GUIDEBOOK
PROGRAM: INFRASTRUCTURE
DOCUMENT NUMBER: X165

NUCLEUS
RESEARCH

ROI GUIDEBOOK
DATABRICKS LAKEHOUSE

ANALYST

Alexander H. Wurm, Samuel Hamway

Nucleus Research, Inc.
5555 Biscayne Boulevard, Miami, FL, 33137
+1 (617) 720-2000
NucleusResearch.com

EXECUTIVE SUMMARY

482%
Average ROI

4.1 Months
Average payback period

$30.5M
Average annual benefit

Organizations, especially leading enterprises, have reached an inflection point in their data and artificial intelligence (AI) strategies. Enterprise data storage has scaled from terabytes to petabytes, and data processing continues to climb at an even more rapid pace.

Databricks is well suited to support data storage and processing at any scale, up to and beyond multiple petabytes of data. Its flagship offering, The Databricks Lakehouse Platform, is built upon a differentiated lakehouse architecture that integrates the best qualities of data lakes and data warehouses. The architecture facilitates cost-effective data storage on low-cost object storage while leveraging high-performance, elastically scalable query engines to ensure timely and cost-efficient data processing.

The Databricks Lakehouse Platform also integrates extensive capabilities to support the development of artificial intelligence (AI) and large language models (LLMs), streaming data ingestion and transformation, and data governance across structured, semi-structured, and unstructured data.

- MLflow, embedded within Databricks, offers detailed tools for AI and LLM model tracking, versioning, and deployment while also enabling consistent lifecycle management through its experiment logging, model registry, and deployment capabilities.
- The platform's ability to handle batch and streaming data pipelines is bolstered by Delta Live Tables (DLT), which simplifies ETL development and management, maintains data consistency, and ensures fault tolerance.
- For comprehensive data governance, Unity Catalog provides an expansive view of data assets, enhancing data lineage and discovery through its metadata management, integrated search, and access control mechanisms that function uniformly across varying data types and sources.
- For data warehousing and business intelligence (BI) use cases, the vendor also offers Databricks SQL (DB SQL), a serverless data warehouse on the Databricks Lakehouse Platform that lets organizations run all of their SQL and BI applications at scale on the entirety of their data without complicated ETL.

To better understand the benefits and costs associated with an investment in the Databricks Lakehouse Platform, Nucleus conducted an in-depth return on investment (ROI) assessment of five customers representing five distinct industries. These customers realized an average ROI of 482 percent over a three-year period, with an average annual benefit of $30.5M and a payback period of 4.1 months. On aggregate, customers also referenced 49 percent time savings across their data teams, 52 percent shortened time to production for data and AI projects, $2.6M in annual infrastructure savings, $1.1M in annual administrative cost savings, and improvements to data ingestion, ETL, business intelligence (BI), and machine learning operations (MLOps) processes.

$2.6M
Average annual infrastructure cost savings

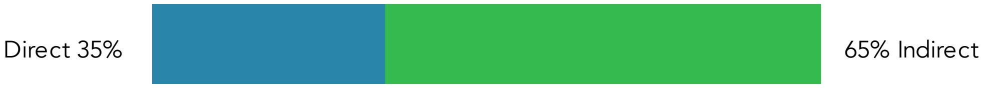
49%
Improved data team productivity

52%
Shortened time to production for big data projects

KEY FINDINGS

Certain benefits directly translate to financial value, while others need one or two additional steps to impact an organization's financial statements. These quantifiable benefits can be categorized as direct and indirect. The following direct and indirect benefits represent those most commonly experienced by the companies analyzed in this report and comprise the largest share of returns.

TYPES OF BENEFITS



DIRECT BENEFITS

Direct benefits include cost savings, cost avoidance, and other changes that have a direct impact on a budget or profit and loss (P&L) statement:

- **Eliminated or avoided technology costs.** With Databricks' highly integrated lakehouse platform, customers can retire various physical and cloud-based systems related to data processing and data management. This generates direct cost savings from avoided license and subscription spending.

> *" After grappling with countless siloed systems across our global brands, we unified our analytics tech stack with Databricks and slashed costs related to redundant software. "*

- **Eliminated maintenance and administrative costs.** Databricks features a variety of automated capabilities to streamline platform administration, yielding direct cost savings in the form of avoided DBA hires. Customers who previously relied on on-premise solutions also see hardware maintenance cost savings by moving to Databricks' cloud-native, fully-managed platform.

> *" We reduced our platform maintenance costs by 50 percent and more than doubled our administrative team's productivity while requiring fewer specialized technical skill sets. "*

- **Processing cost savings.** The Databricks Lakehouse Platform also features multiple capabilities to drive efficient data processing. This can translate into immediate cost savings as customers move workloads from less performant systems to Databricks. Additionally, the lakehouse architecture allows data to be re-used without needing to be moved to different systems, which can dramatically reduce infrastructure costs. These savings continue to ramp up as customers scale their data processing and are also reflected in their cloud investment.

> *" Our move from Snowflake to Databricks resulted in 4x cost-effectiveness on the same budget, allowing us to consume and share more data while incorporating more complex sources. "*

- **Storage cost savings.** The Databricks Lakehouse Platform features various storage options, including block storage and low-cost object storage. This gives customers the flexibility to use high-performance storage for low-latency, mission-critical workloads while most of their data sits in lower-cost object storage, translating into direct cost savings on the customer's cloud investment.

> *" We treat our data lakehouse like a data warehouse, but we pay a tenth of the cost for storage. "*

INDIRECT BENEFITS

Indirect benefits include time savings from accelerated processes that can be quantified but have an indirect impact on a budget or P&L:

- **Improved user productivity.** Databricks elevated users' productivity across various roles, including data scientists, engineers, analysts, and consumers, by accelerating general data access while providing automated capabilities to streamline manual work for data practitioners. This translates into time savings for these affected users and indirectly impacts an organization's budget through avoided hires.

> *" We are able to do a lot more with less, which has been the big motif of this deployment. Realistically, we wouldn't have allocated the resources necessary to get this done in Snowflake because of the time sink. "*

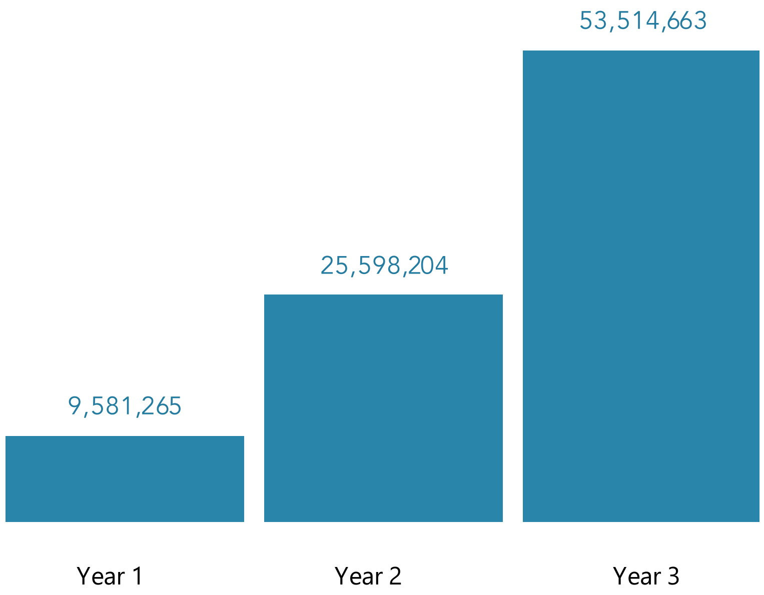
- **Better control over processing timelines.** By leveraging Databricks' elasticity and department-level controls, customers gained improved control over their costs and latency and could better optimize processing for the specific requirements of any given workload. This flexibility ensures that individual teams can better manage their consumption with a more relevant understanding of necessary timelines for workloads they own. This means that customers are better able to project usage and avoid additional burst charges.

> *" Databricks has fundamentally changed how we view data and analytics. Each individual team now controls, owns, and runs its own data projects with a consistent knowledge base, best practices, economies of scale, and reuse of data while avoiding the disadvantages and bottlenecks of a central IT ownership model. "*

- **Reduced time to production for data and AI projects.** AI models have become a catalyst for financial success, with organizations finding new ways to harness proprietary data every day to capture business opportunities and improve profitability. Databricks provides tools to accelerate time to production, whether it's a model for classification, real-time anomaly detection, time-series forecasting, or, more recently, generative applications.

> *"Our new SMS marketing model was up and running end to end in a couple of weeks. Otherwise, we would have had to build everything out ourselves on a timeline of two to three months."*

CUMULATIVE NET BENEFITS



SUMMARIZED CHALLENGES

Customers interviewed by Nucleus encountered multiple challenges that spurred Databricks Lakehouse adoption. Some of these challenges were consistent, while others depended on the scale of the organization's data operations and the maturity of its data infrastructure.

- **Tech-stack consolidation.** Organizations, especially those with multiple internal departments and acquisition-led growth strategies, often rely on multiple analytics tech stacks. Organizations contending with this challenge seek to unify systems to simplify various internal operations, from system administration and maintenance to collaboration across teams. Before adopting Databricks, a media conglomerate contended with significant system sprawl as each of its brands maintained its own tech stack. This disjointed approach limited the organizations' ability to drive user engagement, create personalized content recommendations, and restricted AI development. A sports franchise noted that transformations were challenging in Snowflake, which forced the organization to use a separate tool for its ETL pipelines. This approach was brittle and inefficient, limiting the types of data sources that could be used while eating up compute. The organization turned to Databricks as a replacement for both Snowflake and its ETL provider.

> *"Managing diverse brands operating globally on different tech stacks, our company needed to consolidate all of this data into a centralized location to break down data silos and establish a single source of truth."*

- **Support for advanced analytics.** Organizations interviewed in this guidebook noted challenges supporting their real-time streaming and data science initiatives with their previous data platforms. This challenge limited the scope of enterprise data science initiatives and made some projects impossible. A sports franchise that previously relied on Snowflake saw this challenge first-hand. It noted that Snowflake made it challenging to connect more complex data sources, including external systems and streaming sources. The organization also cited governance and permissioning issues in Snowflake, which drew out data science initiatives. Transformations to support these initiatives were also brittle and expensive in Snowflake, accounting for 40 percent of the platform's total usage. With a limited budget for R&D, these costs restricted the quantity and scope of analytics projects.

> *" Our data scientists were building a lot of AI models locally, and trying to get data in from Snowflake for training was unwieldy. Now, with Databricks, our analysts are doing stuff on the cloud, it's backed up, it's in GitHub, we're using MLflow heavily, and we can allocate permissions and governance at a scale we never had before. "*

- **Controlling costs at scale.** One of the most pervasive challenges involved limitations related to the cost-efficiency of prior data infrastructures. Organizations would often experience excessive latency and cost with significant time and resources dedicated to tuning analytic systems. A biotechnology company previously used on-premises Cloudera clusters for data processing alongside its cloud data platform. These static clusters made it hard to manage costs and resources across multiple departments. The biotechnology company also noted that the clusters were costly to scale as the number of users increased, and there was never a good time to patch or update a cluster version. Similarly, a sports franchise noted that processing was more cost-efficient in Databricks, referencing 4x cost efficiency relative to Snowflake. This benefit was especially pronounced for data ingestion use cases, which were 10x more cost-efficient. The organization noted that savings from this alone provided the necessary budget for its novel AI use cases. Another organization in the e-commerce industry relied on a combination of cloud data warehouses across multiple business units. However, it faced cost and data quality concerns as it prototyped new technologies and deployed new services into production. The organization also noted that data engineers had a hard time partitioning and optimizing query tables, which created a bottleneck for most processing.

> *" Whenever data scales by an order of magnitude, it's time to reevaluate infrastructure design. "*

- **Data accessibility and governance.** Organizations increasingly contend with poorly optimized compute engines and analytics infrastructure, especially as they scale. This can extend the timeline for accessing relevant data, impairing the performance of various roles, from data scientists and data engineers to business analysts and data consumers. A biotechnology company experienced latency on weekend batch processing workloads that made certain reports unavailable until multiple days into the following week. This latency impaired various users within the organization, from business analysts to sales professionals. Similarly, an organization in the e-commerce industry noted restricted data availability as there was previously no process to automatically pick up experiment data when marketing and communications campaigns were configured and run. Instead, the organization relied solely on data engineers to write and release Spark jobs to retrieve data. Organizations also experienced challenges related to data governance, which reduced data trust and impaired data quality. This challenge was especially pressing for a sports franchise that required many layers of complex transformations for its data products. Without a good understanding of data lineage, data consumers had to spend additional time validating data, and analysts would sink large amounts of time to identify the root cause of poor data quality.

> *"Inefficiencies in data processing created additional latency that constrained the information available for our users."*

ANALYSIS OF BENEFITS

Nucleus found that companies deploying the Databricks Lakehouse Platform experienced a range of benefits across several areas, which were largely dependent on the size and complexity of their data environment, their level of analytic maturity, the number of Databricks products deployed, and the rate of technology adoption. The best technology business cases focus on a select number of key benefits that can guide deployment and adoption efforts. To guide organizations in building their business cases, Nucleus has presented the benefits most commonly experienced by Databricks customers with guidance ranges based on what customers typically experience.

52%
Data scientist time savings

51%
Data engineer time savings

45%
Data analyst time savings

23%
Data consumer time savings

USER PRODUCTIVITY IMPROVEMENTS

Users within organizations who adopted the Databricks Lakehouse platform noted productivity improvements across various roles.

Data Scientists. Data scientists achieved time savings with Databricks by leveraging its integrated environment for model development to facilitate the rapid creation, testing, and deployment of AI models. An equipment manufacturer saw 66 percent time savings for its data scientists as they no longer had to go between different systems to pull the data they needed. Instead, data scientists could access data and start working with it in hours rather than days or weeks. Another organization in the biotechnology industry cited 21,840 work hours saved annually from improved data scientist productivity in Databricks. This was achieved through a mix of automated features and processing time savings. Lastly, a sports franchise reported 50 percent improved data science productivity by leveraging MLflow and Databricks AutoML to automate various processes. The organization also noted further improvements over time by creating and using templates to drive 2.5x faster model development.

Data Engineers. Data engineers achieved time savings with Databricks by utilizing its unified environment to build transformation pipelines, leveraging Apache Spark. The support for batch and streaming processing within Spark allows data engineers to create more flexible and responsive data transformation workflows. One organization in the e-commerce industry saw 60 percent improved productivity across its data engineering teams, noting that nearly all of its time was previously spent building and maintaining ETL pipelines.

Data Analysts & Consumers. Data analysts and consumers leverage Databricks to achieve significant reductions in processing latency, enabling swift data access. A sports franchise noted substantial productivity improvements for its business analyst teams, citing 75 percent time savings due to the ease of bringing in new data sources and reduced time spent on engineering workloads. Similarly, a biotechnology company achieved data analyst and data consumer time savings equivalent to 24,960 work hours annually with improved data access and reduced latency. This organization also noted a 10 percent improved sales rep productivity from not having to wait on weekend batch processing, yielding time savings equivalent to 83,200 work hours per year.

48%
data ingestion improvement

33%
ETL/ELT efficiency gains

28%
Improved BI efficiency

60%
More efficient MLOps

PROCESS IMPROVEMENTS

The Databricks Lakehouse Platform drove efficiency across multiple processes, including data ingestion workloads, ETL pipelines, business intelligence (BI), and MLOps.

Data Ingestion. An organization in the biotechnology industry experienced 30 percent more efficient data ingestion and avoided purchasing an additional tool by aligning with Lakehouse components. Similarly, an equipment manufacturer was able to create streaming ingestion pipelines that were previously impossible. Since adoption, the organization has seen 2x to 3x growth in its data ingestion, receiving massive amounts of data from its IoT ecosystem every 200ms, which are then mapped to hundreds of features. Likewise, an e-commerce business cited 25 percent more efficient data ingestion with Databricks. Previously, it couldn't do a full pipeline due to data volume constraints with Amazon S3's connector to its cloud data warehouse. Now, the organization can perform batch ingests in hours rather than days, referencing Databricks' connector to S3 as a driver. Another organization in the media industry saw similar benefits, noting that setting up streaming pipelines to new data sources in Databricks was instantaneous. Lastly, a sports franchise moved from twice-a-day batch ingestion to a continuous ingestion model with Databricks and noted 90 percent improved processing efficiency. This latency reduction was crucial to drive synergy within the organization as players, coaches, and business teams could access their reports immediately after a game concluded rather than waiting on batch jobs.

Data Transformation. An organization in the e-commerce industry experienced 30 percent more efficient ETL using Spark to make everything process faster, including some heavily calculated silver and gold tables. The organization also noted that it uses Databricks to build its ETL pipelines into Snowflake for other departments. Similarly, a biotechnology company saw 20 percent faster transformations in their ETL pipelines. Another organization in the media industry cited 50 percent time savings when building new transformation pipelines with Databricks. Finally, a sports franchise integrated PySpark and Spark SQL into its ETL pipeline to bring in data sources that were previously inaccessible.

Business Intelligence (BI). Customers who adopted Databricks noted improved efficiency for BI workloads with better control of costs and delivery timelines. A biotechnology company reported 35 percent more efficient BI workloads with dynamic resource allocation in Databricks. This prior latency extended batch weekend processing multiple days into the following week and limited the productivity of the organization's business analyst and sales teams. Similarly, an e-commerce platform provider saw 20 percent more efficient business intelligence as business analysts could access data faster and preferred working with the platform due to its intuitive look and feel.

MLOps. Customers reported substantial improvements to their MLOps with a better understanding of their data, transformations, features, and model development. One biotechnology company noted 40x faster time to experimentation, driving 2x faster time to live for data and AI projects. Likewise, an equipment manufacturer reduced AI model deployment time by 80 percent and realized a 2x faster delivery of data products. Another organization in the e-commerce industry experienced 60 percent more efficient MLOps, specifically citing the ease of creating a POC for its new NLP interface. Finally, a sports franchise went from no operations or best practices around their AI initiatives to established

MLOps in Databricks built around MLFlow and Unity Catalog. In streamlining these operations, the organization noted that it moved models to production 2 to 3x as fast, specifically referencing model templates and cataloging as key drivers. Databricks has also enabled the franchise to shift from local to centralized model development, which has been crucial as operations have scaled. Customers interviewed by Nucleus have expressed their intentions of leveraging the Databricks Lakehouse Platform to develop and optimize large language models (LLMs) and generative AI use cases, highlighting the platform's scalability and integrated AI capabilities as driving factors.

INFRASTRUCTURE SAVINGS

Every customer who adopted the Databricks Lakehouse Platform noted infrastructure savings in the form of storage and processing cost savings. On the storage side, this was primarily the result of moving to low-cost object storage and eliminating redundant data. Processing savings stemmed from a combination of factors, including more efficient performance tuning, better partitioning, and improved resource elasticity. A biotechnology company noted $300,000 in annual savings by migrating to the Databricks Lakehouse Platform. This organization also experienced 25 percent compute cost savings by changing from static clusters to dynamic clusters. Likewise, an equipment manufacturer cited 30 percent processing cost savings relative to its previous cloud data platform deployment, referencing Databricks' efficiency in running spark workloads. A sports franchise experienced 4x compute efficiency relative to Snowflake. This benefit was especially pronounced for data ingestion use cases, which were 10x more efficient. The organization noted that savings from this alone provided the necessary budget for all of its novel machine learning use cases. Data storage was also more efficient with Databricks' medallion architecture, generating $96,000 in annual savings.

$2.6M

Average annual infrastructure cost savings

REDUCED PROCESSING LATENCY

Utilizing the Databricks compute engine with dynamic allocation of compute resources, customers gained better control over their processing timelines. A biotechnology company saw 75 percent reduced processing latency with individual ownership of projects and timelines. This enabled respective teams to better balance

48%

Average processing latency improvement

the cost and latency of workloads rather than relying on the IT department to control budgets and timelines. This organization also saw a 36-hour reduction in processing latency for weekend loads, allowing analysts and sales professionals to work on relevant data faster.

ADMINISTRATIVE COST SAVINGS

$1.1M

Average administrative cost savings

With Databricks' elastic and serverless deployment options, customers don't need to worry about tuning and scaling, runtime, resource allocation, and upgrades. Organizations were able to allocate compute resources as needed and scale elastically to meet latency requirements. A biotechnology company noted $710,000 in annual administrative cost savings, including 50 percent less time spent on platform management relative to Hadoop. Similarly, an equipment manufacturer saw $3.4M in annual savings with reduced maintenance and administrative expenditure across its analytics enablement ops and enterprise data lake teams. Another organization in the e-commerce industry saved $480,000 in annual administrative costs and noted 20 percent administrative time savings with reduced time and effort spent troubleshooting.

ACCELERATED TIME TO PRODUCTION FOR DATA AND AI PROJECTS

52%

shortened time to production

By using the Databricks Lakehouse Platform for large-scale data processing and model training and deployment, Nucleus found that organizations significantly shortened the time to production for their data and AI projects. A biotechnology company saw 2x faster time to production for its data and AI projects. This organization also noted 40 percent less wait time for AI and data science jobs, which allowed the organization to extend the scope of these workloads and act on insights faster. Another organization in the e-commerce industry noted 1.8x faster time-to-live for AI use cases, including 60 percent reduced time-to-live for the organization's NLP interface. The organization also got its SMS marketing model up and running in a couple of weeks rather than the 3+ months it would've taken with its prior platform. Lastly, a sports franchise achieved two to three times faster time to production for its AI models. On average, development dropped from roughly ten days hosted on a data scientist's own computer to three to four days with integrated MLOps.

ANALYSIS OF COSTS

Nucleus analyzed the initial and ongoing costs of software, hardware, personnel, and consulting over a three-year period to quantify the return on investments that the Databricks Lakehouse Platform delivered to customers interviewed in this report.

COST CATEGORY	COST RANGE	AVG. COST	COST FACTORS
Annual License Costs	$283,500–$2.9M	$1.98M	Number of Databricks units for consumption-based pricing
External Consulting and Support	$0–$875,000	$292,267	Scale and complexity of the modernization initiative, external consulting spend
Initial Personnel	$320,000–$1.02M	$488,167	Internal team skill levels, scale, and complexity of the implementation
Ongoing Personnel	$200,000–$1.05M	$744,000	Internal team skill levels, scale, and complexity of data storage and processing

FINANCIAL SUMMARY

Nucleus found that the average return on investment (ROI) from a Databricks deployment was 482 percent, with a high of 1023 percent and a low of 276 percent. ROI was calculated over a three-year time horizon, projecting costs and benefits forward on a straight-line basis for organizations that had not yet reached three years of deployment.

KEY FINANCIAL METRICS:

- The payback period for a Databricks Lakehouse deployment ranged from 2.4 months to 6.0 months, with an average of 4.1 months.
- The average annual benefit of a Databricks Lakehouse deployment ranged from $2.4M to $114.6M, with an average of $30.5M.
- The net present value (NPV) of a Databricks Lakehouse deployment ranged from $2.5M to $134.1M, with an average of $33.3M.

FINANCIAL METRICS	HIGH	LOW	AVG.
ROI	1023%	276%	482%
Payback (months)	6.0	2.4	4.1
Annual Benefit	$114,602,000	$2,398,667	$30,487,642
Benefit to Cost Ratio	6.4	1.9	3.6
Present Value	$134,068,800	$2,493,442	$33,331,220

CUSTOMER PROFILES

AUDITED ORGANIZATIONS

For the development of this ROI Guidebook, Nucleus spoke to multiple Databricks customers and conducted in-depth ROI assessments of five deployments.

INDUSTRY	EMPLOYEES	INTERVIEWEES
Manufacturing	82,000+	Technical Product Manager, Enterprise Data Platforms
Biotechnology	22,000+	Chief Architect for Data and Analytics Platforms
E-Commerce	16,000+	Senior Engineer, Special Projects
Media	7,000+	Data Engineering Manager
Sports	230+	Assistant Director, R&D

EQUIPMENT MANUFACTURER

This American equipment manufacturing company specializes in creating industry-leading machinery products. The organization began its transition to the cloud in 2014, moving from mainframe databases to a hybrid data analytics environment that consisted of IBM Netezza, Db2, and cloud data warehouses, as well as Hadoop clusters on-premises. In late 2017, the organization had increased its data storage and utilization by an order of magnitude and decided to reevaluate its infrastructure design to ensure cost-efficiency at scale. The organization explored multiple solutions such as open-source Spark, EMR, Qubole, Domino, and C3 IoT, using industry vetting and performance benchmarking to evaluate each solution on a variety of factors, including version control, scalability, openness, cost-effectiveness, performance, and ease of use. After running a pilot with multiple data engineers and data scientists, the organization found that Databricks was best aligned with their organizational needs and fit well with user skill sets. The organization also specifically noted developer familiarity with Spark SQL and Delta Lake's advanced product vision as key deciding factors.

The equipment manufacturer's implementation of Databricks took approximately six months, from initial engagement to production, and was performed by the organization's enterprise data lake team, analytics enablement team, data governance team, and cloud security team. Following deployment, the manufacturer noted $5M in annual savings by sunsetting its IBM Netezza, Db2, and on-premises Hadoop clusters. Additionally, the organization saw a 30 percent improvement in the cost-efficiency of its data processing and noted $3.4M in annual savings with reduced maintenance and administrative expenditure across its analytics enablement ops and enterprise data lake teams. The organization also experienced significant improvements in its data operations, allowing users, now several thousand in number, to access and work with data much faster. Data scientists and data engineers noted 66 percent time savings, while data analyst productivity improved by 50 percent. Databricks also enabled various high-value use cases, including enhanced inventory optimization and customer lead generation, yielding $112.5M and $57.5M in additional revenue to date, respectively. The manufacturer also used the Databricks Lakehouse Platform to develop a precision analytics warehouse by integrating machine and environmental data to enable unified data access for end-users and better train AI models at scale.

BIOTECHNOLOGY COMPANY

This American biotechnology company offers various pharmaceutical products and stores over 20 PBs of data to support its drug development, manufacturing, and commercial delivery. Before adopting Databricks, the organization used on-premises Cloudera clusters for data processing alongside its cloud data platform. These were static clusters, making it hard to manage costs and resources across multiple departments and often increased latency for weekly reports. The biotechnology company also noted that the clusters were costly to scale as the number of users increased, and there was never a good time to patch or update a cluster version.

To address these challenges, the organization evaluated various solutions, including cloud data warehouses, Qubole, and Databricks. The organization decided to adopt the Databricks Lakehouse Platform, citing Delta Lake as a key differentiator. In 2019, the organization began moving its data to Databricks with AWS Glue, as performed by a team of internal and external personnel at a 70 percent time commitment. The biotechnology company noted that SparkSQL queries were easily moved using a code conversion tool; however, changing Hive and Impala queries to SparkSQL required additional work. Within eight months, the organization had moved everything into the Databricks Lakehouse Platform.

Following its Databricks deployment, the organization sunsetted multiple solutions, including Cloudera, Snaplogic, and Hadoop systems, and reduced its prior cloud data platform's usage, generating $2.4M in annual cost savings. The biotechnology company also lowered its administrative costs, yielding $710,000 in annual savings. Deploying Databricks also drove 25 percent compute cost savings with up to 75 percent reduced latency with dynamic cluster allocation and individual ownership of workloads by department. The organization also lowered its annual data storage costs by $300k with object storage and reduced data redundancy. Now, the biotechnology company notes 2x faster time-to-production for data and AI projects as business units could assign compute resources as necessary to manage execution timelines. This includes high-value use cases such as clinical trial optimization, where Databricks is used to speed up clinical trials and accelerate approvals.

E-COMMERCE PLATFORM

This American e-commerce platform provider delivers an online experience to streamline the purchasing, selling, and financing of used cars. As a digital-first business, the organization collects, stores, and maintains over 215 TBs of data to better engage customers and elevate their experience with specialized communications and AI services. Before adopting Databricks, the organization relied on a combination of cloud data warehouses across different business units but faced cost and data quality concerns as they prototyped new technologies and deployed new features and services into production. The organization noted that data engineers had a hard time partitioning and optimizing query tables, creating a bottleneck for most processing. Data availability was also restricted within the organization as there was no process to automatically pick up experiment data as campaigns were configured and run, and the organization relied solely on data engineers to write and release Spark jobs to retrieve data.

To address these challenges, the organization decided to adopt the Databricks Lakehouse Platform in March 2022. The organization spent the following four months implementing the platform, which also included Delta Live Tables and Databricks SQL Serverless. This was performed by a combination of internal engineering teams, DevOps teams, and external consulting. After deployment, the organization reduced its previous cloud data warehouse spend by 90 percent, generating $762,000 in annual savings as well as $160,000 in annual administrative savings. The organization also gained more efficient processing as it scaled, saving $1.1M in the first year and over $7.7M in projected savings for the three-year period following deployment. The organization's internal personnel also referenced significant productivity improvements, with data scientists and data engineers citing 60 percent time savings, data analysts noting 30 percent savings, and data consumers referencing five percent time savings. The online e-commerce platform provider also enabled multiple high-value business cases with Databricks. The first involved improved price optimization for its end-users, generating $2.5M in additional profit over the first year following deployment. The organization also used Databricks to train its own natural language processing (NLP) interface to improve customer support and engagement, yielding $12M in additional revenue annually. Another significant business case involved using Databricks to build an audience segmentation and personalization model, generating $3.7M in additional revenue one year following the model's launch.

MEDIA CONGLOMERATE

This US-based media conglomerate operates over 30 global print and digital media brands and receives over 100 million visits, generating 800 million page views and over one trillion data points per month. Before adopting Databricks, the organization struggled with multiple siloed systems, with each brand maintaining its own tech stack. This disconnected approach limited the organizations' ability to drive user engagement, create personalized content recommendations, and drive AI development. Faced with the challenge of consolidating its data operations, the organization considered various solutions, including Databricks and Snowflake, to create a unified data hub. It selected Databricks due to its cost-effectiveness, platform flexibility, and support for AI relative to Snowflake, as well as the decision-makers' previous experience with the environment.

The media conglomerate ran an initial pilot of Databricks in late 2021 and spent half a year moving over relevant data and workloads before going fully live in May 2022. The organization cited a smooth implementation with minimal roadblocks related to training less technical analysts and updating a few incompatible queries. Post-deployment, Databricks reduced the conglomerate's data infrastructure costs significantly, saving the organization $1.1M annually on storage and $60,000 in monthly usage by reducing EC2 instances. The organization was also able to eliminate its Presto expenditure, saving $500,000 annually. Furthermore, the organization's internal personnel noted substantial productivity gains, with a reduction in data analyst query times and data pipeline construction by 50 percent and a decrease in time for data discovery from two days to two hours. Productivity benefits were evenly distributed across various roles, with data scientists, data engineers, and data analysts all reporting time savings of 25 to 30 percent. The organization's adoption of Databricks also spurred revenue growth by enabling behavior-driven sales and supporting acquisitions, particularly offering personalized incentives for the organization's subscription-based initiatives. The media conglomerate also achieved a significant reduction in carbon usage following deployment through more insightful emissions tracking. Going forward, the organization has identified generative AI initiatives powered by Databricks as potential areas for future exploration. The company recommends the immediate adoption of Databricks SQL Serverless and Unity Catalog, managing operations in code, and maintaining a close partnership with Databricks as best practices.

PROFESSIONAL SPORTS FRANCHISE

This American professional sports franchise owns and operates four teams across the United States and the Dominican Republic. The organization relies heavily on data to differentiate its operations across its player analytics, scouting, player development, and sports science teams, with 59.4 TBs of current storage. Before adopting Databricks, the organization leveraged a Snowflake data warehouse built on AWS and S3 storage but noted that the platform was brittle and expensive, with limited support for advanced analytics. Transformations were also challenging in Snowflake, forcing the organization to use a separate tool for its data transformations, which was inefficient and ate up compute. The organization started exploring new solutions in late 2021 and began evaluating Databricks as a potential replacement for Snowflake and its transformation point solution. The organization decided to adopt Databricks in November 2022 and spent the following six months moving relevant data and workloads over to the platform. This process was performed by the organization's sports analytics team, data engineering team, and sports systems team in conjunction with multiple side projects to develop net-new pipelines.

After deployment, the sports franchise was able to retire its transformation tool as well as various cloud-native services, saving over $150,000 per year. The organization also substantially reduced its Snowflake spend and projects $350,000 in total savings through 2025. Although this benefit is realized gradually, the organization noted that it plans to cut its Snowflake spend entirely by 2026. Additionally, the sports franchise reduced its maintenance and administrative expenses across its data engineering and governance teams, yielding $320,000 in annual savings. Processing was also more cost-efficient in Databricks, with the organization noting 4x cost efficiency relative to Snowflake. This benefit was especially pronounced for data ingestion use cases, which were 10x more cost efficient. The organization noted that savings from this alone provided the necessary budget for all of its novel AI use cases. Data storage was also more cost-efficient with Databricks' medallion architecture, generating $96,000 in annual savings. Users within the organization also cited time savings across various internal roles. Data scientists saw 50 percent time savings with MLflow and AutoML in Databricks and brought AI models to production 2.5x faster. Data engineers saw even more significant time savings of at least 60 percent with up to 15x faster development for certain workloads. Data analysts saw the most improvement, with 75 percent time savings due to the ease of bringing in new data sources and less time spent on engineering workloads. This has improved the synergy between players, coaches, scouts, and operations teams, with data consumers seeing 2x more data with less latency to iterate faster on ad-hoc analysis. The sports franchise also established MLOps practices, which have been essential as it scales its analytics operations. This approach was largely enabled by Unity Catalog, which provided governance and permissioning for the franchise's global workforce. This implementation of Unity Catalog not only managed dataset availability but also governed which features and AI models certain teams and user groups could access.

THE ROI GUIDEBOOK METHODOLOGY

Based on the ROI assessments developed through Nucleus's in-depth interviews with Databricks customers, Nucleus has developed an ROI framework for organizations that are considering a Databricks investment. The framework can be used by potential and existing customers to understand the cost, benefit, and deployment factors that impact their potential return on investment. The Nucleus ROI Guidebook development process includes the following:

Technology review. Nucleus interviewed Databricks product managers and subject matter experts, participated in product demonstrations, and conducted a full review of technical documents and data sheets to gather data on the Databricks Lakehouse Platform.

Customer interviews. Nucleus analysts conducted in-depth interviews with five organizations that were using the Databricks Lakehouse Platform to understand their business challenges, strategy, deployment processes, costs incurred, benefits achieved, and best practices learned from their deployments.

ROI assessments. Based on the data collected from customers, Nucleus completed an ROI assessment of each customer's deployment and validated that ROI audit with each customer's project team leadership.

$$\text{ROI} = \frac{((\text{net benefit in year one} + \text{net benefit in year two} + \text{net benefit in year three}) / 3)}{\text{total initial cost}}$$

Construction of aggregate ROI framework and analysis. Nucleus constructed a financial model based on its NASBA-registered ROI methodology, using the data from Nucleus's ROI business case assessments of the customers that were interviewed. All financial metrics presented in this report are calculated based on standard NASBA accounting principles commonly used by certified finance professionals.

Benefits guidance. Based on the variability and clustering of benefits in the aggregate, Nucleus provides appropriate averages, ranges, and estimation factors to guide other customers in using the framework to develop their own ROI projections.

NUCLEUS
RESEARCH

Nucleus Research, Inc. | Miami, FL

Nucleus Research provides the ROI, insight, benchmarks, and facts that allow clients to understand the value of technology and make informed decisions. All research is built on an in-depth, case-study research approach that analyzes the results of actual deployments to provide technology advice built on real-world outcomes, not analyst opinions. Learn more at

NucleusResearch.com

Nucleus Research Inc. | 5555 Biscayne Boulevard, Miami, FL, 33137 | +1 (617) 720-2000