FORRESTER®

Databricks Platform Increases Business Agility And Drives Better Outcomes

As Forrester reports in its insights-driven business playbook, data leaders are quickly discovering that "simply putting lots of diverse data into the cloud or a data lake won't magically create meaningful insights — not without further integration, transformation, enrichment, and orchestration."[1] But doing all of that at scale is far from trivial, and doing it poorly leads to "poor business decisions, bad customer experience, reduced competitive advantage, and slower innovation and growth."[2]

To get the most value, firms must democratize all data (structured, semi-structured, unstructured, and streaming), evolve processes, restructure teams, challenge cultures, and rethink their data platforms to support data teams and business users alike.

The value of open-source standards

Data analysts, data engineers, data scientists, business users, and machine learning (ML) models all need real-time access to the most up-to-date data. To ensure they're maximizing their data's value, data leaders should make their data open and accessible across different tools, programming languages, and systems. Increasing data availability, upskilling data analysts and data engineers, and empowering them to do data science and analytics work will lead to more insights and more performant ML models faster than before.

Databricks platform

Achieving a data-driven culture at scale is beyond most organizations' reach because of the complexity and cost of their existing methodologies and technology stacks. That's why data leaders are rapidly adopting Databricks' platform for their organizations. The Databricks platform is an open platform that brings together the performance of a data warehouse and the scalability and cost-efficiency of a data lake. The Databricks platform simplifies analytics and AI, enables large-scale data engineering, improves collaboration across teams, and supports ML development across the full lifecycle.

KEY STATISTICS

Return on investment (ROI)
417%

Net present value (NPV)
$23.3 million

Increased data scientist productivity:
25%

Increased revenue:
5% by Year 3

Decreased compute time:
40%

To better understand the benefits, costs, and risks associated with the Databricks platform, Databricks commissioned Forrester Consulting to interview four customers and conduct a Total Economic Impact™ (TEI) study.[3] Forrester interviewed an additional three customers to understand the specific challenges that data leaders face. These additional interviewees included:

- The vice president of architecture for a financial services firm. The vice president's department has 300 data team members (70 of whom are data scientists or data engineers) and 10 terabytes of data.
- The director of data science and analytics for a technology company. Their organization has 12 data scientists or engineers and 5 petabytes of data.
- The head of AI engineering for a retail firm. The retail firm has 70 data scientists and data engineers.

This abstract focuses on the challenges data leaders face and how the Databricks platform helps them overcome them.

> **"If you have a business whose lifeblood is data, and you limit the number of people who can leverage the data, you're creating bottlenecks and harming the business."**
>
> *Vice president of architecture, financial services*

INVESTMENT DRIVERS

The data leaders interviewed explained that a combination of technological complexity and organizational bottlenecks reduced the speed of innovation and impeded their organizations' abilities to deliver on new business initiatives. These challenges included the following:

- **Legacy solutions limited adoption and hindered innovation.** The technical complexity of the previous data analytics platforms reduced the number of people who could leverage their organization's data, reducing the impact these solutions had and creating bottlenecks. The vice president of architecture said: [Although there were over a thousand data users,] only the 40 data scientists and data engineers could answer a question or review a process. This really limited the operationalization and adoption [of our previous solution]."

  Because it was impractical to train users and even many developers on these legacy platforms, data scientists spent more time supporting data users to create or troubleshoot reports, and they spent less time creating or iterating on ML models.

  The interviewed data leaders said they hope to increase innovation by making it easier for more users to leverage company data to create insights or develop data applications. Democratizing analysis and development would also free up valuable data scientist time.

- **Traditional data governance applications failed to meet the organizations' needs.** The interviewees noted that quality issues could go undetected for more than a month at a time. Such issues degraded the performance of ML models and reduced the value of analyses. The longer these data issues went unnoticed, the more time-consuming and expensive it was to remediate them. Data quality issues can also increase customer churn and harm the reputation of organizations that are commercializing their data.

- **Legacy processes created bottlenecks and lengthened project timelines.** Previously, data science teams relied on engineering teams to

create test environments and production environments. These processes were slow and time-consuming. Since data scientists and engineers used different programming languages and methodologies, they also struggled to understand each other.

Business users — even those with experience with a programming language such as Structured Query Language (SQL) — faced similar slowdowns because they had to rely heavily on data analysts and data science teams to perform analyses or to troubleshoot existing analyses.

- **Previous modes of collaboration were ineffective.** In a Forrester Analytics Business Technographics® survey, 22% of enterprise technology decision-makers cited a lack of collaboration as the biggest challenge in executing their vision for data, data management, and analytics (ranking second behind maturity of technology around security).[4] Creating effective data processes, analytics, and ML models that drive business outcomes requires collaboration among data scientists, data engineers, data analysts, app developers, and business stakeholders. Without the right tools to collaborate, organizations will struggle to maximize the value of their data.

  The data leaders also explained that existing collaboration tools would often fail security policies. Data scientists shared notebooks and passwords through email, and they lacked version control capabilities. This led to rework.

  The actual experimentation process was also slow and inefficient. Data scientists spent significant amounts of time cleaning and preparing data, waiting for processes to finish, or manually tracking experiments. The head of AI for a retail firm explained that 10% to 20% of a data scientist's time could be spent keeping track of one's experiments.

- **Existing infrastructure failed to meet organizational needs while being costly to maintain.** Infrastructure constraints prevented workers from starting new projects, and they slowed the progress of existing projects. Despite massive infrastructure investments, data leaders lacked the computing resources their teams needed to finish processes in a timely fashion. For example, the vice president of architecture for a financial services firm explained: "We had to reaggregate the 12 years of data. Using our legacy relational database and the app built around it, it would have taken us two years to go through that historical data." Meanwhile, the head of AI engineering reported that the company's data analytics infrastructure went unused 90% of the time.

Increased revenue due to the faster creation and optimization of ML models with Databricks:

5%

KEY RESULTS

By leveraging the Databricks platform, interviewees were able to achieve the following key results:

- Increased revenue by 5%.
- Gained up to 25% of productivity.
- Reduced infrastructure costs by $11 million.

Increased revenue by 5%. Databricks enabled the interviewees' organizations to spend more time creating and improving ML models. Databricks also enabled data teams to use cutting-edge ML models (e.g., deep-learning models) that were previously not accessible to them. The combination of more and better ML models and insights enabled the data leaders to increase revenue by implementing the following:

- **Increased sales through improved customer experience.** By increasing the number and efficacy of their recommendation engines for various buying personas, both retailers were able to increase order frequency. One retail leader calculated that every percentage increase in order frequency resulted in a 2.8% increase in revenue. That data leader could also identify customer pain points that led to lost sales. For example, the retailer found that more than a third of cart abandonments were due to forced account creation. By eliminating account creation requirements, the retailer recognized a 4% increase in revenue.
- **Optimized pricing.** The principal architect for a heavy equipment manufacturer said: "Our parts pricing team moved from a manual process to a more automated process using Databricks. [Team members] estimate that the optimized pricing has resulted in a few million dollars in profit, but they haven't priced near the number of parts that they want to in the long run. So, we believe that this opportunity is much more significant than a few million dollars a year."

Data scientist productivity increase

25%

Gained up to 25% of productivity. Democratizing data analytics and empowering users with the right tools enabled data analysts, data scientists, and data engineers to be 25% and 20% more productive. The Databricks platform enabled data scientists and data engineers to spend less time searching for and cleaning data and creating and maintaining extract, transform, load (ETL) pipelines. They also spent more time building and improving ML models that could drive business outcomes. The interviewees' organizations achieved these benefits by doing the following:

- **Improved data quality.** The interviewees' organizations leveraged Delta Lake (which adds reliability, performance, and lifecycle management to data lakes) to keep their data clean and performant by enforcing governance standards and by monitoring their data quality more frequently. These actions improved the quality of the analyses conducted and ML products created.

  Identifying data quality issues faster reduced the internal labor and computing costs required to remediate these issues. For instance, the director of data science and analytics explained: "Before we had the monitoring capabilities on Databricks, we had a month where it cost us $17,000 to fix a data issue. Since [using] Databricks, our cumulative cost for all of 2020 was less than $5,000."

> **"We can now identify data quality issues immediately and rectify them without adversely impacting our customers. This improves our reputation with our customers. It also helps us when talking to potential customers since we can show them our dashboards to prove our data is high quality."**
>
> *Director of data science and analytics, software company*

- **Made data more performant.** After migrating to Databricks, data analysts and data scientists spent less time searching for data, cleaning it, and sifting through results, and they spent more time on activities that drove business outcomes. Once they could analyze their entire data set, users uncovered new insights and business opportunities.

> **"We were able to get a product to market in a month using only Databricks. Without [the Databricks platform], the product wouldn't have been feasible; the R&D alone would have taken us two years at our previous pace."**
>
> *Director of data science and analytics, software company*

- **Lowered the barrier to entry to work with data.** Databricks supports a wide range of modern programming languages (including R, Python, Scala, and SQL), and this reduces the technical requirements for data teams to leverage their data. This allows users to work in the programming language they're most comfortable with, which increases engagement. The vice president of architecture said, "[It] opens the doors to all sorts of use cases, applications, and innovations that you might not have seen prior."
- **Supported self-service.** Providing data scientists and data analysts self-service capabilities enabled users to be more self-sufficient, which reduced friction and accelerated project timelines. Data scientists could now create test environments and orchestrate data without interfacing with infrastructure teams. Likewise, business users could now access data in real time or near-real time and create their own dashboards and insights without relying on valuable resources for standardized analytics and business intelligence (BI) requests.
- **Optimized data scientist workflows.** Making analysts and business users more self-sufficient reduced the time data scientists had to spend supporting these users.

  The collaborative notebooks available through Databricks made teams more productive. For instance, the head of AI engineering explained that data teams created notebooks to track noteworthy data, and that made the initial discovery process four times more efficient.

  Meanwhile, Managed MLflow from Databricks saved data teams significant amounts of time managing the complete ML lifecycle. The head of AI estimated that MLflow reduced the time data scientists spend tracking their experiments by half while reducing the possibility of rework from poor project tracking.

  These efficiency gains enabled data scientists to spend more time developing and iterating on ML models. And, as the vice president of architecture explained, this has the added benefit of increasing job satisfaction. They said: "[Data scientists] didn't get drawn into data science to write ETL jobs basically or to struggle finding and sourcing data. That's not what drives them or what excites them. [Reducing the time they spend on these tasks] dramatically improved their job satisfaction."

> **"Without Delta Lake, we wouldn't have been able to learn enough about our data to know if a recently developed product was viable. But even if we could have done that, it would have taken us years to build, and it would have been much more expensive to get anywhere near as performant as we've got it using Delta Lake."**
>
> *Director of data science and analytics, software company*

Reduced infrastructure costs by $11 million. By moving to Databricks, the interviewed data leaders noted that their organizations could retire their on-premises infrastructures, cancel redundant software licenses, and reallocate IT resources. Managing the

platform proved substantially more straightforward than with prior data platforms.

- **Decommissioned legacy infrastructure and services.** Before adopting Databricks, most of the interviewees' organizations managed extensive on-premises data analytics environments. The organizations either had hundreds or thousands of commodity servers running open-source solutions or expensive servers running proprietary software — or both in some cases. To continue to meet their organizational needs, decision-makers had to keep buying more servers, which created more management overhead. Moving to the cloud did little to bend the cost curve. One executive noted that cloud costs increased by more than 16% each year to keep pace with growing demands.

  With Databricks, the data leaders said their organizations retired their on-premises infrastructures and began reducing or canceling third-party licenses and services. Moreover, Databricks reduced administrative costs. Engineers no longer had to worry about maintaining or upgrading the platform.

  Databricks further reduced costs by enabling organizations to retire adjacent systems. For example, some interviewees reported copying data from their data lakes to relational databases in data warehouses to support business users who relied on SQL. This further increased organizations' infrastructures and support costs while failing to adequately meet business users' needs since they didn't have real-time access to all of the data they needed, nor could they adequately troubleshoot issues when they arose. Instead, Databricks enabled users to perform these SQL queries directly on the data in the data lakes.

- **Provided a more stable and performant environment.** Processes on the Databricks platform were finished in a fraction of the time they took with previous solutions. For example, the heavy equipment manufacturer's principal architect noted a 40% decrease in compute time and a 97% decrease in time needed for the Apache Spark processes to finish. The vice president of architecture explained that a process that would have taken two years to complete on the legacy relational database took eight days to complete on Databricks.

KEY TAKEAWAYS

By harnessing and applying data analytics, data science, and ML at every opportunity with quality data, organizations can grow revenue, optimize, and differentiate existing products and identify new business opportunities. Based on the success the interviewed data leaders spoke about, organizations should:

- **Focus on a unified and open data architecture for data usability.** Giving users access to company data isn't enough to drive key business outcomes. Data leaders need to support a wide range of workflows within their data architectures with real-time access to quality data. For example: allowing data analysts to use SQL and data engineers to use Python to maximize their data's value. Similarly, upskilling users will be relative based on a user's technical experience and interest.

  Enable a self-serve and collaborative data-driven culture. Data leaders need to make their data architecture an enabler for innovation by providing collaborative and self-service capabilities to both data teams and business users. Reducing complexity by automating data delivery and enabling teams to operate from the same view of data will reduce rework and result in more timely and better-informed decisions and more performant products and ML models. The goal is to maximize an organization's data literacy and data culture by maximizing who can leverage its data.

TOTAL ECONOMIC IMPACT ANALYSIS

For more information, download the full report "The Total Economic Impact™ Of Databricks Lakehouse Platform," commissioned by Databricks and delivered by Forrester Consulting.

STUDY FINDINGS

Forrester interviewed decision-makers from organizations with experience using the Databricks Lakehouse Platform, and combined the results into a three-year composite organization financial analysis. Risk-adjusted present value (PV) quantified benefits include increased revenue (5%), improved data scientist and engineer productivity (25% and 20%, respectively), and retiring of on-premises infrastructures. This saves the composite millions of dollars annually.

Return on investment (ROI)

417%

Net present value (NPV)

$23.3 million

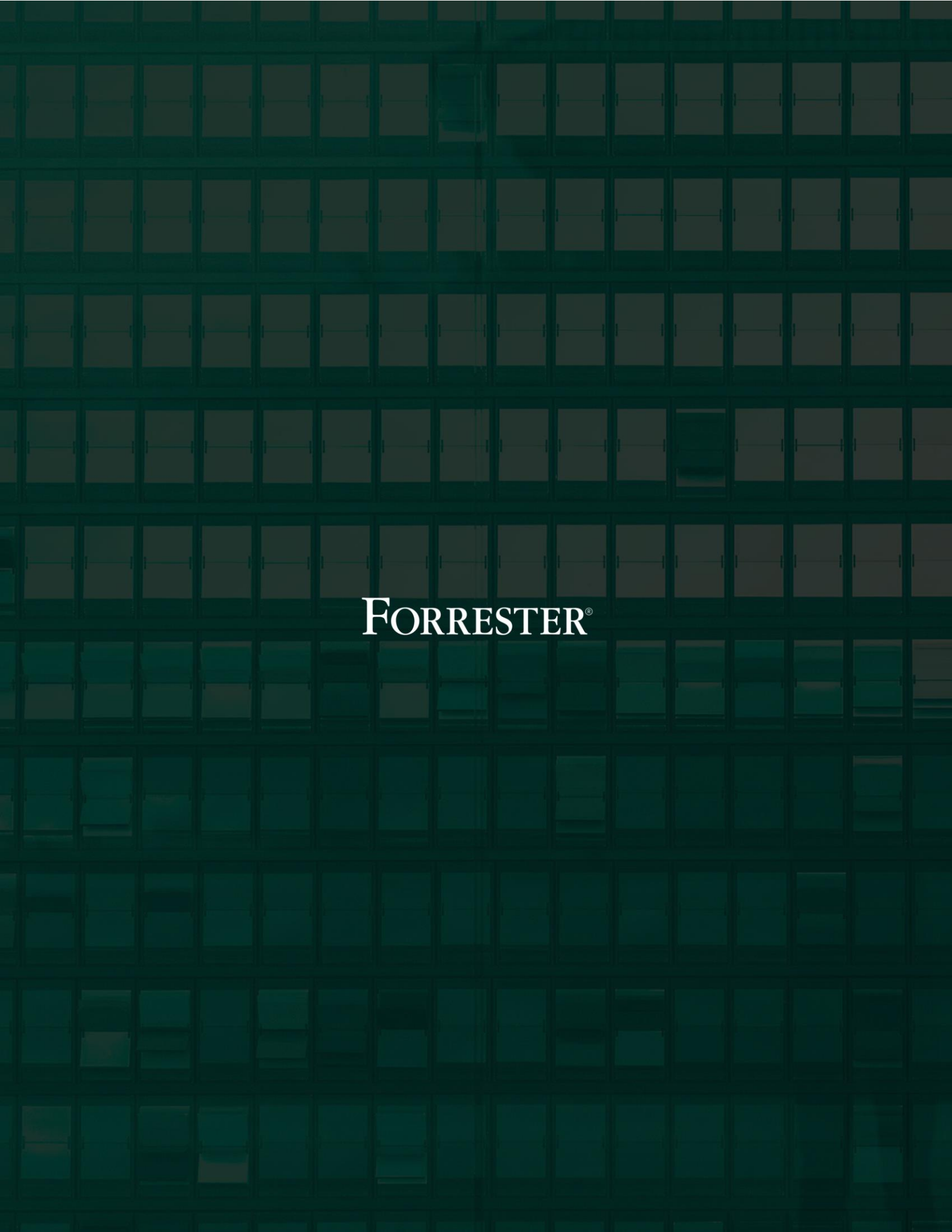
Appendix A: Endnotes

[1] Source: "Enterprise Data Fabric Enables DataOps," Forrester Research, Inc., December 23, 2020.

[2] Ibid

[3] Total Economic Impact is a methodology developed by Forrester Research that enhances a company's technology decision-making processes and assists vendors in communicating the value proposition of their products and services to clients. The TEI methodology helps companies demonstrate, justify, and realize the tangible value of IT initiatives to both senior management and other key business stakeholders.

[4] Source: Forrester Analytics Business Technographics® Data And Analytics Survey, 2020.


FORRESTER®