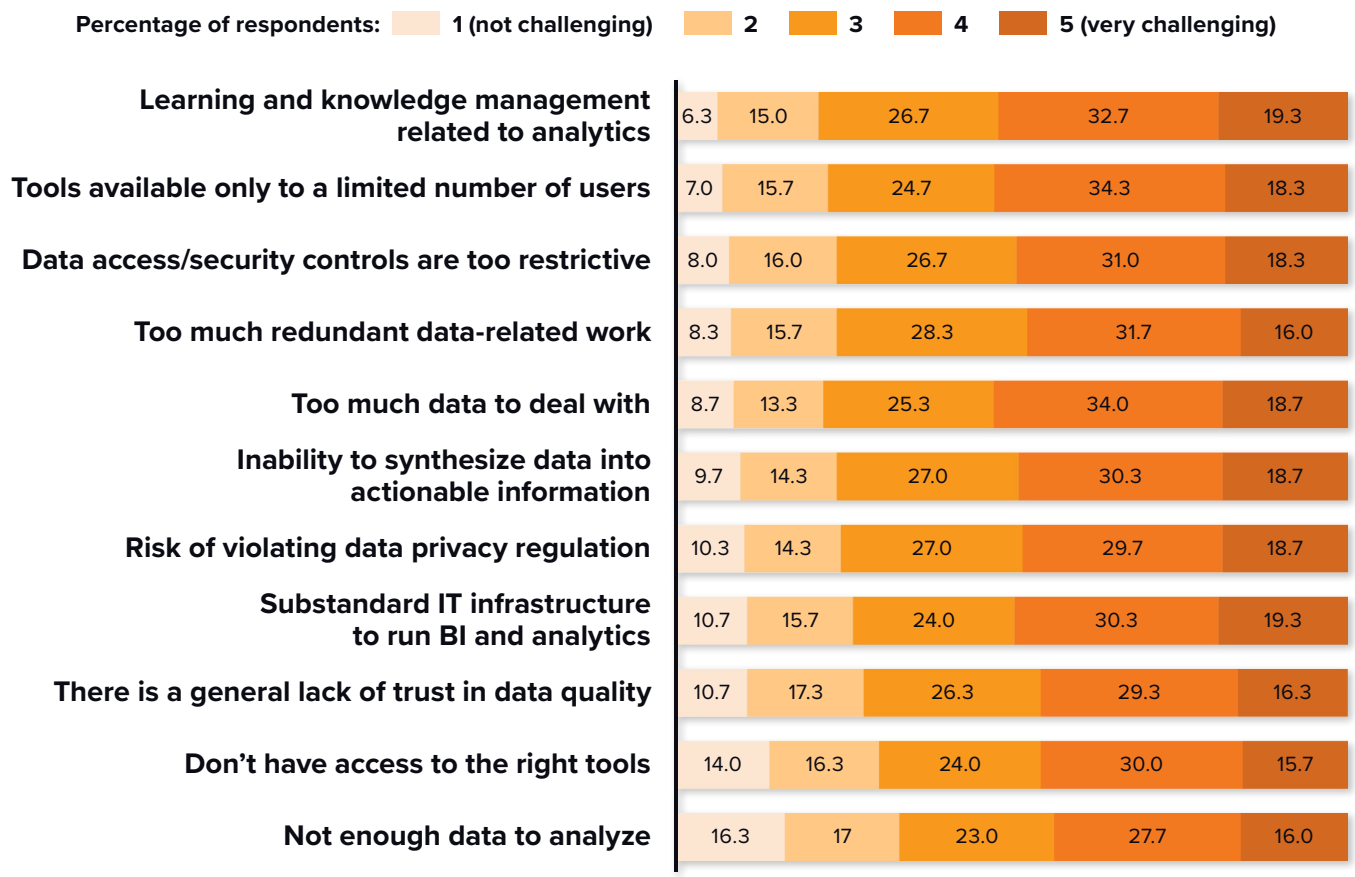
# Challenges and Roadblocks

There is no shortage of challenges facing today's data engineering teams. Today's demands on data engineering teams and chief data officers span a broad range of usage patterns with their varied data access and processing requirements. These demands vary and can be conflicting. As shown in Figure 3, which presents results from a 2020 IDC study on analytics, some groups complain about too much data while others complain about not enough data. Some groups are challenged by privacy and security policy issues while others, including 20% of respondents, complain about substandard IT infrastructure to run analytics jobs as being very challenging.

FIGURE 3

## Technical Challenges with Respect to Analytics

**Q. Please rate each of the following technical challenges with respect to analytics in your organization.**




Source: IDC's *Business Intelligence End-User Survey,* February 2020 | n = 310 (United States only)

**While data warehouses continue to play an important role in any comprehensive analytics and data architecture, they can't address all user requirements. Over the past several years, more organizations have incorporated data lakes — as well as multiple data movement, integration, and data intelligence technologies — into their architecture.**

# Architecture

Many organizations are saddled with a legacy thinking that for years (even decades) relied on a data analytics architecture that included a tool for data extraction, transformation, and loading (ETL) into a relational enterprise data warehouse, which in turn is accessed by business intelligence (BI) and analytics tools. While data warehouses continue to play an important role in any comprehensive analytics and data architecture, they can't address all user requirements.

Over the past several years, more organizations have incorporated data lakes — as well as multiple data movement, integration, and data intelligence technologies (e.g., data catalog, master data, metadata, data quality) — into their architecture. But not all data lakes are created equal. Early in the use of data lakes, many were "homegrown" or internally developed and required substantial ongoing maintenance from internal data engineering teams. For example, this was the case at a digital experience services company with a team of seven data engineers that was struggling to support 20+ data scientists who were part of the company's professional services group, building analytics solutions as a service for clients.

The reality for many data engineering teams is that they not only must maintain data warehouses, data marts, and data lakes but also must develop and maintain connection among all these various data management solutions, both in the cloud and on premises. The aforementioned digital experience service provider has data flows where results of data analysis in a data lake are loaded into the data warehouse; in other cases, processed data in a data warehouse becomes a source for data science workloads on a data lake.

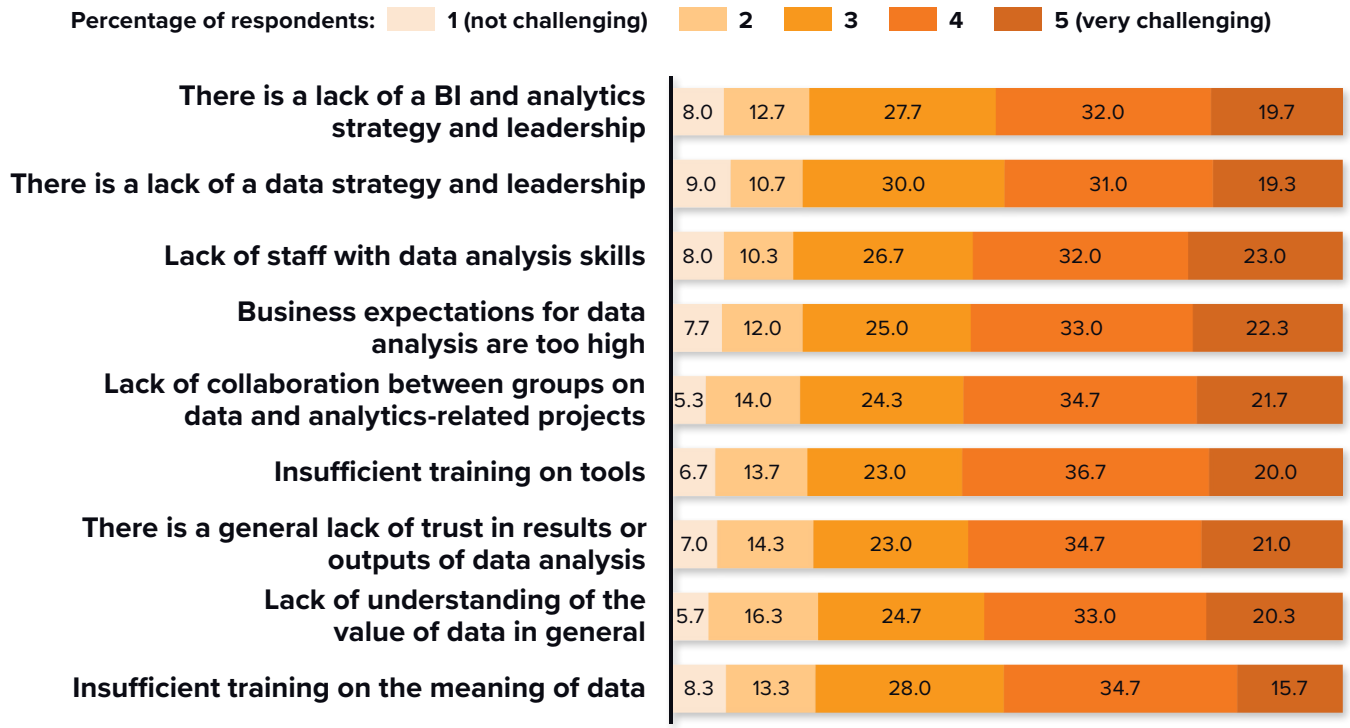
# Culture, Skills, and Staffing

The biggest roadblocks are never only about architecture and technology; some of the biggest challenges are related to lack of an enterprise-wide data strategy, inconsistent data culture, and lack of staff with relevant skills. Over the past several years, the latter issue has focused almost exclusively on the lack of data science skills. However, the lack of top-level data engineering skills is as or even more acute (based on average salary trends and the number of data science and business analytics education initiatives over the past decade). Figure 4 provides a sample of common business challenges and the extent to which they affect organizations that participated in the 2020 IDC study on analytics.

FIGURE 4

## Business Challenges with Respect to Analytics

**Q. Please rate each of the following business challenges with respect to analytics in your organization.**




Source: IDC's *Business Intelligence End-User Survey,* February 2020 | n = 310 (United States only)

# Financial Intelligence

We have all heard the adage that "you can't manage what you can't measure." In today's market, we should also add that "you can't automate what you can't measure." Data platform management through automation is top of mind for many of today's chief data officers and other data engineering leaders. However, in our interviews and studies of organizations across the globe, we have found that many of these technology managers still lack the financial insight and governance support needed to make effective decisions. Ironically, people who support analytics and decision making for others in their organization don't always have that functionality for themselves.

The data management team, like other users within the organization, needs functionality for usage monitoring to be able to analyze user-interaction and system-performance data for uncovering past trends, report on this analysis, predict future outcomes, and optimize system administration and management decisions taken by either humans or the machine. Ultimately, this type of financial intelligence about the data and analytics platform is required for effective automation of many of today's manual administrative, tuning, and maintenance tasks.

To address these challenges, a growing number of organizations are turning to a new generation of cloud-based data management technology.

**Financial intelligence about the data and analytics platform is required for effective automation of many of today's manual administrative, tuning, and maintenance tasks.**

# Technology Needs and Requirements

Modern analytic data management needs and requirements are driven by elevated awareness of the value of analytics in today's uncertain market; growing volume, variety, and velocity of data; and availability of new data analysis and AI/ML tools and techniques, many of which are based on open source technology.

## Platform Agility

In a 2015 article by McKinsey, the consulting firm defined *agility* as a combination of speed and stability. In this IDC Spotlight, we expand on that definition to combine several of the common system characteristics under platform agility: performance, scalability, manageability, and visibility.

To address all of the major usage patterns discussed previously, today's analytic data platforms must be able to scale to handle dozens of terabytes of data daily. For variable use cases, the platform must be able to scale up and down as needed to provide optimal price/performance outcomes. Also, given today's constraints on IT budgets and staffing, the platform should incorporate automation capabilities that release data engineers and system administrators from certain mundane daily tasks, allowing the technical staff to focus on higher value-added projects.

## Openness

When the senior vice president (SVP) of the marketing cloud for a major digital experience service provider started looking for a new platform to replace the company's homegrown legacy technology, openness was a key requirement. As a former chief technology officer (CTO), this executive was initially concerned about the ability to deploy the analytic data management platform for the company's data scientists on multiple clouds. This was partly a strategic architectural requirement, partly a risk-management decision, and partly a decision driven by customers in some industries that prefer or require one cloud over another.

However, an open platform means more than having the ability to deploy on multiple clouds. To some, openness means having open source components. But even that designation is limiting. Openness can also mean the option to use multiple storage formats, multiple data science and analytics tools and languages, and multiple AI frameworks, as well as the ability to develop multiple upstream and downstream data pipelines via connectors and application programming interfaces (APIs) and the ability to extend the solution by internal IT staff who are able to rely on industry-standard development tools and techniques.

Ultimately, openness means freedom of choice, which is what this SVP and his team achieved with the data lake platform from Qubole running on Google Cloud. Today, data scientists of this digital experience service provider use Jupyter Notebook, PySpark, Airflow, and MLflow, among other tools.

# Cloud Deployment and Migration

Today's requirements are dominated by cloud deployment. According to IDC research, during the full year of 2019, global spending on cloud-based analytic data management solutions grew 47% while spending on similar noncloud technology declined by 1%. While some organizations are still constrained from moving to the cloud by regulatory requirements or internal security policies, the vast majority of organizations are allocating their net-new spending to cloud-based analytic data management solutions.

However, 50+ years of legacy doesn't change overnight. We expect that most organizations will continue to have a hybrid data environment. In this context, we use *hybrid* to mean a mix of in-the-cloud and on-premises deployments. In the meantime, the speed of migrating existing data platforms to the cloud has emerged as an important variable in the new platform selection process. In its cloud migration, the digital commerce company interviewed by IDC needed only two months to migrate its legacy data platform to the new, production-ready platform on Qubole and Google Cloud.

# Augmented Administration

A modern analytic data management platform should not only support requirements of data scientists for various advanced analytics techniques but also incorporate some of those techniques within its own operations. In other words, just like we talk about autonomous cars as a future endpoint on a spectrum from fully human-operated cars to cars that augment some driving functions (e.g., self-parking, cruise control, proximity monitoring) to fully autonomous cars, so too should we view analytic data management platforms through the same lens.

### A modern analytic data management platform should include capabilities such as:

- **Workload (and service-level-agreement)-aware auto-scaling for downscaling and upscaling**
- **Intelligent spot management**
- **Dynamic workload packing**
- **Automated cluster life-cycle management**
- **Multitenant and single-tenant cluster management for applications**

Capabilities such as these and others ensure that some tasks and activities previously performed by database or systems administrators can be automated. However, at this point, the whole process of managing the data platform is not automated — nor should it be. Today's platforms need to strike a balance between available and proven automation techniques and human experience and expertise that can add contextual intelligence still not grasped by today's "intelligent machines."

# Benefits

Although some market pundits espouse the notion of many failed data analytics projects, we at IDC don't see evidence of this. Yes, there are roadblocks and challenges to overcome, and some projects may go through growing pains. However, IDC research over the past 20+ years has consistently shown high value derived from analytics projects. In a recent IDC study, 75% of respondents indicated that the value derived from analytics projects exceeded their expectations.

That does not mean that there is no room for improvement. As mentioned in the "Challenges and Roadblocks" section of this IDC Spotlight, business and technology issues are part of any data and analytics initiative. In our interviews with Qubole and Google Cloud Platform clients, we heard similar stories and evidence of deriving benefits and overcoming challenges.

## Scalability, Performance, and Reliability

As the associate vice president (AVP) of data engineering at the digital commerce company said, *"Qubole running on Google Cloud gives a much better ability to automatically start and terminate jobs."* Prior to the migration, the company ran auto-scalability tests with different constraints. The capabilities of the solution proved superior to those of the prior on-premises technology (using Hadoop, Kafka, and Spark) as well as those of a previous cloud solution. For example, the previous cloud solution wasn't able to auto-terminate jobs (only auto-start them). As this data engineering AVP said, *"Instead, we had to manually spin up a cluster, then delete it."* The company had to always overprovision its environment, which resulted in higher than necessary costs. With the previous cloud solution, the company was able to run fewer than 100 jobs on a single cluster; today, with Qubole on Google Cloud Platform, the company is running 1,000 jobs.

The company also experienced reliability improvements. With the previous systems, there were daily internal client complaints from data scientists and analysts; with the new Qubole on Google Cloud, complaints have decreased by 80%.

In another example, the digital experience service provider that builds customer data platforms for its clients and then provides related analytics services on this platform was able to improve support for internal data scientists and, by extension, for external clients. For some clients, this service provider now processes dozens of billions of rows of data per day, and because the platform is external-customer-facing, there is an expectation for reliability characteristics akin to those of an

operational system (rather than an internally facing data warehouse — which, while important, is typically not mission-critical from the perspective of operational system requirements).

The senior vice president of this digital experience service provider also counted among the benefits of its new platform the ability to extend from its current use of batch processing to streaming data processing. In another IDC study, conducted in October 2019, 43% of organizations stated that their data environment incorporates streaming data. As the demand for speed increases, more organizations are future-proofing their data and analytics architecture by including technology that can process both data at rest and data in motion.

## Control and Efficiency

Although the digital commerce company was not able to do an "apples to apples" cost comparison between its old platform and the new cloud data lake platform from Qubole on Google Cloud, it estimates a projected 40% monthly cost savings. The savings are the result of not only lower technology costs but also, just as importantly, lower labor costs. In the past, the company had two to three dedicated administrators for the analytic data platform. Today, the company has no dedicated administrators for maintaining the Qubole solution on Google Cloud. Anyone on the data engineering team can step in, review logs, and address any identified issue. In this scenario, if we assume that the company went from using three dedicated data administrators to 0.25 administrators, then the number of users per administrator increased 12 times.

However, the productivity improvements accrued not only to the data engineering team. Now, all data scientists are able to use one consistent data and analytics platform. Data scientists, who work on such initiatives as search algorithms, recommendations, and trend analysis, can use Jupyter Notebook to connect to the data in Qubole and don't have to rely on any custom machines. Some of them have also started to use Google's Looker business intelligence software, which the company plans to roll out to a broader user base. Most of the jobs, which process around 10TB of data per day, are handled by data scientists, while data engineers provide the foundation and the raw data. This provides the company with the most efficient division of labor based on a single consistent platform.

Another element of control has been security management. As the associate vice president explained, initially, the Qubole cluster used external IPs. This practice was inconsistent with the company's security policy, and with Qubole's help, the company is now using only internal IPs.

Similarly, the digital experience service provider was able to quadruple the number of predictions it was creating on its new Qubole platform on Google Cloud without increasing its data engineering or data science staff. Its ability to operationalize more models developed by data scientists enhanced this company's ability to service and retain customers as well as use this capability to create competitive differentiation in pursuing new customers.

# Considering Qubole and Google Cloud

Qubole, founded in 2011, provides customers with a modern data lake platform. From its origin, Qubole has embraced open source componentry (its founders cowrote Apache Hive) and today it incorporates Spark, Presto, Hive, and Airflow. As a data lake software provider, Qubole itself doesn't provide the underlying storage and compute infrastructure; instead, that infrastructure is provided by the cloud platform provider—in this case, Google Cloud. All Qubole customers access its technology via application programming interfaces (APIs) or native user interfaces.

In addition to its core data management functionality supporting multiple analytic, business intelligent (BI), and machine learning (ML) workloads, Qubole provides a module called Cost Explorer, which is an internal system and usage performance analysis and reporting tool that data teams can use to track spend with fine granularity, monitor showback, and budget and allocate costs in support of their ongoing decision making. Combined, these monitoring and analysis features underpin Qubole's capability to automate myriad data engineering and administration tasks and activities. This automation provides organizations with valuable leverage in deploying their overextended data engineering teams, thus increasing the productivity of these IT teams as well as data science and analytics teams.

## Challenges

As with any technology provider, it's important to assess not only the technical functionality of specific tools and support services provided by Qubole but also its vision and partnerships. It's also important to evaluate the performance, availability, scalability, and security characteristics of the preferred cloud platform and the commitment of both Qubole and the cloud platform vendor, such as Google, to the partnership. As always, IDC recommends proof-of-concept projects and in-depth reference calls for any enterprise considering new technology from solution providers in the IT market.

# Recommendations

Whether your organization has extensive data management experience and has "felt the pain" of previous missteps or is a relative newcomer, we hope the information contained in this IDC Spotlight helps guide your decision-making process about new data and analytics technology solutions. We find that the more experienced organizations and their data professionals tend to focus on cost and efficiency, while less experienced organizations tend to focus more on time to value. However, this segmentation is not binary.

## We recommend that everyone consider the following:

- **Implement "matched triple" leadership and operational project teams** to harness business, data engineering/IT, and analytics/AI expertise. In this case, the "triple" refers to emerging growth in the number of chief data officers (or equivalent roles) and the need to expand the traditional IT-business relationship to incorporate this third partner in data, analytics, and decision-support strategy development and execution.
- **Drive new knowledge and information sharing,** including data platform performance and financial metrics, across the enterprise to decrease learning cycle times in different parts of the organization.
- **Consider a technology partner that addresses a broad set of decision-making capabilities** needed by executives, managers, business analysts, data scientists, frontline employees, and automated systems/bots.
- **Conduct a proof of concept and check references.** This is perhaps the most obvious recommendation, but we emphasize it because of the overwhelming number of claims by various technology vendors. For example, one of our reference interviewees, a data engineer, stated that "one of the solutions that we tested, which claimed auto-scalability, could only scale up, but not down."
- **Consider—again—the total cost of ownership,** including cost of compute and storage for highly variable as well as stable workloads, solution maintenance costs and staff utilization efficiency, and software subscription costs. Invest in system monitoring and a financial analysis and reporting solution for the data engineering team.

**Drive new information sharing across the enterprise to decrease learning cycle times.**

# About the Analyst

**Dan Vesset**
**Group Vice President, Analytics and Information Management, IDC**

Dan Vesset is group vice president of IDC's Analytics and Information Management market research and advisory practice, where he leads a group of analysts covering all aspects of structured data and unstructured content processing, integration, management, governance, analysis, visualization, and monetization. Dan also leads IDC's global Big Data and Analytics research pillar.

More about Dan Vesset

# Message from the Sponsors

Market leaders throughout the world have one common denominator: they are all winning with data. They leverage their data assets to learn from the past using business intelligence tools; they also focus on what is happening today and predict the future using real-time and streaming data sources combined with historical batch-type data sources.

Qubole is the open data lake company that provides a simple and secure data lake platform for machine learning, streaming, and ad hoc analytics. Qubole's Data Lake Platform provides openness and data workload flexibility, radically accelerating data lake adoption, reducing time to value, and lowering cloud data lake costs by 50 percent. Qubole's Platform provides end-to-end data lake services such as cloud infrastructure management, data management, continuous data engineering, analytics, and machine learning with near-zero administration.

Qubole partners with Google Cloud to enable companies to spur innovation and to transform their businesses for the era of big data.

**Learn more at qubole.com**

The Google Cloud strategy is to enable our customers to digitally transform, using our smart and innovative Google Cloud Platform, and to do this rapidly with our deep analytics and AI/ML capabilities. Google Cloud is widely recognized as a global leader in delivering a secure, open, intelligent, and transformative enterprise cloud platform. Our technology is built on Google's private network and is the product of nearly 20 years of innovation in data and analytics, artificial intelligence, machine learning, security, network architecture, collaboration, and open source software. We offer a simply engineered set of tools and unparalleled technology across Google Cloud Platform and G Suite that help bring people, insights, and ideas together. Customers across more than 150 countries trust Google Cloud to modernize their computing environment for today's digital world.

## About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

## IDC Custom Solutions

This publication was produced by IDC Custom Solutions. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

**IDC Research, Inc.**
5 Speen Street
Framingham, MA 01701
USA
508.872.8200

idc.com @idc



Doc. #US46799120