The 2021 State of Data and What's Next

ENTERPRISE MANAGEMENT ASSOCIATES® (EMA™)

By John Santaferraro

Prepared for Starburst and RedHat

February 2021

EMA™

IT & DATA MANAGEMENT RESEARCH,
INDUSTRY ANALYSIS & CONSULTING

The Great Digital Shift

The COVID-19 pandemic, along with universal global shutdowns, changed the way people live and do business. Organizations with strong digital heritage or successful digital transformation fared well in the downturn. As a result, competitors have been forced to quickly move to modern architectures that enable intelligent digital decisions for all.

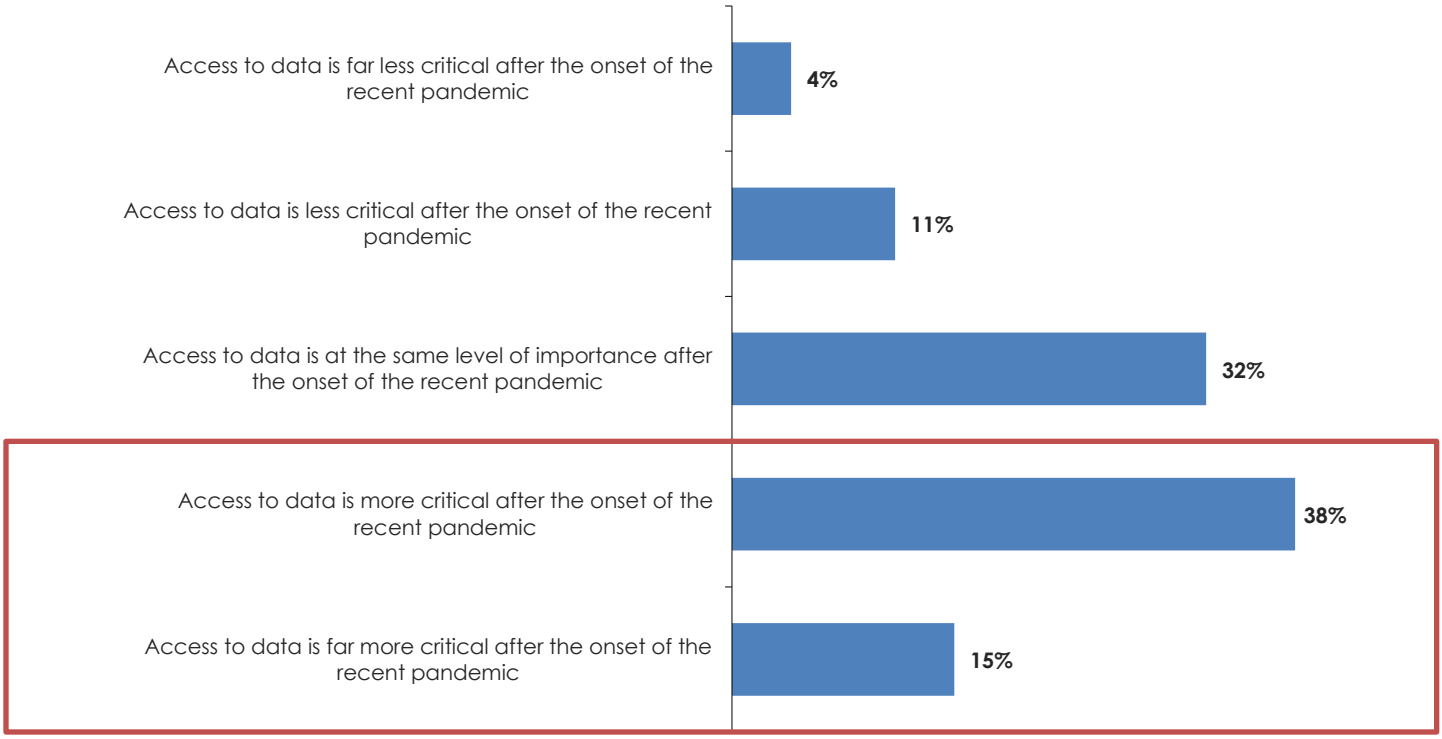
EMA predicts that 2021 will be the year of convergence. Organizations that modernize and converge on technologies like unified analytics and connected intelligence will be the winners. 2022 will emerge with clear winners and losers.

Convergence poses opportunities for consolidation, cost-cutting measures, and efficiency gains for these modern, post-pandemic organizations. The winners will operate as intelligence curators and brokers. Leading organizations will crush the competition with speed, depth, and agility.

THE IMPACT OF THE PANDEMIC ON DATA ACCESS

The road ahead is fraught with landmines and pitfalls. In the wake of the pandemic, access to data for decisions is even more critical. Fifty-three percent of respondents said that data access is more critical or far more critical after the pandemic.

What impact has the recent pandemic had on the need for data access at your company?



CONFIDENCE IN DATA ACCESS IS WANING

While data access is more critical than ever, many organizations struggle with the retrieval of timely, relevant data for analytics and decision-making. Thirty-seven percent of respondents are only somewhat confident, or not confident, in their ability to find and utilize insight.

37%

of respondents are only somewhat confident, or not confident, in their ability to find and utilize insight.

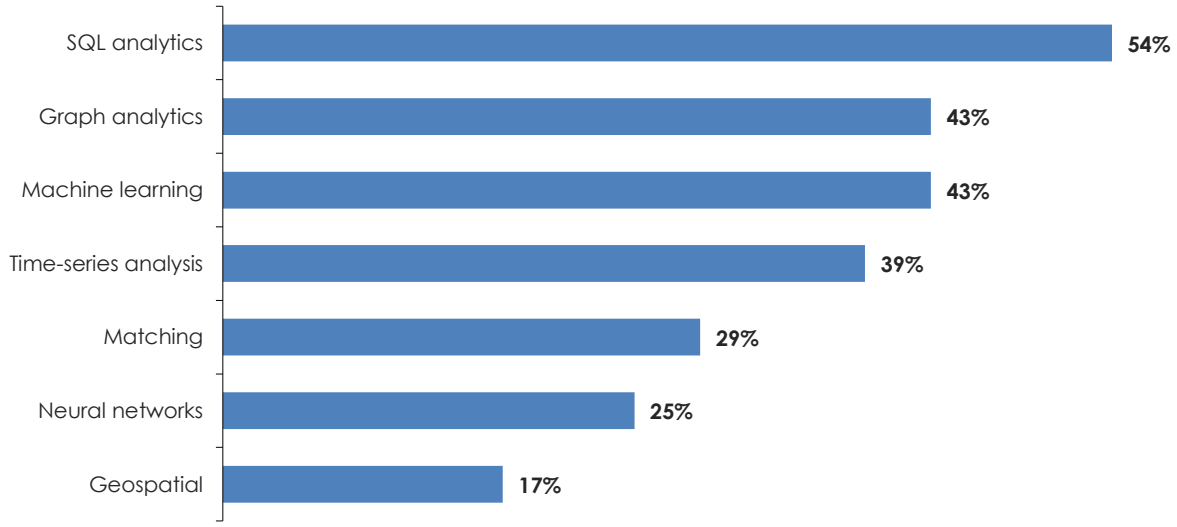
Analytical Priorities Remain Strong

Some things change and some things seem to stay the same. While the world shifts to digital business models and data changes drastically in both time and type, SQL remains the *lingua franca* of analytics. This unexpected surprise comes as companies shift major spending to new platforms that bypass the need for SQL.

EMA posed two questions regarding priorities for general analytics programs and priorities specifically for machine learning. In both cases, SQL emerged as the preferred language for business and technical users alike.

When asked, "Which types of analytical workloads are important to your overall analytics program?" 54% chose SQL analytics over options like graph, machine learning, and time-series analysis.

Which types of analytical workloads are important to your overall analytics program?



Even with the knowledge that digital data tends to include massive amounts of semi-structured data, there still remains the need for SQL access to multi-structured data.

EMA also delved into the overall machine learning operations process to discover what matters most to today's technologists. When asked, "Which aspects of machine learning are most important to your overall analytics program?" 46% selected support for SQL over choices like operations, monitoring, and model selection.

51%

of respondents indicated that they have five or more different data platforms.

Post-Pandemic – Risk Mitigation Has Risen as an Analytics Priority

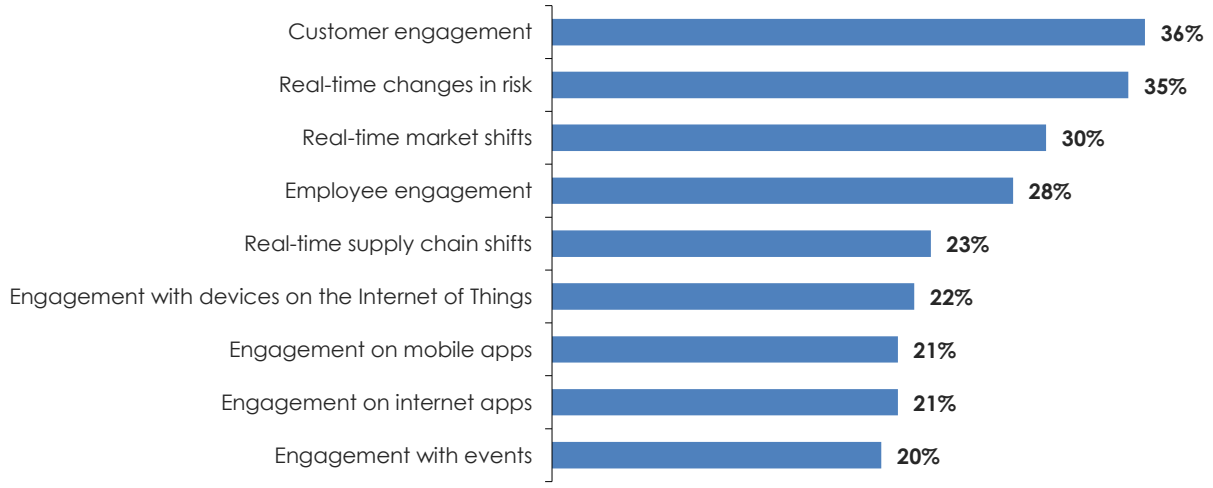
In 2020, the word "unprecedented" was used incessantly to describe the unexpected impact of the pandemic and shutdowns. Most business leaders had not anticipated the severity of government regulations or the length of regulatory controls. They were caught off guard.

As a result, 2021 brings an increased concern in risk at the same level of importance as customer relationships. Since the origin of popular analytics back in the nineties, customer analytics has remained the most important driver for investments in decision support. Customer analytics remains in the number-one position, but surprisingly, risk is now equally important to data access and analytics.

When asked, "What is driving your company's need for more real-time access to data or analytics?" the top reasons were customer engagement at 36% and real-time changes in risk at 35%.

Modern businesses are focused on two opposite but critical business initiatives: one about protecting the organization, the other directly tied to growth and revenue generation. Both areas require immediate responses to triggers and indicators as they occur.

What is driving your company's need for more real-time access to data or analytics?



The Great Data Dispersion

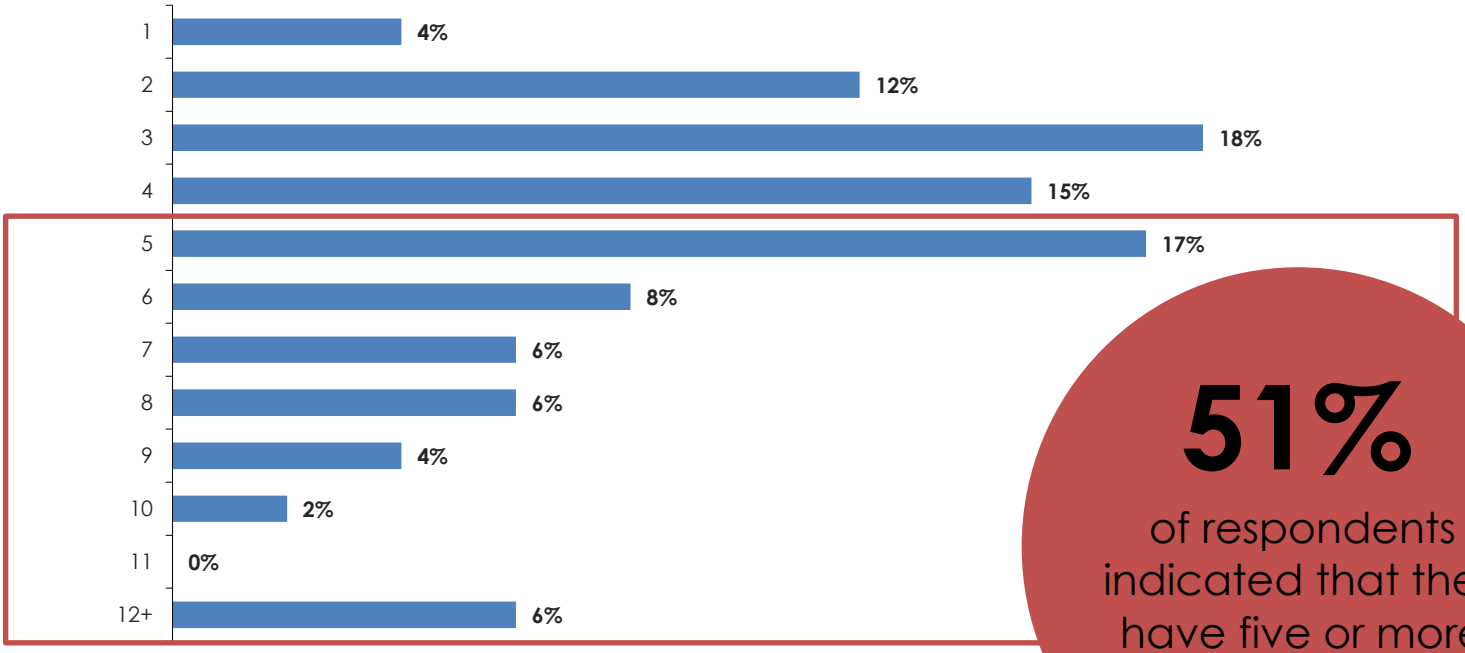
The lack of confidence in finding data, combined with the preeminence of analytics and the rise of risk, elevates the importance of data access and location. Unfortunately, data continues to be highly dispersed.

A decade ago, relational databases struggled to handle new digital data. As a result, many organizations added data lake technology. Since that time, data continues to expand, and the platforms used to harness and harvest insight also continue to multiply.

In understanding the trend toward more data platforms, it is important to realize that some data platforms can have hundreds of instances. Additionally, every new platform or instance adds more cost and complexity. The result is a quagmire of information scattered throughout organizations that have spent millions on data warehouses and data lakes.

When asked about the current number of data platforms, 51% of respondents indicated having data in five or more different platforms, which is an increase of almost 60% from 2019.[1] When asked about the number of platforms in the next 12 months, the number of respondents with five or more platforms increased to 56%, which is an additional increase of 10%.

How many different data platforms do you currently have in your data ecosystem?



51%

of respondents indicated that they have five or more different data platforms.

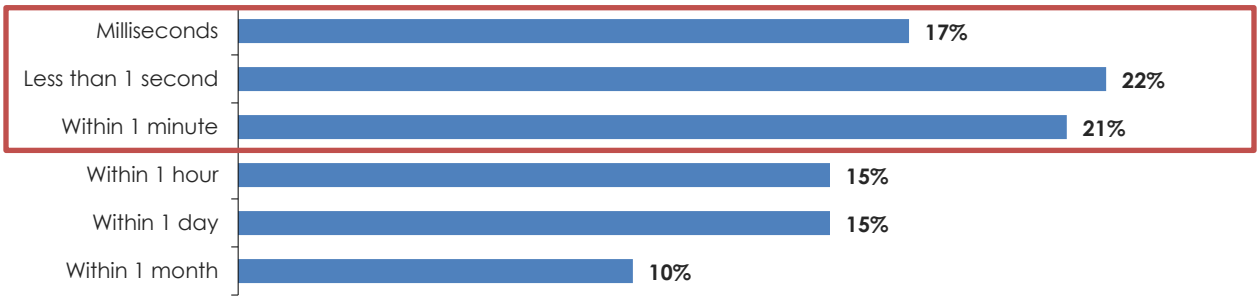
[1] "Modernization and the Operation of Hybrid Data Ecosystems," EMA, July 2019

The Need for Speed

To further complicate the challenge of dispersing data, the need for speed is vital for organizations pursuing digital business success in the post-pandemic world. Digital supremacy requires rapid, intelligent responses to engagement events with customers, partners, and employees. In addition, complex markets that have traditionally been difficult to understand also now require immediate understanding to avoid excessive risk.

EMA asked participants to break down their business decisions by the amount of time they had to respond to critical events. On average, 50% of all business decisions need a response within one minute. Seventeen percent of all decisions need a response in milliseconds.

What percentage of your business decisions require latency at the following levels?

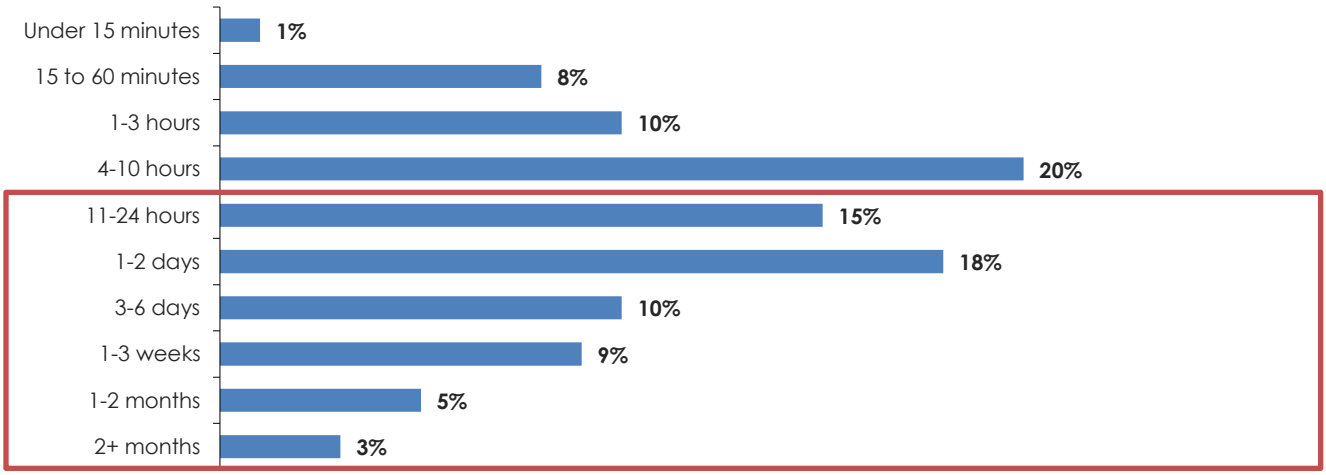


The Data Pipeline Dilemma

Digital success requires rapid responses to business events, and data pipelines are needed to process and deliver insight. Unfortunately, data pipelines can take too much time to develop and place into production, creating a backlog of data and preventing real-time decisions.

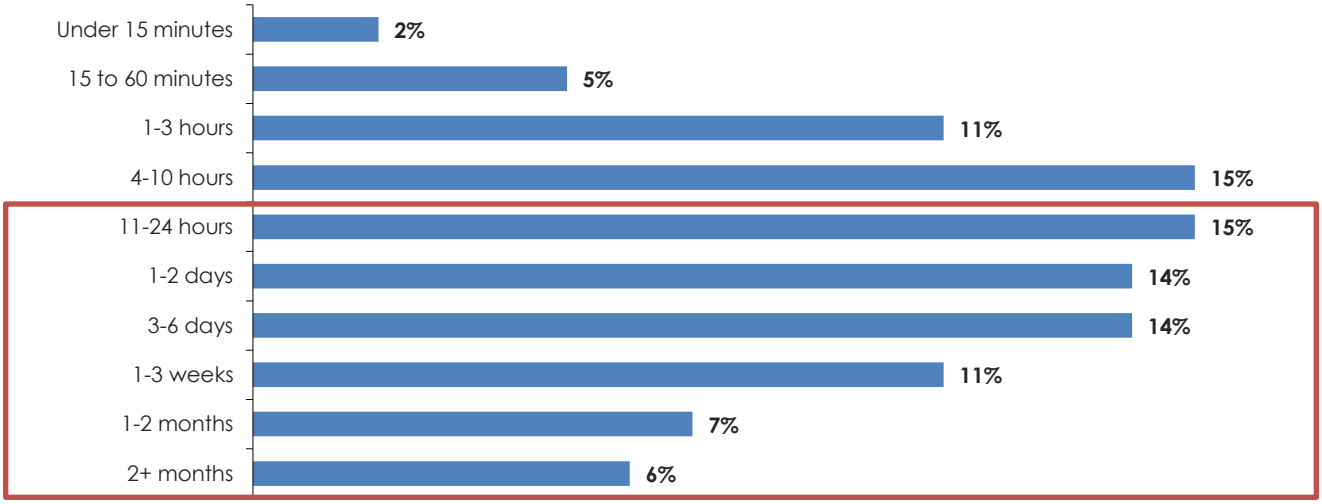
Considering that many data pipelines are created by developers without the right tools or platforms to speed and automate development, 60% of respondents said that they take more than a business day to develop a data pipeline, with 27% in the three days to two months range.

On average, how long does it take to develop a data pipeline?



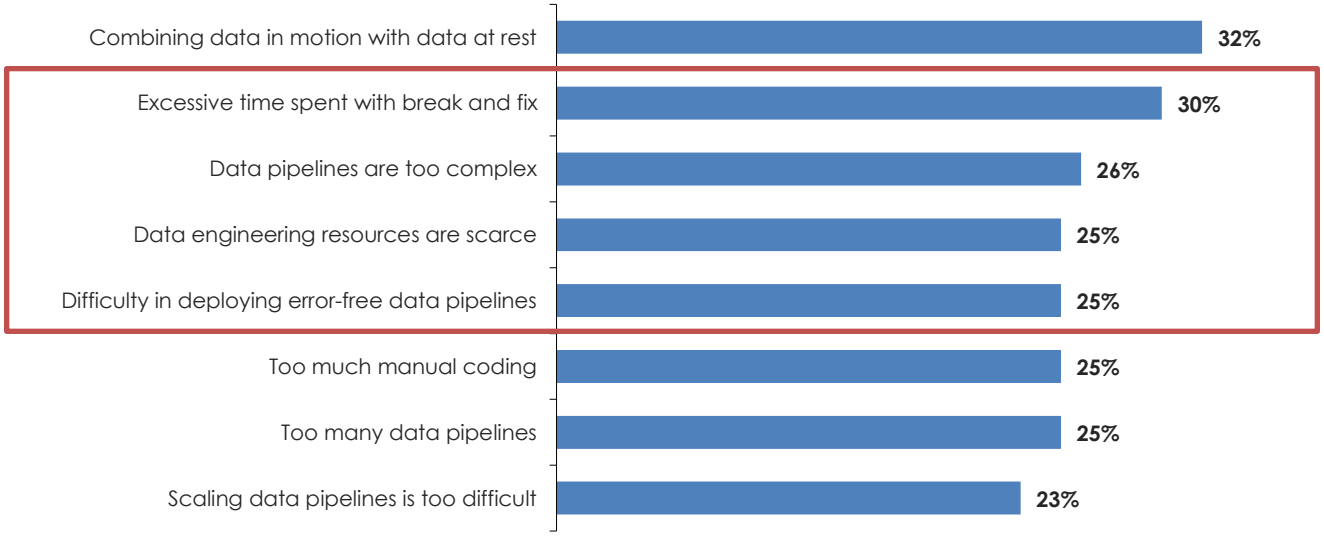
In addition, making a data pipeline operational takes even more time. A full two-thirds of respondents said it takes at least another full business day to get new data pipelines into production, with 24% indicating that production takes more than a week.

On average, how long does it take to make a data pipeline operational in production?



What is it that makes data pipelines so difficult? While the number-one answer given by participants was combining data in motion with data at rest, the next four answers give the real story. Data pipelines are inherently complex, and most companies are using too much manual coding. As a result, they have difficulty deploying error-free data pipelines and end up spending excessive time on fixing broken pipelines.

What are the biggest challenges you face in building and deploying data pipelines?



The Reign of the Cloud

The first and most obvious solution for data dispersion and multi-platform complexity is a move to the cloud. This move to the cloud has been accelerated by the testing of digital business models in the global shutdowns.

Respondents cited that today, 56% of their data is in the cloud versus 44% on-premises. In 12 months, they expect 62% of their data to be in the cloud.

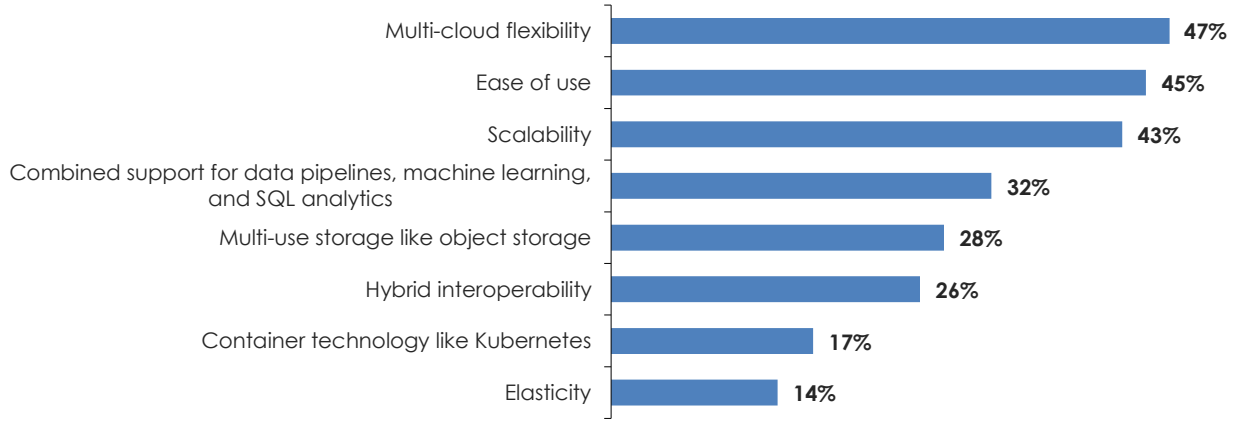
56% of respondents' data is currently in the cloud.

62% of respondents' data will be in the cloud in 12 months.

To further support this massive move to the cloud in 2021, EMA conducted separate research on the impact of the pandemic on IT in general. In that research, 71% indicated that they have accelerated their cloud migration schedule.[2]

Cloud migration also includes a search for solutions for multi-cloud interoperability. As part of the journey to the cloud, the number-one impact on buying decisions is multi-cloud flexibility. This anticipates EMA's recommendation of unified analytics solutions in the next section.

Which aspects of cloud data storage and access most impact your buying decisions?



[2] Best Practices for the Enterprise: Information Security and Technology Trends Responding to the Pandemic," EMA, January 2021

The Rise of Unified Analytics

To address the complexity and challenges of operating hybrid data ecosystems, many organizations are modernizing their ecosystem by running advanced analytics through a distributed SQL query engine running on a unified analytics platform.

When asked which technologies were the most important for making decisions based on data located across multiple platforms, the top three answers defined the most suitable modern architecture for digital decisions:

- 34% Advanced analytics platforms
- 27% Unified analytics platforms
- 26% Distributed SQL query engines

> The ability to store data once and query it many different times for all types of analytical workloads is critical to a successful analytics program.

The Emergence of Autonomous Analytics

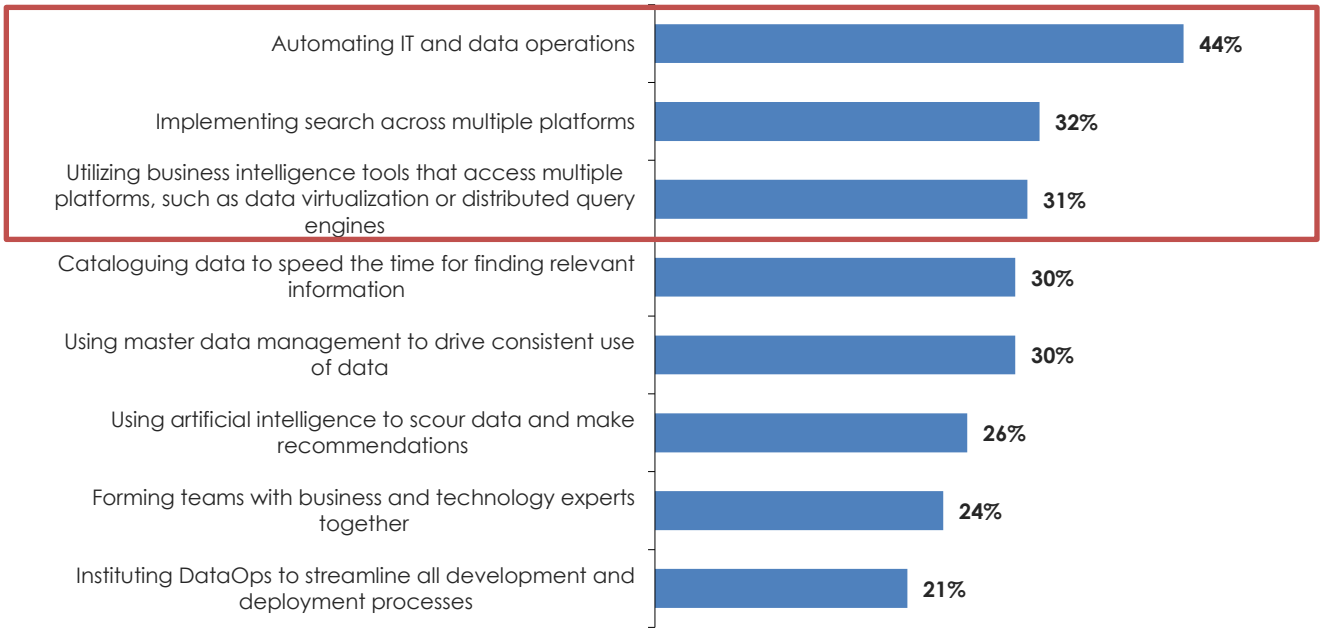
The current push toward digital superiority as the post-pandemic mandate has multiplied the need for advanced analytics in decision making. As a result, most IT organizations are unable to keep up with demand. Data workers of all types are being asked to do more with less. It is no surprise that resource constraint is the number-one driver for automation across the entire information supply chain. Sixty-five percent of IT leaders and practitioners indicated that the need for additional resources is the main driver behind the use of AI-enabled analytics.[3]

EMA has been tracking the use of machine learning in data management and analytics platforms for the last two years. During that time, there has been an increase in the number of formerly manual functions that are now completely automated. AI-enabled analytics promise to scale up data and analytics programs to impact more business decisions with fewer resources.

With data spread throughout organizations, the top requirements for platform interoperability are the automation of data pipelines, the adoption of intelligent search, and the implementation of distributed query engines.

[3] "Innovation in the Use of Artificial Intelligence and Machine Learning for Metadata Services," EMA, 2019.

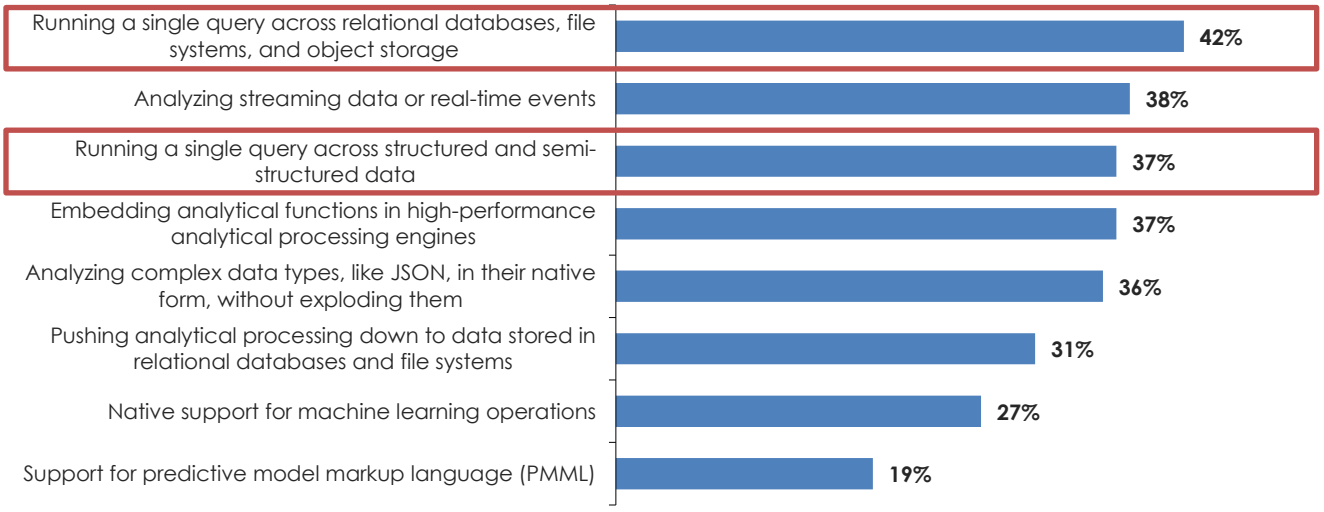
With data spread across your company in different systems, what practices are most important to making the systems work together?



The Power of Modern Analytics

Modernization is at its peak, with new technology speeding access to data stored in disperse locations and lowering the cost of storage. The combination of distributed query engines and object storage may be the wave of the future. The top priority for analytics programs in 2021 is the ability to run queries on multi-structured data across multiple tiers of data storage.

Which modern analytical capabilities are important to your overall analytics program?

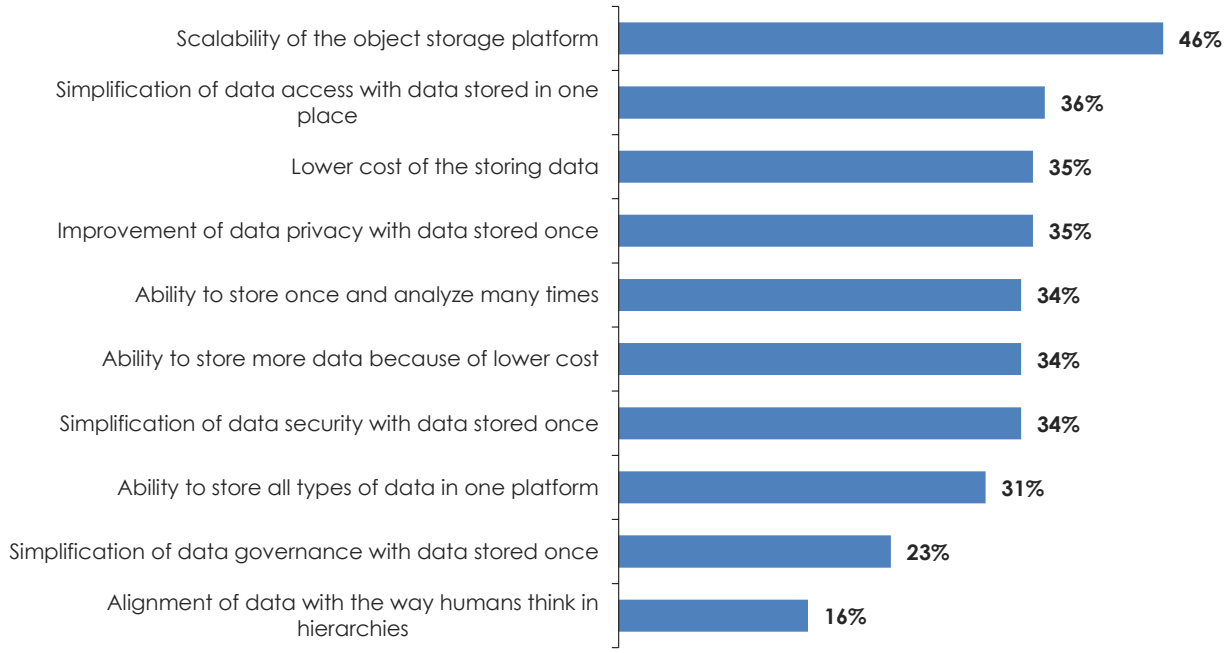


While the analytical focus remains on SQL access to multi-structured data anywhere, many companies are making a move to object storage. Out of 402 participants, 149 indicated that they have already invested in object storage, or that they plan to do so in the next 12 months.

Although object storage was introduced in the mid-2000s, it has gained considerable adoption in the last five years with the introduction of analytical products that are able to analyze stored data. Scalability, simplification of data access, and a significant reduction in cost led buying decisions for object storage in 2020.

Based on the ability to store data once in object storage and analyze it many times with a distributed query engine, EMA expects to see a rapid increase of adoption from 2021 forward. The fact that organizations can deliver multiple analytical solutions and a high volume of data pipelines without having to copy and move the data will be the key driver.

Which capabilities drove your decision to utilize object storage?



(Sample Size = 149, Valid Cases = 149, Total Mentions = 482)

About Starburst and RedHat

Every large organization in the world suffers from a data silo problem. Traditional data warehouse products approach the problem with outdated, monolithic solutions that breed inefficiency and ultimately cannot help business analysts run fast analytics on all their data. This prevents the business from making better and more timely decisions to improve their company's performance.

Companies everywhere are building distributed cloud and hybrid cloud applications. Many of these organizations rely on both traditional applications and modern applications to run their business and make critical business decisions. Likewise, their data sources include both traditional data sources and new data sources located everywhere - in data centers, cloud, and even vendor environments.

Instead of rebuilding data infrastructure from scratch add a straightforward, easy-to-manage and operate, real-time, distributed query engine that accesses your data no matter where it resides, in whatever form it appears in.

Starburst & RedHat provide modern solutions that address data silo & speed of access problems, With Starburst's fast, distributed query Trino engine, Starburst Enterprise, and the leading enterprise Kubernetes platform, Red Hat OpenShift organizations can now run analytics anywhere to make better business decisions with minimal additional load on your operations teams.

Starburst Enterprise provides distributed query support for varied data sources, such as Apache Cassandra, Hive (HDFS), S3 (HDFS), Microsoft SQL Server, MySQL, and PostgreSQL data sources. Starburst Trino Operators delivered with Red Hat OpenShift Container Platform automate installation, upgrades, and lifecycle management throughout the container stack.

Together, Red Hat & Starburst provide a simple, cost-effective, straightforward way to manage architecture that gives organizations fast access to all their data to make better business decisions on more complete data.

About Enterprise Management Associates, Inc.

Founded in 1996, Enterprise Management Associates (EMA) is a leading industry analyst firm that provides deep insight across the full spectrum of IT and data management technologies. EMA analysts leverage a unique combination of practical experience, insight into industry best practices, and in-depth knowledge of current and planned vendor solutions to help EMA's clients achieve their goals. Learn more about EMA research, analysis, and consulting services for enterprise line of business users, IT professionals, and IT vendors at www.enterprisemanagement.com or blog.enterprisemanagement.com. You can also follow EMA on Twitter, Facebook, or LinkedIn.

Corporate Headquarters:
1995 North 57th Court, Suite 120
Boulder, CO 80301
Phone: +1 303.543.9500
www.enterprisemanagement.com

0000.02092021