databricks

Guide

Cloud Modernization

A business guide to the hidden value of migrating from Hadoop to Databricks

GAVIN EDGLEY | ANAND VENUGOPAL | BRIAN DIRKING | MANVEER SAHOTA

databricks

Contents

Introduction 3

Infrastructure Costs: Much More Than Licensing 4

CAPEX vs. OPEX: Pay Only for What Is Used 6

Raising Productivity 7

Business-Impacting Use Cases 7

Customer Examples 11

- Nationwide 11
- Scribd 12
- Picsart 13

Don't Two-Step It 14

Building Internal Expertise 14

databricks

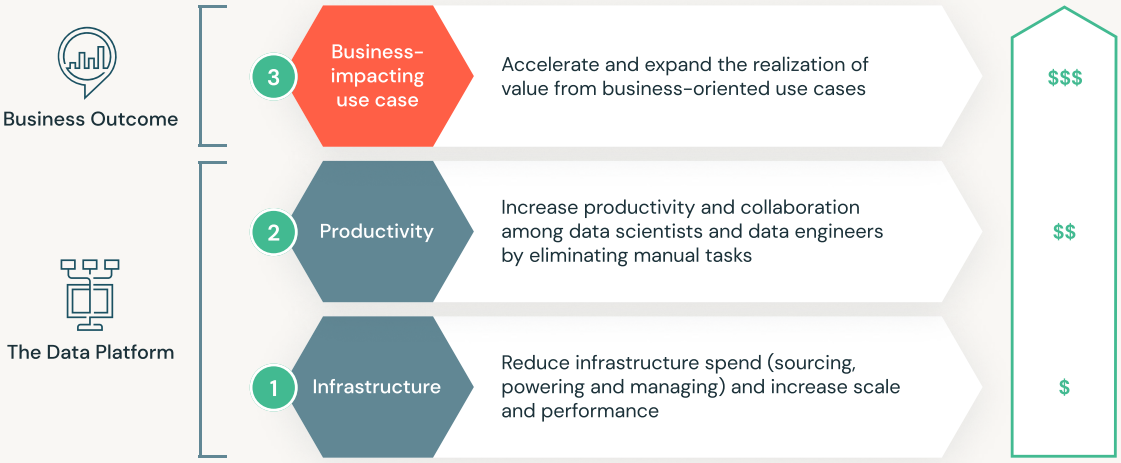
Introduction

Running on-premises or legacy Hadoop-based architectures is no longer feasible for any organization whose data and AI strategies index on growth and innovation. Over the past year, we have seen an acceleration in the number of customers migrating from a Hadoop architecture to a modern cloud lakehouse architecture. Many organizations have made the move in order to reduce the operational costs of licenses and maintenance. But they've also discovered that the new capabilities and use cases unlocked with modern cloud-based analytics and an AI platform quickly outweigh the cost of migration and create long-term exponential growth.

This guide will help you uncover some of the hidden value your organization can reap when it migrates from Hadoop to the Databricks Lakehouse Platform. In fact, by modernizing your Hadoop environment, you can realize value across three areas: infrastructure, productivity and business outcomes.

databricks

The value of migrating to Databricks



Infrastructure Costs: Much More Than Licensing

Most organizations have aspirations to become analytics and AI-based companies, but they have found that Hadoop has not delivered. Some of the inherent limitations of a Hadoop architecture are:

1. **Data processing:** Development SLAs are too slow to provide data in a timely manner. Often there is a backlog of use cases waiting to be addressed, and the cost of delivering business-critical data sets is just too expensive.
2. **Governance:** The systems don't provide the governance and management needed to truly build a self-service data culture driven by analytics and AI
3. **Machine learning:** The systems do not include an embedded machine learning platform
4. **Evolving data and AI landscape:** The platforms can't keep up with the rapidly evolving tools and frameworks for ML and AI
5. **Management overhead:** Software upgrades take significant time and are resource intense, robbing the team of time to innovate by just trying to keep up
6. **Capacity problems:** Hadoop clusters have to be provisioned for peak loads and result in wasted costs or end up being overutilized becoming growth blockers. Compute and storage need to grow independently.

databricks

As more companies migrate to modern cloud data and AI platforms, Hadoop providers have raised licensing costs to make up for their losses, which is only accelerating the migration. Organizations tend to focus on the comparative costs of licensing, and Hadoop's subscription fees alone make a compelling case to migrate. But what gets lost in such a comparison is that a platform change is really an urgent strategic initiative that can ensure the long-term success of the organization. To get a true sense of what Hadoop is costing your organization, you have to step back. From a benchmark of 10 Databricks customers, we found that licensing accounts for less than 15% of the total cost — it's the tip of the iceberg. The other costs are made up of the following:

Data center management. This accounts for nearly half the total cost. It includes property costs, cooling and management. Power often costs $800 per server per year, based on consumption and cooling, leading to an $80K annual bill for a 100-node Hadoop cluster. There is also the cost of environmental by-products from running data centers, which counters any enterprise ESG initiatives.

Hardware growth and upgrades. There are additional significant hardware costs for servers, storage and networking with data growth because storage and compute cannot be separated. Periodic maintenance and upgrade costs also add up.

Administration of the Hadoop clusters. Many organizations assume four to eight full-time high-value resources for every 100 nodes.

8%

"Only 8% of the big data projects are regarded as VERY successful."

— CAPGEMINI

85%

"Close to 85% of big data projects fail."

— GARTNER

95%

More than 95% of committed Databricks customers meet their objectives and timelines.

Customers see a number of ways their total cost of ownership is lower with Databricks in the cloud than running Hadoop on-premises.

databricks

CAPEX vs. OPEX: Pay Only for What Is Used

Databricks is priced based on consumption — you only pay for what you use.
But Databricks is a more economical solution in other ways too:

1. Autoscaling ensures customers only pay for the infrastructure they use
2. With a cloud-based platform, capacity can scale to meet changing demand instantly, not in days, weeks or months
3. Storage and compute are kept separate, so adding more storage does not require adding expensive compute resources at the same time
4. Databricks gives users the flexibility to select GPUs and other high-performance processing options to increase performance even further, as well as to select lower-performance processing for lower-cost daily jobs
5. Expensive data center management and hardware costs disappear entirely

Faster processing with Databricks means beating SLAs *and* keeping costs down:

Founded by the original creators of Apache Spark™ and Delta Lake, Databricks delivers the most highly tuned processing engine in the world — up to 50x faster than open source Spark

Databricks has also introduced open source Delta. Running on Databricks as the Delta Lake service, it provides many more performance improvements, optimizing data through features such as Z-Ordering, data skipping and file compaction

Take advantage of all data types (streaming, structured, semi-structured and unstructured) immediately as they're introduced

databricks

Raising Productivity

Data scientists and data engineers are scarce and expensive resources. One of the best ways organizations can improve ROI and boost their bottom line is to maximize the productivity of data teams.

The big surprise for many customers is how the Databricks platform facilitates collaboration. Databricks makes data teams more efficient by providing one unified platform for data management, data analytics and BI, data science and AI/ML while creating a seamless connection to popular third-party BI and ML tools. For example, Databricks notebooks break down silos and cumbersome processes to enable seamless collaboration. In the past, teams would typically correspond via email or tickets, but now they can ideate and comment directly in the notebooks. Many teams talk about how this has had a 10x impact on accelerating innovation.

Business-Impacting Use Cases

With Databricks, customers are able to move beyond the limitations of Hadoop and finally address business-critical use cases. These organizations find that the power of a modern cloud-based data and AI platform quickly over-delivers on the actual cost of migration due to the platform's ability to address more advanced use cases by using a larger amount of data at lower cost and with greater speed.

- **Deliver data to business users** faster for better and more timely business decisions
- **Deliver consistent data** from a shared data lake that's properly governed to ensure the entire organization is working off the same data
- **Achieve greater scalability** to take on the largest AI and ML use cases — such as for curing diseases and detecting intrusions — and enable more effective analytics and AI by finding new markets, increasing revenue, reducing costs and lowering risk
- **Make larger data sets available** to business decision-makers — and give them the ability to visualize your entire data lake — while keeping costs low through a pay-for-what-you-use architecture

databricks

By migrating workloads and data from a legacy Hadoop environment, data teams can begin delivering production use cases at scale across industries to streamline business operations and, more importantly, build data products that can be quickly monetized. Here are a few use cases customers are unlocking with the Databricks Lakehouse Platform:

- Customer lifetime value
- Subscriber churn prediction
- Customer retention
- Recommendation engines
- Faster, more accurate demand forecasting
- Clinical data lake
- Safety stock analysis
- Alternative data for investing
- ESG investing
- Predictive maintenance (IoT)
- Risk/value at risk calculation
- Quality of service video streaming analytics
- Ad effectiveness with forecasting and attribution
- Threat detection at scale with DNS analytics
- Disease prediction
- Digital pathology image analysis
- Anomaly detection with geo clustering
- Reputation risk
- Transaction enrichment
- Rules-based AI for financial fraud prevention
- Product matching with ML
- Building forward-looking intelligence with external data
- Modernizing investment data platforms
- Toxicity detection in gaming
- Cyber analytics with Splunk and Databricks
- Multi-touch advertising attribution

Visit databricks.com/solution/accelerators to see how you can begin doing more with your data in production and at scale.

Databricks drives value for customers in three areas: infrastructure, productivity and business impact.

1. Databricks lowers infrastructure costs with optimized Apache Spark and Delta Lake performance gains, and through automation that manages clusters more efficiently and cost effectively

2. Databricks makes data teams more productive through a unified, open and collaborative platform that reduces the complexity of multiple tools and handoffs that plagues many data science and data engineering teams

3. Databricks scales to handle the most impactful use cases — which typically require huge amounts of data — and provides collaborative tools that enable teams to work together more effectively and accelerate innovation

Organizations find that migrating to Databricks pays for itself quickly and puts them on course to having a much bigger impact as an organization driven by data and AI. In many cases, we've been able to automate portions of the migration process, reducing migration costs and duration significantly.

One way we help customers identify the benefits of migration is by making a thorough assessment of the current and projected costs of their existing platform and then comparing them to the costs of a cloud-based platform.

An example of some of the inputs we use in the model are shown in Figure 1. The numbers in the Value column represent an average we have seen in a group of customers (hence, numbers like 5.2 employees).

	UNITS	VALUE
How many **nodes** in your Hadoop cluster?	*# nodes*	156
How many **people** support your Hadoop cluster?	*# FTE*	5.2
When is your Hadoop **renewal**?	*Months from today*	3
How do you expect your **capacity** needs to grow?	*% growth per year*	20%
Total professional services costs for migration	*$*	$450,000
How long do you expect your migration to take?	*Months*	3

Figure 1: Model inputs

Figure 2 shows a typical customer result. These numbers are based on the averages from a set of real-life customers.

	UNITS	YEAR 1	YEAR 2	YEAR 3	TOTAL
DO-NOTHING SCENARIO	**$**	**$5,343,515**	**$7,418,418**	**$9,838,102**	**$22,600,035**
Total – Hadoop	$	$5,343,515	$7,418,418	$9,838,102	**$22,600,035**
Hardware	$	$624,000	$1,372,800	$2,271,360	**$4,268,160**
Hadoop administration	$	$1,178,315	$1,413,978	$1,696,774	**$4,289,067**
Data center costs	$	$2,574,000	$3,556,800	$4,736,160	**$10,866,960**
Hadoop license	$	$967,200	$1,074,840	$1,133,808	**$3,175,848**

Figure 2: Forecasted costs of current platform

Many customers experience similar savings. Figure 3 shows dollar amounts of cumulative costs from Hadoop ($18.7 million) and the corresponding Databricks cost ($4.9 million) over three years. Most organizations see payback within two quarters.

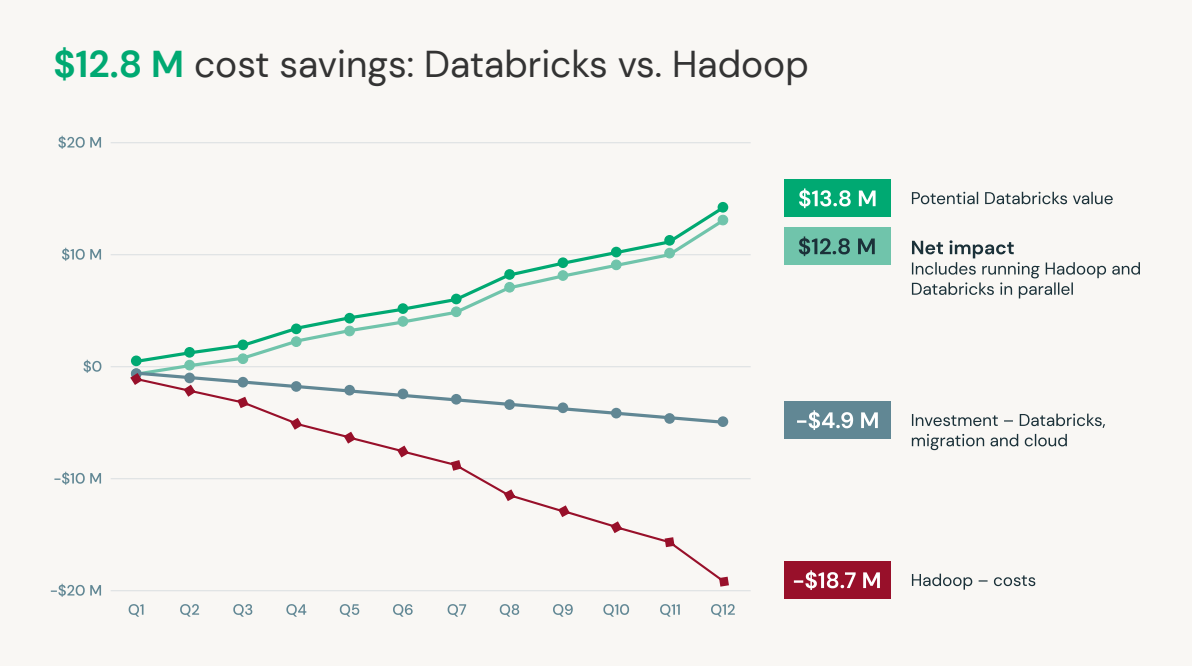


Figure 3: Hadoop costs vs. Databricks value and investment

Customer story

Nationwide®

Accurate insurance pricing with data and ML

Nationwide chose Databricks for actuarial modeling and is able to optimize insurance pricing by leveraging data and machine learning. Nationwide uses Databricks because they provide:

1. A unified data and AI platform to simplify infrastructure management, enabling fast data pipelines at scale and streamlining the ML lifecycle
2. A deep learning platform using hierarchical neural networks to provide more accurate pricing predictions, resulting in more revenue

Nationwide has benefited in a number of ways:

1. **Self-service:** Actuaries are now able to make decisions based on large volumes of data previously locked away in silos
2. **9x faster data pipelines**, improving runtime from 34 hours to less than 4 hours
3. **5x improvement** in featurization speeds for downstream ML
4. **50% reduction in time** to train and deploy ML models
5. **25%+ improvement in productivity** of high-value data engineers and data scientists

Customer story

SCRIBD

Migrating to the cloud to enable the world's largest digital library

Scribd is an eBook and audiobook subscription service that offers 1 million titles and hosts 60 million documents on their open-publishing platform. Scribd had a Hadoop-based "conventional data platform" with a Hadoop Distributed File System (HDFS) and a smattering of Hive.

Over time the business changed, and Scribd needed more machine learning, more real-time data processing and more support for teams collaborating to deliver new data products.

Their data platform now consists of a combination of Airflow, the Databricks Lakehouse Platform, Delta Lake and AWS Glue Catalog, a powerful data platform that has improved development velocity and collaboration significantly.

Scribd has backfilled their entire data warehouse into Delta Lake and has deployed new projects on Databricks while continuing the migration. Scribd **automated the migration of about 80%** of their Hive workloads over to Databricks with their own tooling.

"Databricks claimed an optimization of 30%–50% for most traditional Spark workloads. Out of curiosity, I refactored my cost model to account for the price of Databricks and the potential Spark job optimizations. After tweaking the numbers, I discovered that at a 17% optimization rate, Databricks would reduce our Amazon Web Services (AWS) infrastructure cost so much that it would pay for the cost of the Databricks platform itself. After our initial evaluation, I was already sold on the features and developer velocity improvements Databricks would offer. When I ran the numbers in my model, I learned that I couldn't afford not to adopt Databricks!"

R. Tyler Croy
Director of Platform Engineering
Scribd

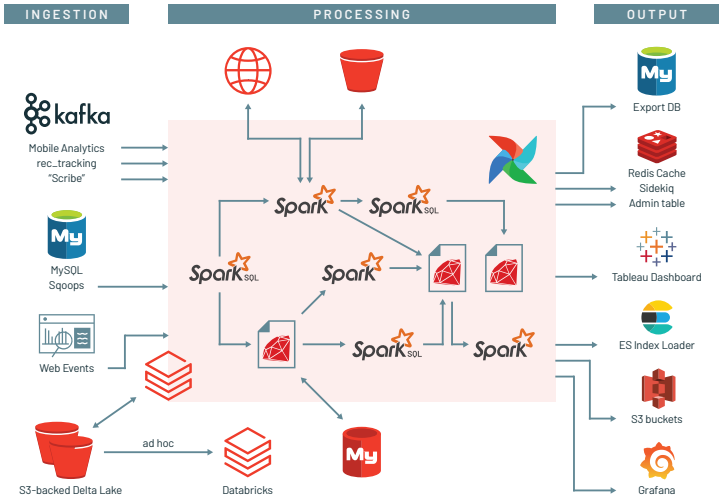


Figure 4: The Scribd data platform on AWS

Customer story

Picsart

A picture-perfect data architecture that drives growth and improves conversion

Picsart makes an app for photo editing. With 500 million-plus installs and 140 million monthly users, Picsart spans the globe. Their Hadoop architecture led to slow business insights and missed conversion opportunities. Their strategic objectives include:

1. Building a scalable data infrastructure to support the company's rapid growth
2. Getting insights faster so they can add critical new features to **increase engagement** and **improve conversion, such as**:
   - Exploratory analysis and A/B tests
   - Fraud detection ("copycats")
   - Recommendations (stickers, feed, etc.)

In their Hadoop-based environment, **storage** and **compute** were **tightly coupled**:

1. Adding new infrastructure was slow and expensive as physical servers are spun up
2. Stability was at risk as business demands fluctuated

Significant efforts were spent on **performance management** and maintenance. The infrastructure **limited A/B tests** to approximately 15 in parallel. There was a high potential to increase conversions with many more experiments. Just one test led to changes that increased **subscription rates by 15% per day**. PMs wanted to draw insights on user behavior but were limited by **analytics delays of one to three days**. The analytics team needed to build capabilities in **modern technologies**, such as structured streaming and an optimized data lake. The organization wanted to avoid rework when migrating to a modern architecture, and to leverage expertise to **get it right the first time**.

How Picsart used Databricks to increase conversions:

A scalable, reliable, managed architecture built natively in the cloud (Databricks Runtime, Delta Lake, etc.)

Collaborative workspaces that enable their data scientists, engineers and analysts to collaborate, accelerating their rate of innovation

Databricks is providing production support and expertise in professional services and training

Don't Two-Step It

Many organizations try to take their Hadoop experience and re-create it in the cloud, which re-creates the same problems and hurdles but in a new environment. Once again the promises of Hadoop go undelivered. To maximize business value, organizations should skip Hadoop in the cloud and future-proof their data and AI architecture by adopting the Databricks Lakehouse Platform.

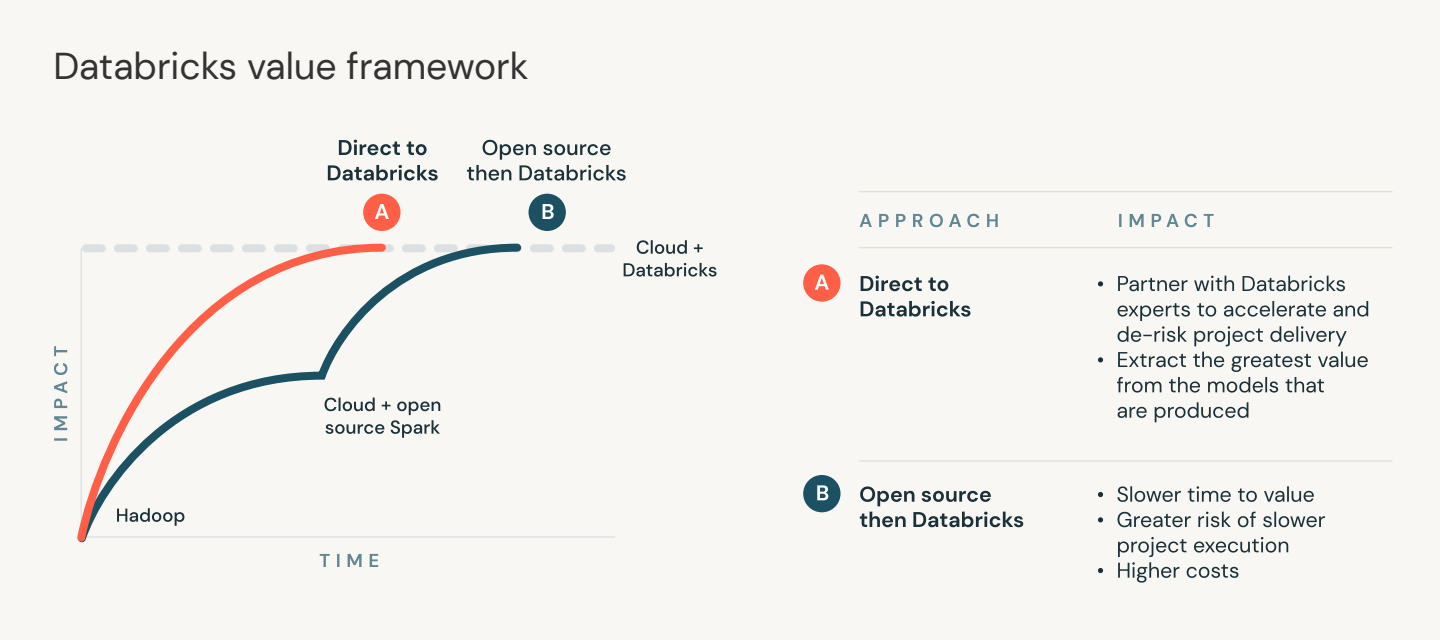


Figure 5: Value impact of direct migration

Building Internal Expertise

Your teams can take their knowledge of data architectures and apply it to Databricks. Over 100,000 people register for Data + AI Summit annually to find out how companies are moving forward with a modern cloud-based data, analytics and AI architecture. The combined open source projects created by Databricks, Apache Spark, Delta Lake, MLflow, Koalas and Redash have over 30 million monthly downloads, much larger compared to other tools — making it much easier to build teams. Databricks also provides free training by role.

databricks

About Databricks

Databricks is the data and AI company. More than 5,000 organizations worldwide — including Comcast, Condé Nast, H&M and over 40% of the Fortune 500 — rely on the Databricks Lakehouse Platform to unify their data, analytics and AI. Databricks is headquartered in San Francisco, with offices around the globe. Founded by the original creators of Apache Spark™, Delta Lake and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems. To learn more, follow Databricks on Twitter, LinkedIn and Facebook.

Evaluate Databricks for yourself

Start your free trial

Contact us for a personalized demo at databricks.com/contact

To learn more about migration, visit databricks.com/migration