A Forrester Total Economic Impact™
Study Commissioned By Databricks
April 2020

The Total Economic Impact™ Of The Databricks Unified Data Analytics Platform

Cost Savings And Business Benefits Enabled By Databricks Platform

FORRESTER®

Table Of Contents

Project Director:
Edgar Casildo

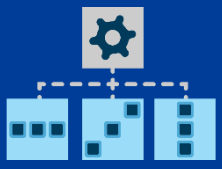
ABOUT FORRESTER CONSULTING

Forrester Consulting provides independent and objective research-based consulting to help leaders succeed in their organizations. Ranging in scope from a short strategy session to custom projects, Forrester's Consulting services connect you directly with research analysts who apply expert insight to your specific business challenges. For more information, visit forrester.com/consulting.

FORRESTER®

Benefits

Increased revenue:
$9,895,520



Increased operating efficiency:
$7,677,643

Legacy data analytics platform cost savings:
$11,298,202

Executive Summary

Forrester reports that: "In today's hypercompetitive business environment, harnessing and applying data, business analytics, and machine learning at every opportunity to differentiate products and customer experiences is fast becoming a prerequisite for success."[i] Much like the internet before them, AI and machine learning (ML) represent both the biggest threat and the biggest opportunity for enterprises today.[ii] Organizations will need to leverage AI and ML to optimize products, services, and operations to succeed.

Companies need to be able to effectively align analytics and ML investments with key business priorities to drive success. Similarly, they need to align data management strategies with business priorities and remove any constraints or barriers to ensure the success of data teams as they execute on these priorities.

Organizations must democratize data (both structured and unstructured), evolve processes, restructure teams, challenge cultures, and re-architect technology stacks to support data team success. To make real-time decisions, organizations will need access to the most up-to-date data. Moreover, the data should be open and accessible across different tools and systems within an organization — not locked into closed formats or proprietary systems inaccessible by other technologies or teams. Training and empowering data engineers and analysts to do data science work will increase productivity by eliminating bottlenecks caused by relying solely on data scientists.

Achieving all this at scale is beyond the reach of most organizations today because of the complexity and cost involved in meeting these requirements. That's why organizations are rapidly adopting Databricks' open, unified data platform that simplifies data and AI for massive-scale data engineering, collaborative data science, full-lifecycle machine learning, and business analytics. Databricks commissioned Forrester Consulting to conduct a Total Economic Impact™ (TEI) study of Databricks' current customers and determine the return on investment (ROI) enterprises may realize by deploying the Databricks Unified Data Analytics Platform. The purpose of this study is to provide readers with a framework to evaluate the potential financial impact of the Databricks platform on their organizations.

Prior to using Databricks, the customers had various data processing, business analytics, and ML technologies spread across on-premises and cloud environments. Siloed data, difficult-to-use systems, operational friction, and lack of scalability hindered data teams' abilities to collaborate effectively to unlock business value from their data. Data teams found themselves spending most of their time setting up and managing systems instead of driving data-driven innovation and business outcomes.

Key Findings

Quantified benefits. The following risk-adjusted present value (PV) quantified benefits are representative of those experienced by the companies interviewed:

ROI
417%

Benefits PV
$28.9 million

NPV
$23.3 million

Payback
<6 months

- **Increased revenue by 5%.** Databricks enabled the interviewed customers' data scientists to spend more time creating and improving ML models. Databricks also enabled data scientists to use cutting edge ML models, such as deep learning models, that were not accessible to them previously. Additionally, by democratizing data access, the interviewed organizations saw new users create a diverse set of new ML models and derive more insights than before. The combination of more — and better — ML models drove increased revenue for many of the interviewees.
- **Improved data team productivity by 25% and 20%, respectively.** Databricks enabled teams of data scientists and data engineers to spend less time searching for and cleaning data, and creating and maintaining extract, transform, load (ETL) pipelines, and spend more time building and improving ML models that could drive business outcomes. Databricks also helped remove technical barriers that limited collaboration between analysts, data scientists, and engineers, enabling data teams to work together more efficiently.
- **Retired their on-premises infrastructures, saving millions of dollars annually.** By migrating to Databricks, the interviewed organizations were able to retire their on-premises infrastructures and cancel their now-redundant data analytics licenses.

Unquantified benefits. The interviewed organizations experienced the following benefits, which are not quantified for this study:

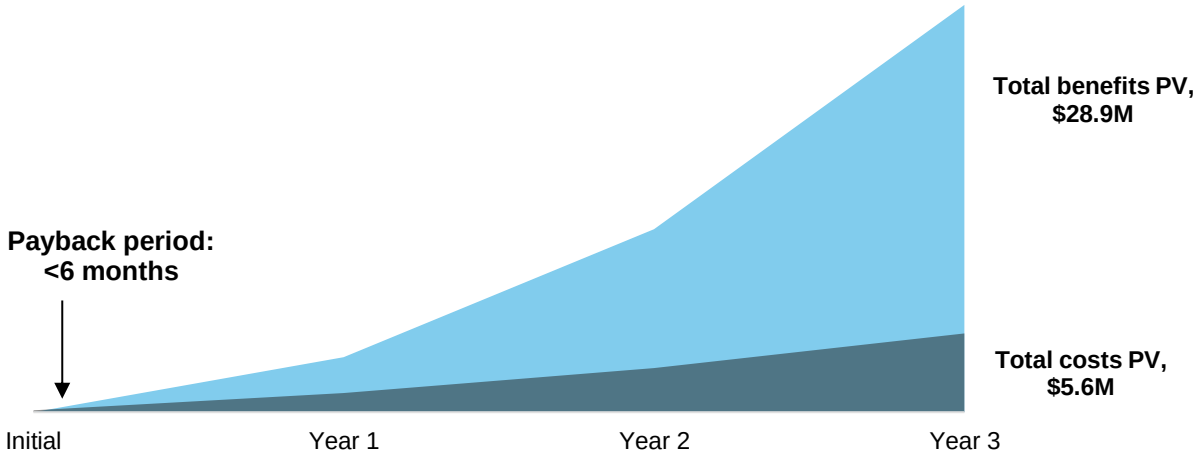
- **Reduced operating costs.** The variety of features that Databricks offers enabled the interviewed organizations to reduce their spending on raw materials and identify manufacturing issues earlier, which reduced warranty repair costs and return rates.
- **Improved security.** Before Databricks, the interviewed organizations lacked security standards across their data analytics environments. Databricks provided native security protection, which alleviated security concerns and allowed the interviewed customers to focus resources on other activities.

Costs. The interviewed organizations experienced the following risk-adjusted PV costs:

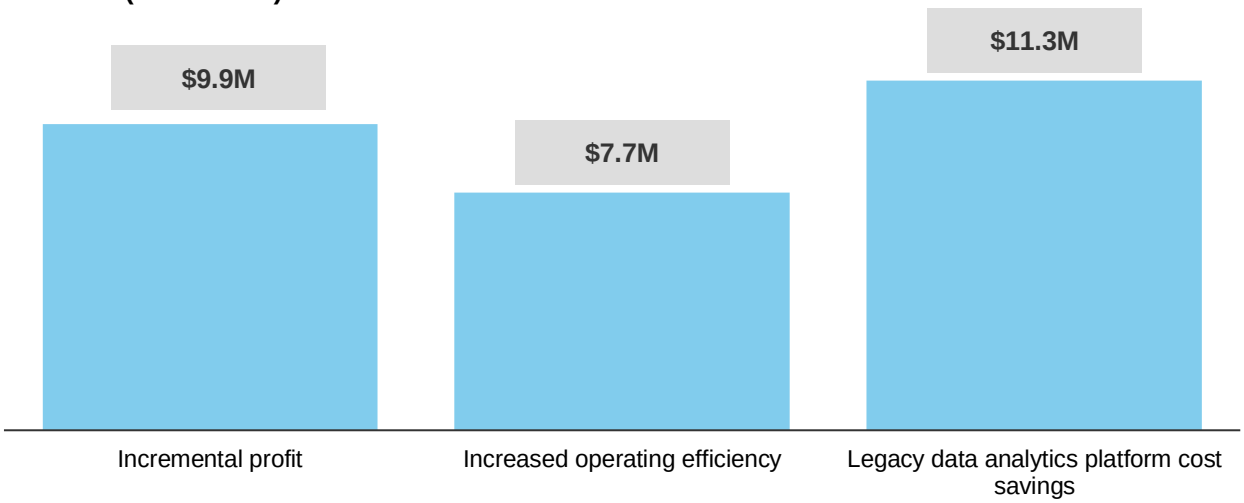
- **Databricks Unified Data Analytics platform, training, and storage costs of $4.5 million.** These are the annual usage and training costs users pay to Databricks along with third-party cloud storage costs.
- **Databricks administrative costs of $815.4 thousand.** This is the organizational salary burden incurred to manage the implementation and operation of the platform.
- **Databricks training internal labor costs of $263.6 thousand.** This is the organizational salary burden incurred to train users on the platform.

Forrester's interviews with four existing customers and subsequent financial analysis found that a composite organization based on these interviewed organizations experienced benefits of $28.9 million over three years versus costs of $5.6 million, adding up to a net present value (NPV) of $23.3 and an ROI of 417%.

Financial Summary



Benefits (Three-Year)



FORRESTER®

TEI Framework And Methodology

From the information provided in the interviews, Forrester has constructed a Total Economic Impact™ (TEI) framework for those organizations considering implementing the Databricks Unified Data Analytics Platform.

The objective of the framework is to identify the cost, benefit, flexibility, and risk factors that affect the investment decision. Forrester took a multistep approach to evaluate the impact that Databricks platform can have on an organization:

The TEI methodology helps companies demonstrate, justify, and realize the tangible value of IT initiatives to both senior management and other key business stakeholders.

DUE DILIGENCE
Interviewed Databricks stakeholders and Forrester analysts to gather data relative to the platform.

CUSTOMER INTERVIEWS
Interviewed four organizations using the platform to obtain data with respect to costs, benefits, and risks.

COMPOSITE ORGANIZATION
Designed a composite organization based on characteristics of the interviewed organizations.

FINANCIAL MODEL FRAMEWORK
Constructed a financial model representative of the interviews using the TEI methodology and risk-adjusted the financial model based on issues and concerns of the interviewed organizations.

CASE STUDY
Employed four fundamental elements of TEI in modeling Databricks' impact: benefits, costs, flexibility, and risks. Given the increasing sophistication that enterprises have regarding ROI analyses related to IT investments, Forrester's TEI methodology serves to provide a complete picture of the total economic impact of purchase decisions. Please see Appendix A for additional information on the TEI methodology.

The Databricks Unified Data Analytics Platform Customer Journey

BEFORE AND AFTER THE PLATFORM INVESTMENT

Interviewed Organizations

For this study, Forrester conducted interviews with four Databricks customers. Interviewed customers include the following:

INDUSTRY	REGION	INTERVIEWEE	ENVIRONMENT
Retail	Headquartered in EMEA	Lead data scientist	• 9 data scientists • 220 TB of data • 450 Databricks users
Pharmaceutical	Headquartered in North America	Manager, data lake and analytics	• 100 data scientists • 300 TB of data • 300 Databricks users
Heavy equipment	Headquartered in North America	Principal architect, data lake and analytics	• 10 data scientists • 120 TB of data • 120 Databricks users
Media	Headquartered in North America	Vice president, data science	• 200 data scientists • 10 TB of data • 400 Databricks users

Key Challenges

The interviewed customers knew that their organizations needed to be more data-driven to drive key business outcomes. However, their existing technology stacks proved to be more hindrances than assets. The interviewed organizations cited the following problems with their previous technology stacks:

- **Siloed data slowed the development process, led to less-accurate ML models, and wasted data teams' time.** Data scientists and analysts struggled to collect the relevant data to begin a new project. The principal architect in the heavy equipment industry said: "Data was scattered across the organization. You had to know someone who knew someone to know where the data was. We were stuck with what we had, making it more difficult to change according to our research." The limited data and relatively small amount of time data scientists could spend on ML models or deriving insights muted their overall efficacy.
- **Existing infrastructure slowed projects at best and blocked them at worst.** The limited, static, on-premises infrastructure prevented workers from starting new projects and slowed the progress of existing projects. The interviewed organizations' on-premises environments:
    - **Lacked the compute power to process full datasets.** The vice president of data science for the media company explained that "there was nowhere for us to analyze our extremely large datasets. We were stuck on-prem with our expensive [legacy] serves; we needed something that could scale to meet our needs." Data teams were forced to work with portions of data, impeding their ability to derive meaningful insights from experiments or analysis.

> "There was a lack of willingness to do anything on our previous environment because it was so slow."
>
> *Principal architect, data lake and analytics,*
>
> *heavy equipment manufacturer*

- **Couldn't scale to meet times of high demand.** The principal architect for the manufacturer said: "We didn't have any real scalability or elasticity built in to meet our needs during our busy season. During these periods, we'd be worried whether or not our [on-premises] servers would survive or be able to handle the volumes of data that are coming in."

  The manager at the pharmaceutical organization echoed these concerns: "There are sets of use cases that don't make sense in a static cluster environment. For example, we get data sets that are brought in on a weekly or monthly basis. Ideally, we'd spin up a cluster or multiple clusters to process the datasets as rapidly as possible and then get the insights to our users. But that isn't possible with a static cluster."

› **Struggled to keep up with climbing on-premises and cloud costs.** The data lake and analytics manager at the pharmaceutical company said, "Our [on-premises] cluster costs, from both an infrastructure and licensing perspective, were increasing dramatically." Other interviewed data executives noted that they had to continuously buy new on-premises servers to keep up with their growing business needs. The ever-increasing infrastructure and maintenance needs associated with this growth proved daunting for some organizations.

  Organizations that moved to the cloud continued to see their compute and storage costs increase. One executive noted that their organization's storage costs rose at a rate of 16% per year.

› **Lacked efficient collaboration tools to drive business outcomes.** Previously, the data science, engineering, and analytics teams lacked modern collaboration tools in their development stacks, so data teams would often work independently on projects. The lead data scientist for the retailer explained: "Before, somebody would just build something, hope it worked, and be done. Then, when things went wrong, nobody followed up to fix the errors." Creating effective ML models that drive business outcomes requires collaboration among data scientists, engineers, app developers, and business stakeholders. Without the right tools to collaborate, data scientists' teams cannot effectively create models and iterate over them. Moreover, without adequate documentation, data science teams risk having models that no one understands, forcing them to spend substantial amounts of time understanding the models or remaking them.

> "Our [on-premises] infrastructure and licensing costs were increasing pretty dramatically as our storage needs went up, but our compute needs didn't really change. Trying to manage that growing infrastructure was daunting."
>
> *Manager, data lake and analytics, pharmaceutical*

Solution Requirements

The interviewed organizations searched for a solution that could:

› Provide a modern technology stack to foster collaboration and drive business outcomes.

› Free up data scientist and analyst time to focus on higher-value activities.

› Enable a move to the cloud.

› Keep data clean and performant through improved governance policies.

Key Results

The interviews revealed that key results from the Databricks platform investment include:

› **Democratized data access drove various business outcomes.** Allowing people throughout the organization to leverage the Databricks platform brought in a diverse set of perspectives and skillsets. These new users increased revenue, decreased costs, and improved performance.

- **Faster time-to-market.** Adopting Databricks enabled the interviewed organizations to address bottlenecks in their previous environments. The lead data scientist for the retail firm said: “People from marketing could now build software and tools themselves — and bring them into production — without needing data scientists or developers. [That led to] a faster time-to-market with fewer resources needed.”
- **Increased revenue by identifying new opportunities.** The marketing team within the retail organization identified an opportunity to increase sales by predicting returns on sold-out items. The lead data scientist said the marketing team "built a solution out of the Databricks notebooks entirely by themselves.”

  The principal architect echoed these points and said: “We’re going to market faster because I’m able to do things that I struggled with before. They’re just working. A lot of that comes with scale and performance. I can process more data that I couldn’t process before. I can deal with bigger datasets that I couldn’t deal with before and had real challenges with. I have opened up possibilities I didn’t have before, but I also just have fewer issues. My output goes to market quicker. My ideas go quicker. I’m able to get to other ideas and other problems that I historically just wouldn’t have time to get to because I was struggling just to [finish] the base I was working on.”
- **Upskilled employees discovered new opportunities while reducing data science team burdens.** The lead data scientist explained that one marketer within the organization identified a way to optimize an operation by 25%. "They didn't have any data science or data engineering experience. They just picked it up. Then they created a blog post with a nice example of how to do this for everybody."

  The principal architect for the heavy equipment manufacturer said that adopting Databricks enabled a mindset shift across their organization. "Our customer support team is trying to figure out, 'How do we use the information we have to support our customers and machines in the field better?' To that end, they're creating models to identify early indicators of problems to alert our dealers to make proactive repairs. Our engineering team is asking, 'How do we use our information to improve the manufacturing process?' The parts department is asking, 'How can we automate this manual process?'" While it’s still too early to measure the impact of some of these initiatives, the principal architect reported that the parts pricing department has already driven millions of dollars in additional revenue while only optimizing the price for a relatively small percentage of the organization’s overall catalog.

> “Databricks lowers friction, democratizes control, and enables data science to do R&D without any blockers.”
>
> *Vice president, data science, media*

> “Beforehand, we had a select number of analysts who knew how everything worked, so everything would go through them. But now, basically everybody is learning. Everybody in the company wants to do something with Databricks.”
>
> *Lead data scientist, retail*

- **Improved collaboration and experimentation capabilities led to better ML products that drive revenue and decrease costs.** The Databricks platform enabled the interviewed customers to speed up their development cycles and refine their ML models over time. The lead data scientist for the retailer said, "Databricks allows us to shorten the time needed to complete an ML model, check that it works, then have someone else review and improve on the model while also tracking the model over time.". With these added capabilities, the interviewed organizations were able to create or optimize the following applications in their companies:
    - The pharmaceutical company created an AI engine to identify which prospective customers that salespeople should focus their selling efforts on — and when they should do this. The organization also found that it was overpaying for raw materials, enabling it to renegotiate contracts and save tens of millions of dollars.
    - The retailer increased sales by improving its search engine and creating recommendation engines for various customer personas. It was also able to reduce its operating costs by driving down returns and identifying fraudulent orders.
    - The heavy equipment manufacturer improved its ability to identify manufacturing issues, reducing downtime for its clients, and decreasing warranty repair costs.

> "We're recognizing huge benefits across the organization. People are less dependent on developers or analysts and more self-sufficient. People can build projects on their own."
>
> *Lead data scientist, retail*

- **Enabled the creation of more ML models.** Access to ML libraries, experimentation tracking features provided through native inclusion of MLflow (an open source platform to manage the ML lifecycle), and collaborative notebooks helped the interviewed organizations accelerate the creation of ML models. Furthermore, by democratizing access to the Databricks platform, more people within the organizations were able to contribute to the creation of ML models and algorithms.
- **Provided more stable and performant environments.** Processes on the Databricks platform finished in a fraction of the time they took on the interviewees' previous solutions. For example, the principal architect noted a 40% decrease in compute time and a 97% decrease in time needed for Apache Spark (an open source, general-purpose distributed computing engine used for processing and analyzing a large amount of data) processes to finish. Also, the principal architect observed a reduction in failed jobs from a 0.4% to 1.0% failure rate before Databricks to a 0.1% failure rate with Databricks. Meanwhile, the vice president of data science noted that processes went from taking a week to finish before Databricks to taking 1.5 hours with Databricks. These performance increases helped data scientists and analysts perform more experiments during sprint cycles and helped the organizations get a better sense of their models.

> "Before, we had to wait two weeks for a process to finish before we could analyze and iterate on a ML model. Now the same job takes an hour. We can iterate over it multiple times a day, allowing us to really understand the outcomes of our research, tweak it, and make a new model without interruption."
>
> *Vice president, data science, media*

- **Reduced infrastructure and simplified management.** By moving to Databricks, the interviewed organizations were able to retire their on-premises infrastructures, cancel redundant software licenses, and reallocate IT resources. Managing the platform proved substantially easier than it was with prior environments. Organizations could move to DevOps environments, further simplifying the management of their environments. The manager of data lake and analytics for the pharmaceutical company stated: "[Databricks] made my life a lot easier. [Nobody] wants to spend all their time dealing with finances and people complaining about upgrades or performance dips. [On Databricks], we adopted DevOps: everyone has their own cluster that they pay for. They manage the size and upgrade when they want."
- **Provided invaluable expertise in the biotech and genetic space.** The data lake and analytics manager for the pharmaceutical company said: "Databricks is really helpful in the healthcare space because it has people who are Spark programmers and geneticists. Instead of spending all our time trying to figure something out by yourself, we can go to Databricks for help. They understand what we're trying to do with DNA sequencing because they've invested in the biotech space."

Composite Organization

Based on the interviews, Forrester constructed a TEI framework, a composite company, and an associated ROI analysis that illustrates the areas financially affected. The composite organization is representative of the four companies that Forrester interviewed and is used to present the aggregate financial analysis in the next section. The composite organization that Forrester synthesized from the customer interviews has the following characteristics:

Description of the composite. The composite organization is a global enterprise in the B2C and B2B products and services business. It's headquartered in the United States and generates annual revenue of $5 billion.

The organization currently has a large on-premises data analytics footprint, 100 data scientists, 200 data engineers, and 450 TB of data. It has a series of ML models that support its recommendation engines and search engine.

Deployment characteristics. The composite organization moves to the Databricks platform incrementally over three years, transitioning 25% of its data, data scientists, and data engineers every year. In addition, the organization also onboards 100 new users every year — from marketers to business analysts, application developers, product managers, and others.

As part of its migration, the composite organization adopts Delta Lake (an open-source storage layer that brings ACID transactions to Apache Spark™ and big data workloads) to enforce governance standards and keep its data performant. The organization uses Databricks' capabilities to refine its existing ML models and create new ones. Examples of new ML models include a pricing optimization engine, fraud detection analysis, optimization of manufacturing processes, and predictive maintenance.

Key assumptions

$5 billion in revenue
100 data scientists
200 data engineers
525 Databricks users by Year 3

Analysis Of Benefits

QUANTIFIED BENEFIT DATA AS APPLIED TO THE COMPOSITE

Total Benefits

REF.	BENEFIT	YEAR 1	YEAR 2	YEAR 3	TOTAL	PRESENT VALUE
Atr	Incremental profit	$843,750	$3,375,000	$8,437,500	$12,656,250	$9,895,520
Btr	Increased operating efficiency	$1,594,219	$3,188,438	$4,782,656	$9,565,313	$7,677,643
Ctr	Legacy data analytics platform cost savings	$1,840,625	$4,417,500	$7,951,500	$14,209,625	$11,298,202
	Total benefits (risk-adjusted)	$4,278,594	$10,980,938	$21,171,656	$36,431,188	$28,871,365

The table above shows the total of all benefits across the areas listed below, as well as present values (PVs) discounted at 10%. Over three years, the composite organization expects risk-adjusted total benefits to be a PV of more than $28 million.

Incremental Profit

The Databricks Unified Data Analytics Platform enabled the interviewed customers to increase revenue by providing them with the tools to create better ML models and analysis faster than before. With Databricks, analytics and data science teams were able to spend less time cleaning data and creating and maintaining ETL pipelines, and spend more time building and improving ML models that could drive business outcomes. The interviewed organizations drove incremental profit through:

- **Increased sales through improved customer experience (CX).** By increasing the number and efficacy of its recommendation engines for various buying personas, the retailer was able to increase order frequency. It calculated that every percentage increase in order frequency resulted in a 2.8% increase in revenue. The retailer was also able to identify customer pain points that led to lost sales. For example, it found that over a third of cart abandonments were due to forced account creation. By eliminating account creation requirements, the retailer would recognize a 4% increase in revenue.
- **Optimized pricing.** The principal architect for the heavy equipment manufacturer said: “Our parts pricing team moved from a manual process to a more automated process using Databricks. They 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.”
- **Augmented salespeople’s capabilities.** The pharmaceutical company created an AI assistant that optimized salespeople’s efforts by prioritizing whom they should speak to and when. “There are always more available prescribers than there are salespeople. We need a way to prioritize whom we engage with… By creating a sales rep AI assistant, we’re helping our sales reps better optimize their time, resulting in higher revenue.”

› **Resulted in ML models put into production faster than before.** Though the interviewees struggled to quantify how much faster they were going to market because of Databricks, they agreed that time-to-market impacted the financial benefits they recognized from Databricks. The interviewed organizations could realize the financial benefits of their ML models faster than before. Additionally, they could make more ML models over the same course of time, leading to additional financial benefits. Instead of creating one ML model that drives increased revenue in a year, an organization could release multiple ML models over the same period.

In modeling this benefit, Forrester made the following assumptions:

› The composite organization increases revenue by improving its conversion rate, optimizing its pricing, and offerings through the creation of new ML models and optimizing its existing ML models.

› As Databricks users continue to collaborate and refine their ML models and uncover new insights, the revenue increase due to Databricks compounds.

› Because Databricks enables the composite organization to develop and release more ML models faster than it could previously, the composite organization recognizes these revenue increases earlier.

The improvement in profit will vary based on:

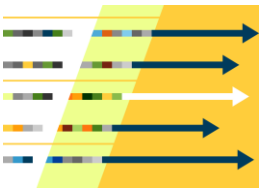
› The ability of Databricks users to create new ML models and derive insights from their data.

› The ability of Databricks users to refine their ML models.

› An organization's operating margin.

› How much an organization prioritizes revenue-increasing initiatives.

To account for these risks, Forrester adjusted this benefit downward by 10%, yielding a three-year risk-adjusted total PV of $9,895,520.



Faster creation and optimization of ML models with Databricks drives incremental profit.

Impact risk is the risk that the business or technology needs of the organization may not be met by the investment, resulting in lower overall total benefits. The greater the uncertainty, the wider the potential range of outcomes for benefit estimates.

Incremental Profit: Calculation Table					
REF.	METRIC	CALCULATION	YEAR 1	YEAR 2	YEAR 3
A1	Annual revenue	Composite	$5,000,000,000	$5,000,000,000	$5,000,000,000
A2	Percent of the organization/revenue affected by Databricks	Composite	25%	50%	75%
A3	Revenue increase due to Databricks	Assumption	1.50%	3.00%	5.00%
A4	Increased revenue due to Databricks	A1*A2*A3	$18,750,000	$75,000,000	$187,500,000
A5	Operating margin	NYU Stern	5%	5%	5%
At	Incremental profit	A4*A5	$937,500	$3,750,000	$9,375,000
	Risk adjustment	↓10%			
Atr	Incremental profit (risk-adjusted)		$843,750	$3,375,000	$8,437,500

Increased Operating Efficiency

Prior to migrating to the Databricks platform, analytics, data science and engineering teams spent the majority of their time on lower-value activities other than creating ML models and deriving insights from data. Data scientists, analysts and engineers spent the majority of their time:

- **Gathering and cleaning data.** Analytics and data science teams could spend weeks just to get ready to experiment on datasets. These data teams spent significant amounts of time looking for data across silos and cleaning data.
- **Creating test and production environments.** Because data science and data engineering teams worked in their silos, it was difficult to collaborate on a project. Data scientists and engineers used different programming languages and methodologies, and they didn't understand what the other was doing. Consequently, working to create a new test environment or production environment required a lot of back and forth.

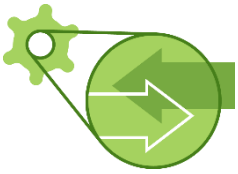
  The vice president of data science for the media company explained the difficulties his team faced: "Historically we would do all of our R&D in our environment, write up a requirements document, and then send it over to engineering for them create a development environment. There was a lot of back and forth as they didn't really understand what we were asking them to implement. It was a really slow and cumbersome development paradigm."

> "My output goes to market quicker. My ideas go quicker. I'm able to get to other ideas and other problems that I historically just have time to get to because I was struggling just to get my base work done."
>
> *Principal architect, data lake and analytics,*
>
> *heavy equipment manufacturer*

Because data scientists had to spend so much time on the tasks above, they had relatively little time to spend on value-added activities. Data scientists spent a small fraction of their time iterating on their ML models, collaborating with stakeholders, or working on new projects. For many interviewed customers, their infrastructure only worsened this problem. Slow processing times and the lack of automation meant that data scientists might only be able to run one experiment over a two-week sprint cycle.

The interviewed organizations leveraged Delta Lake to keep their data clean and performant. By enforcing data quality standards and centralizing their data, the interviewed organizations reduced the time and effort needed to start a new project. Moreover, access to more and higher-quality data improved the quality of the ML models.

After migrating to the Databricks platform and managing their data on Delta Lake, the interviewed organizations reported productivity gains across their data science teams. Analysts and data scientists were able to spend less time searching for data, cleaning it, or sifting through results, and they were able to spend more time on activities that drove business outcomes. Further, the improved collaborative and ML lifecycle management capabilities helped data scientists create and improve their ML models faster than they previously could. The additional features and capabilities of the platform enabled teams to work much more quickly. Instead of running one experiment over a two-week sprint cycle, data science teams could run multiple analyses a day, allowing them to have a better sense of the results. That resulted in better ML models.



25% increase in productivity for data scientists.

The additional languages supported on the platform, helped analysts, data scientists and engineers collaborate better. The vice president of data science for the media company said: “Databricks enabled us to be much more integrated with the engineering team than we were on our old solution. We’re more integrated with their tech stack, work on languages that they’re willing to support, and we share repositories.”

The interviewed organizations noted the following time savings due to the Databricks platform:

CATEGORY	BEFORE DATABRICKS	AFTER DATABRICKS
Creating and maintaining ETL pipelines	Days	Minutes
Configuring server clusters for new data product experiments	Weeks to months	Hours
Verifying and cleaning data	Days	Hours
Model training/hyperparameter tuning and model validation	Limited to nonexistent	Hours to a day
Putting models into production, monitoring, managing, and retraining models	Multiple two-week sprint cycles over months	Days to weeks, depending on the level of self-service enabled
Moving data product from development to production	Weeks to months	Days to weeks, depending on the level of self-service enabled

All of these factors enabled the interviewed organizations to make meaningfully better ML models faster and with fewer resources than before. For the composite organization, Forrester assumes that:

- Data scientists and data engineers spend 75% of their time searching for data, cleaning data, creating test environments, and moving models to production.
- Forrester conservatively assumes that data scientists and engineers reduce the time spent on the above activities by 25% and 20%, respectively. Data scientists achieve these primarily by reducing the time spent searching and cleaning data. Meanwhile, data engineers primarily lessen the time spent cleaning data and creating test and production environments.

The productivity increases will vary based on:

- How data scientists and data engineers currently spend their time.
- Data quality.
- The fully loaded compensation of a data scientist and data engineer.

To account for these risks, Forrester adjusted this benefit downward by 5%, yielding a three-year risk-adjusted total PV of $7.7 million.

20% increase in productivity for engineers

Increased Operating Efficiency: Calculation Table					
REF.	METRIC	CALCULATION	YEAR 1	YEAR 2	YEAR 3
B1	Data scientists affected by Databricks	Composite	25	50	75
B2	Percent of data scientist time affected by adopting Databricks	Based on customer interviews	75%	75%	75%
B3	Data scientist time savings with Databricks	Based on customer interviews	25%	25%	25%
B4	Data scientist salary	Industry average	$150,000	$150,000	$150,000
B5	Data engineers affected by Databricks	Composite	50	100	150
B6	Percent of data engineer time affected by Databricks	Based on customer interviews	75%	75%	75%
B7	Percent of data engineer time affected by adopting Databricks	Based on customer interviews	20%	20%	20%
B8	Data engineer salary	Industry average	$130,000	$130,000	$130,000
Bt	Increased operating efficiency	(B1*B2*B3*B4)+(B5*B6*B7*B8)	$1,678,125	$3,356,250	$5,034,375
	Risk adjustment	↓5%			
Btr	Increased operating efficiency (risk-adjusted)		$1,594,219	$3,188,438	$4,782,656

Legacy Data Analytics Platform Cost Savings

Before adopting Databricks, most of the interviewed organizations managed extensive on-premises data analytics environments. The interviewed organizations either had hundreds to 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, the interviewed customers had to keep buying more servers, creating more management overhead. One executive explained the difficulty of managing their previous environment: “You’re continually buying new machines or adding new instances because your needs are growing. You’re expanding your infrastructure and buying new licenses. And it’s difficult to manage. You have to deal with funding [across various departments], provisioning requests, and upgrading the equipment.” As the environment continued to grow, these challenges would only increase in difficulty.

Moving to the cloud did little to bend the cost curve. One executive noted

that their cloud costs increased by more than 16% each year to keep pace with their growing demands.

With Databricks, the interviewed customers were able to start retiring their on-premises infrastructures and begin reducing or canceling third-party licensing and services. Moreover, Databricks helped reduce administrative costs. Engineers no longer had to worry about maintaining or upgrading the platform. Several interviewees reduced administrative requirements even further by providing self-service portals and moving to DevOps. "We're adopting DevOps, we're moving to a 'you own it, you size it, and you manage your own costs' system. We use API wrappers and orchestration tools to set up new instances with the right permissions, enforce tagging, and log all the relevant activity."

Cost savings extend to the cloud. Interviewees noted that they had seen cloud costs stay steady or grow less rapidly than they had before, despite growing the size of their data lakes substantially. Databricks' autoscaling, coupled with its significantly better performance and stability, resulted in lower cloud infrastructure costs.

Also, the interviewed data executives noted significant stability and computational improvements after moving to Databricks:

- The principal architect for the heavy machinery manufacturer said they observed a 40% to 50% increase in specific processes. Furthermore, they experienced a 97% decrease in the time required for Spark processes to complete, compared to their previous cloud solution using open source Spark. The principal architect also described problems with their previous environment: "Some processes just wouldn't run. It would run for two to three hours and then crash on the server. [On Databricks,] it runs and finishes in a matter of seconds." Finally, the number of failed jobs decreased from "one out of every 100 to 250 jobs [before Databricks] to one out of every thousand [with Databricks]."
- The vice president of data science noted that processes went from taking a week on their previous solution to taking 1.5 hours on Databricks. They said: "Before, a job would have taken a week to compute and a week of manual processes. With data available programmatically in a centralized data lake and with the Databricks cluster management functionality, we're able to automate the entire process from end-to-end."

For the composite organization, Forrester assumes that:

- Its on-premises infrastructure costs $7 million annually to maintain. That includes hardware spending costs to decommissioned servers, networking, cooling, storage, internal administrative labor costs, and software costs not related to the data analytics solution.
- Infrastructure costs increase 20% year-over-year to meet growing needs.

On-premises infrastructure costs will vary with:

- The size of an organization's on-premises data analytics environment.
- Data analytics licensing fees.

To account for these risks, Forrester adjusted this benefit downward by 5%, yielding a three-year risk-adjusted total PV of $ $11,298,202.

> "Our cloud costs have stayed flat year-over-year despite exponential growth and use of cloud compute."
>
> *Vice president, data science, media*

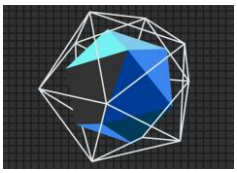
75% of legacy environment decommissioned over three years with Databricks

Legacy Data Analytics Platform Cost Savings: Calculation Table					
REF.	METRIC	CALCULATION	YEAR 1	YEAR 2	YEAR 3
C1	Annual on-premises infrastructure cost for the legacy data analytics solution	Composite	$7,000,000	$8,400,000	$10,080,000
C2	Legacy data analytics platform licensing and support costs	Composite	$750,000	$900,000	$1,080,000
C3	Percent of environment decommissioned	Composite	25%	50%	75%
C4	Hardware and licensing cost avoidance	(C1+C2)*C3	$1,937,500	$4,650,000	$8,370,000
Ct	Legacy data analytics platform cost savings	C4	$1,937,500	$4,650,000	$8,370,000
	Risk adjustment	↓5%			
Ctr	Legacy data analytics platform cost savings (risk-adjusted)		$1,840,625	$4,417,500	$7,951,500

Unquantified Benefits

The interviewed organizations experienced the following benefits that could not be financially quantified in this study. The unquantified benefits for the Databricks Unified Data Analytics Platform, evaluated by Forrester, include:

- **Reduced operating costs.** The new ML models created on the Databricks platform, coupled with the new insights derived by centralizing their data, enabled the interviewed organizations to reduce their spending on raw materials, identify manufacturing issues earlier, and reduce warranty repair costs and return rates.
    - The pharmaceutical company created an AI engine to identify whom salespeople should focus their selling efforts on and when they should do this. The organization also found that it was overpaying for raw materials, enabling it to renegotiate contracts and save tens of millions of dollars.
    - The retailer increased sales by improving its search engine and creating recommendation engines for various customer personas. It was also able to reduce its operating costs by driving down returns and identifying fraudulent orders.
    - The heavy equipment manufacturer improved its ability to identify manufacturing issues, which reduced downtime for its clients and decreased warranty repair costs.
- **Improved security.** Prior to using Databricks, the interviewed organizations lacked security standards across their environments and the internal resources needed to enforce security standards. Databricks provided native security protection.



Better security and data governance

Flexibility, as defined by TEI, represents an investment in additional capacity or capability that could be turned into business benefit for a future additional investment. This provides an organization with the "right" or the ability to engage in future initiatives but not the obligation to do so.

Flexibility

The value of flexibility is clearly unique to each customer, and the measure of its value varies from organization to organization. There are multiple scenarios in which a customer might choose to implement the Databricks platform and later realize additional uses and business opportunities, including:

- **Leverage new ML libraries.** Databricks provides popular ML libraries and frameworks out-of-the-box, enabling data scientists to easily experiment with the latest machine learning techniques.
- **Adopt new programming languages.** Databricks supports a wide range of modern programming languages (including R, Python, Scala, and SQL), empowering teams to explore data using new programming languages within the same notebook.

Flexibility would also be quantified when evaluated as part of a specific project (described in more detail in Appendix A).

Analysis Of Costs

QUANTIFIED COST DATA AS APPLIED TO THE COMPOSITE

Total Costs

REF.	COST	INITIAL	YEAR 1	YEAR 2	YEAR 3	TOTAL	PRESENT VALUE
Dtr	Databricks platform, training, and storage costs	$0	$971,250	$1,732,500	$2,913,750	$5,617,500	$4,503,916
Etr	Databricks administrative costs	$136,500	$273,000	$273,000	$273,000	$955,500	$815,411
Ftr	Databricks training costs	$4,234	$101,606	$105,840	$105,840	$317,520	$263,593
	Total costs (risk-adjusted)	$140,734	$1,345,856	$2,111,340	$3,292,590	$6,890,520	$5,582,920

Databricks Platform, Training, And Storage Costs

To use the Databricks Unified Data Analytics Platform, organizations incurred consumption costs called Databricks Units (DBUs). A DBU is a unit of processing capability per hour, billed on per-second usage. Additionally, the interviewed organizations paid for the onboarding and training of its analysts, data scientists and data engineers, and to democratize usage of the platform to marketers, business stakeholders, and other people throughout the organizations.

Furthermore, the organizations incurred storage costs from their third-party cloud providers.

For the composite organization, Forrester made the following assumptions:

- Usage costs increase every year as more data and business units move to the Databricks platform.

Forrester recognizes that these costs will vary based on:

- Usage of the Databricks platform.
- The amount of data moved to the cloud.
- The number of people trained on the platform.

To account for these risks, Forrester adjusted this cost upward by 5%, yielding a three-year risk-adjusted total PV of $4,503,916.

The table above shows the total of all costs across the areas listed below, as well as present values (PVs) discounted at 10%. Over three years, the composite organization expects risk-adjusted total costs to be a PV of more than $5.5 million.

Implementation risk is the risk that a proposed investment may deviate from the original or expected requirements, resulting in higher costs than anticipated. The greater the uncertainty, the wider the potential range of outcomes for cost estimates.

Databricks Platform, Training, And Storage Costs: Calculation Table						
REF.	METRIC	CALCULATION	INITIAL	YEAR 1	YEAR 2	YEAR 3
D1	Databricks platform and training costs			$800,000	$1,400,000	$2,400,000
D2	Cloud storage costs			$125,000	$250,000	$375,000
Dt	Databricks platform, training, and storage costs	D1+D2	$0	$925,000	$1,650,000	$2,775,000
	Risk adjustment	↑5%				
Dtr	Databricks platform, training, and storage costs (risk-adjusted)		$0	$971,250	$1,732,500	$2,913,750

Databricks Administrative Costs

Before fully adopting the Databricks Unified Data Analytics Platform, the interviewed organizations performed a proof of concept (PoC) to ensure the platform met their needs. After, the interviewed organizations dedicated data engineers to the administration of the Databricks platform; including:

- Management of Delta Lake, including the initial implementation of governance policies and the ongoing migration of data to the platform.
- Assistance in the training of new employees.

In modeling these administrative costs, Forrester assumes that two data engineers are dedicated to the ongoing management of the Databricks platform.

Administrative costs can vary based on:

- The workforce size and existing skillsets.
- The size and scope of a deployment.
- The amount of data moved to the cloud.
- Business requirements.
- Data governance requirements.

To account for these risks, Forrester adjusted this cost upward by 5%, yielding a three-year risk-adjusted total PV of $815,411.

Six months:
Total implementation and deployment time

Two FTEs
spend 100% of their time on ongoing management of the Databricks platform.

Databricks Administrative Costs: Calculation Table						
REF.	METRIC	CALCULATION	INITIAL	YEAR 1	YEAR 2	YEAR 3
E1	Data engineers dedicated to Databricks PoC and implementation	Composite	2	0.00	0.00	0.00
E2	Months required to set up Databricks		6	0.00	0.00	0.00
E3	Data engineers required to maintain Databricks			2.00	2.00	2.00
E4	Data engineer salary		$130,000	$130,000	$130,000	$130,000
Et	Databricks administrative costs	(E1*E2/12*E4) +(E3*E4)	$130,000	$260,000	$260,000	$260,000
	Risk adjustment	↑5%				
Etr	Databricks administrative costs (risk-adjusted)		$136,500	$273,000	$273,000	$273,000

Databricks Training Costs

In addition to the training fees calculated above, Forrester calculated the internal labor costs associated with training the composite organization's employees on the Databricks platform. In calculating this cost, Forrester assumes that:

- As part of the initial PoC, seven FTEs, a blend of data engineers, data scientists, and other users are trained on the Databricks platform.
- Twenty-five data scientists, 50 data engineers, and an additional 100 users are trained on the Databricks platform in Years 2 and 3.
- Users spend an average of 12 hours on instructor-led seminars, lunch and learn sessions, and self-guided courses.
- An average blended salary of $100,000 per year.

Training costs will vary based on:

- The number of FTEs trained per year.
- Existing skill sets, which will either increase or decrease training needs.
- Whether or not users other than data scientists and data engineers are trained.
- FTE salaries.

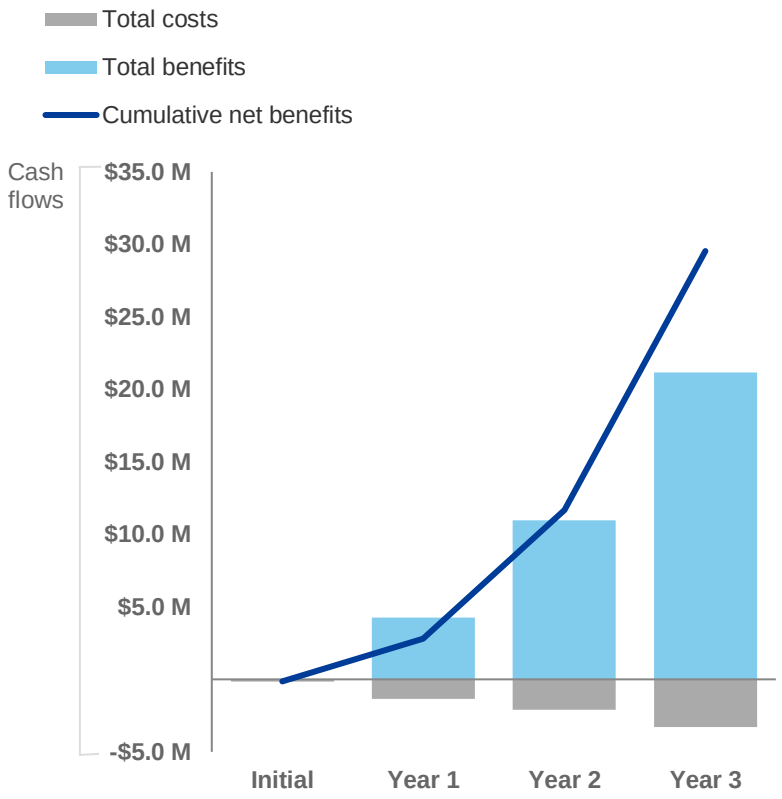
To account for these risks, Forrester adjusted this cost upward by 5%, yielding a three-year risk-adjusted total PV of $263,593.

Databricks Training Costs: Calculation Table						
REF.	METRIC	CALCULATION	INITIAL	YEAR 1	YEAR 2	YEAR 3
F1	FTEs trained on Databricks	Composite	7	168	175	175
F2	Hours spent on training	Assumption	12	12	12	12
F3	FTE blended salary	Assumption	$48	$48	$48	$48
Ft	Databricks training costs	F1*F2*F3	$4,032	$96,768	$100,800	$100,800
	Risk adjustment	↑5%				
Ftr	Databricks training costs (risk-adjusted)		$4,234	$101,606	$105,840	$105,840

Financial Summary

CONSOLIDATED THREE-YEAR RISK-ADJUSTED METRICS

Cash Flow Chart (Risk-Adjusted)



The financial results calculated in the Benefits and Costs sections can be used to determine the ROI, NPV, and payback period for the composite organization's investment. Forrester assumes a yearly discount rate of 10% for this analysis.

These risk-adjusted ROI, NPV, and payback period values are determined by applying risk-adjustment factors to the unadjusted results in each Benefit and Cost section.

Cash Flow Analysis (risk-adjusted estimates)

	INITIAL	YEAR 1	YEAR 2	YEAR 3	TOTAL	PRESENT VALUE
Total costs	($140,734)	($1,345,856)	($2,111,340)	($3,292,590)	($6,890,520)	($5,582,920)
Total benefits	$0	$4,278,594	$10,980,938	$21,171,656	$36,431,188	$28,871,365
Net benefits	($140,734)	$2,932,737	$8,869,598	$17,879,066	$29,540,668	$23,288,445
ROI						417%
Payback period (months)						<6

FORRESTER®

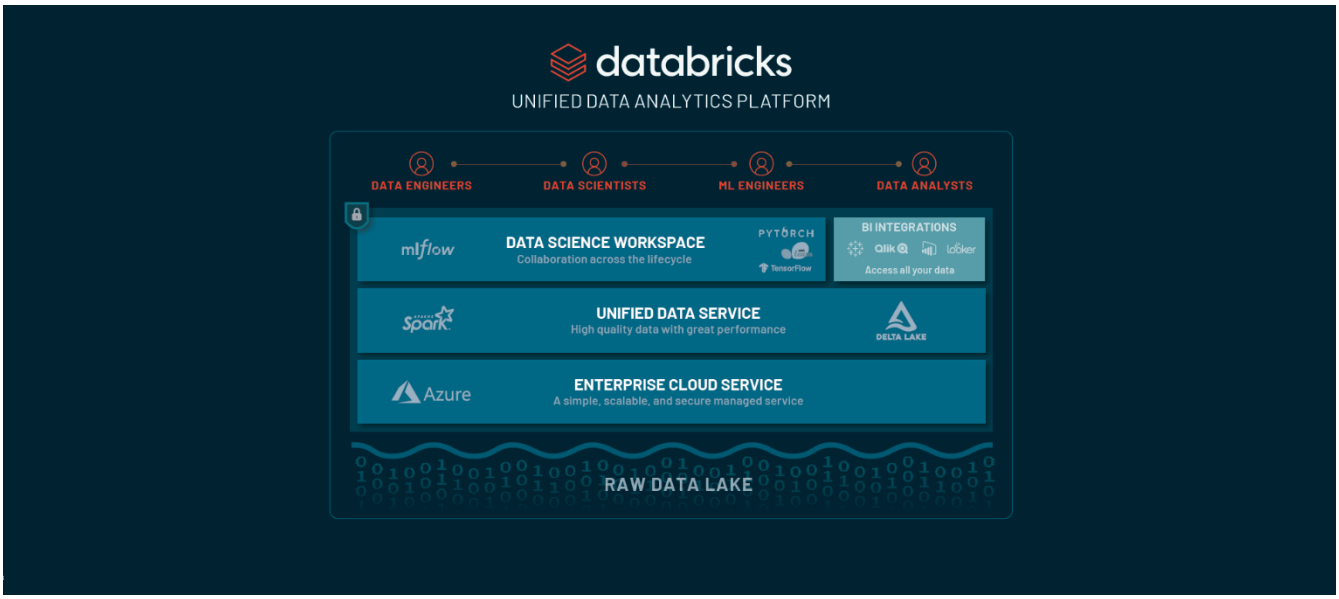
Databricks Unified Data Analytics Platform: Overview

The following information is provided by Databricks. Forrester has not validated any claims and does not endorse Databricks or its offerings.

Accelerating Data-Driven Innovation with Databricks

Databricks simplifies data and AI so data teams can solve the world's toughest problems. Open source and cloud-first, Databricks brings together data, analytics, and business on an open, unified platform where data teams can collaborate and innovate faster than ever. Thousands of organizations worldwide — including Comcast, Shell, Starbucks and Regeneron — rely on Databricks to make all their data ready for analytics, empower data-driven decisions across the organization, and rapidly deploy machine learning. Databricks is venture-backed and founded by the original creators of open source projects including Apache Spark™, Delta Lake, and MLflow.

The Databricks Unified Data Analytics Platform



Data Science Workspace

The Data Science Workspace is a collaborative environment for practitioners to run all analytic processes in one place and manage ML models across the full lifecycle.

- Collaborative Notebooks: Databricks notebooks natively support Python, R, SQL, and Scala so practitioners can work together with the languages and libraries of their choice to discover, visualize, and share insights with stakeholders.
- Machine Learning Runtime: One-click access to preconfigured ML clusters, powered by a scalable and reliable distribution of the most popular ML frameworks, with built-in AutoML and optimizations for unmatched performance at scale.
- Managed MLflow: Built on top of MLflow — an open source platform from Databricks — Managed MLflow helps manage ML models from experimentation to production with enterprise security, reliability, and scale.

Unified Data Service

The Databricks Unified Data Service provides a reliable and scalable platform for your data pipelines, data lakes, and data platforms. Manage your full data journey so you can ingest, process, store, and expose data throughout your organization.

- Delta Lake for Databricks: Delta Lake brings enhanced reliability, performance, and lifecycle management to Data. No more incomplete jobs to roll back for cleanup, suspect data added into your data lake, or difficulty deleting data for compliance changes.
- Databricks Runtime: The Databricks Runtime is a distributed data processing engine built on a highly optimized version of Apache Spark, for up to 50x performance gains. Build pipelines, schedule jobs, and train models with easy self-service and cost-saving performance.
- BI Reporting on Delta Lake: BI Reporting on Delta Lake delivers business analytics on your data lake. Connect directly to your most complete and recent data in your data lake with Delta Lake and SparkSQL, and use your preferred BI visualization and reporting tools for more timely business insights.

Enterprise Cloud Service

Enterprise Cloud Service provides native security, simple organizationwide administration, and automation at scale for the Unified Data Analytics Platform across multiple clouds.

- Enterprise security: With native identity federation, encryption, and access controls, you get the tools you need to secure your data. You can create a safe analytics environment for your users using built-in network, data, and job isolation with various compliance options.
- Simple administration: Audit trails, logs, and billing and usage reports give you full operational visibility. Create individual workspaces for each team, set policies on usage limits, and analyze activities to ensure adherence.
- Production-ready: An API-first approach ensures seamless CI/CD and automation on auto-scaling infrastructure that scales globally. Maintain SLAs with easy application and infrastructure monitoring.

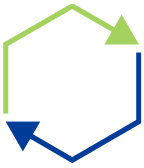
Appendix A: Total Economic Impact

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.

Total Economic Impact Approach

Benefits represent the value delivered to the business by the product. The TEI methodology places equal weight on the measure of benefits and the measure of costs, allowing for a full examination of the effect of the technology on the entire organization.

Costs consider all expenses necessary to deliver the proposed value, or benefits, of the product. The cost category within TEI captures incremental costs over the existing environment for ongoing costs associated with the solution.



Flexibility represents the strategic value that can be obtained for some future additional investment building on top of the initial investment already made. Having the ability to capture that benefit has a PV that can be estimated.

Risks measure the uncertainty of benefit and cost estimates given: 1) the likelihood that estimates will meet original projections and 2) the likelihood that estimates will be tracked over time. TEI risk factors are based on "triangular distribution."

The initial investment column contains costs incurred at "time 0" or at the beginning of Year 1 that are not discounted. All other cash flows are discounted using the discount rate at the end of the year. PV calculations are calculated for each total cost and benefit estimate. NPV calculations in the summary tables are the sum of the initial investment and the discounted cash flows in each year. Sums and present value calculations of the Total Benefits, Total Costs, and Cash Flow tables may not exactly add up, as some rounding may occur.

Present value (PV)

The present or current value of (discounted) cost and benefit estimates given at an interest rate (the discount rate). The PV of costs and benefits feed into the total NPV of cash flows.

Net present value (NPV)

The present or current value of (discounted) future net cash flows given an interest rate (the discount rate). A positive project NPV normally indicates that the investment should be made, unless other projects have higher NPVs.

Return on investment (ROI)

A project's expected return in percentage terms. ROI is calculated by dividing net benefits (benefits less costs) by costs.

Discount rate

The interest rate used in cash flow analysis to take into account the time value of money. Organizations typically use discount rates between 8% and 16%.

Payback period

The breakeven point for an investment. This is the point in time at which net benefits (benefits minus costs) equal initial investment or cost.

FORRESTER®

Appendix B: Endnotes

[i] Source: “Build An Insights-Driven Business” Forrester Research, Inc., December 10, 2019

[ii] Source: “The Future Of Machine Learning Is Unstoppable,” Forrester Research, Inc., April 25, 2019