databricks

WHITEPAPER

The Hidden Value of Hadoop Migration

A Business Value Whitepaper From Databricks

GAVIN EDGLEY | ANAND VENUGOPAL | BRIAN DIRKING

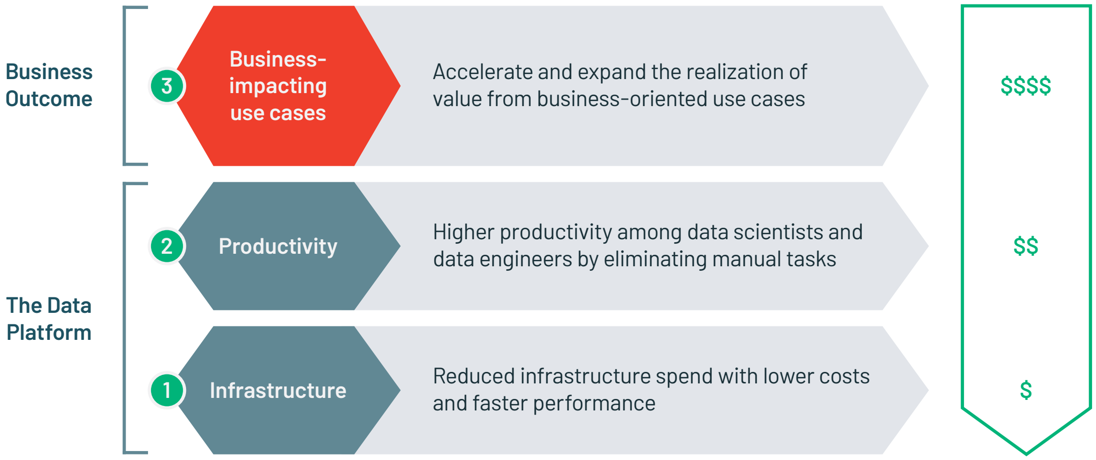
databricks

Contents

Introduction

Over the past year, we have seen an acceleration in customers migrating from a Hadoop architecture to a modern cloud architecture. Many organizations have made the move to reduce the operational costs of licenses and maintenance, but they've also discovered that the power of a modern cloud-based analytics platform quickly outweighs the cost of migration.

This whitepaper will help you to uncover some of the hidden value your organization can reap by migrating from Hadoop to a modern cloud-based analytics platform. You can analyze the value of migration by looking at three areas: infrastructure, productivity and business outcomes.


Business Outcome
3
Business-impacting use cases
Accelerate and expand the realization of value from business-oriented use cases
$$$$
The Data Platform
2
Productivity
Higher productivity among data scientists and data engineers by eliminating manual tasks
$$
1
Infrastructure
Reduced infrastructure spend with lower costs and faster performance
$

Infrastructure Costs: Much More Than Licensing

Many organizations have found that Hadoop has not delivered on their aspirations to become an analytics-based company. Some of the inherent limitations of a Hadoop architecture are:

1. The development SLAs are too slow to provide the data needed 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. The systems don't provide the governance and management needed to truly build an analytics-driven, self-service data culture
3. The systems do not include an embedded machine learning platform
4. The platforms can't keep up with the rapidly evolving tools and frameworks for ML and AI
5. Software upgrades take significant time and are resource intense, robbing the team of time to innovate by just trying to keep up
6. Managing capacity can be difficult and time consuming since adding Hadoop data nodes just for storage is cost prohibitive. Compute and storage need to grow independently.

As more companies migrate to modern cloud platforms, Hadoop providers have raised license costs to make up their losses. This cycle has led more companies to migrate from Hadoop on-premises to lower their total cost of ownership. These organizations focus on the comparative cost of licensing, which alone makes a compelling case to migrate, but doesn't show the true value that makes a platform change an urgent project for 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 is less than 15% of total cost. The other costs are made up of the following:

- **Data center management** is nearly half the total cost. This 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.
- **Excess hardware.** Overcapacity is common for on-premises implementations because it allows you to scale up to meet your largest needs, but much of that capacity sits idle most of the time. The ability to separate storage and compute does not exist, so costs grow as data sets grow, and most organizations have big data sets.
- **Administration** of the Hadoop clusters. Many organizations assume 4–8 full-time employees 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 objective 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.

Lowering Infrastructure Costs — Customers Pay for Only What They Use

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

- Autoscaling ensures customers only pay for the infrastructure they use
- In the cloud, capacity can scale to meet changing demand in minutes, not weeks or months
- Storage and compute are kept separate, so adding more storage does not require adding expensive compute resources at the same time
- Databricks enables users to select GPUs and other high-performance processing options to increase performance even further, but then to also select lower-performance processing for lower-cost daily jobs
- Expensive data center management and hardware costs disappear entirely

The faster processing of 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 data immediately as it is introduced with streaming data, and combine it with historical data to provide instant insights that are critical in security and financial services use cases

Raising Productivity

Data scientists and data engineers are expensive resources. One of the best ways organizations can save money is to maximize the productivity of data teams.

The big surprise for many customers is how the Databricks platform facilitates collaboration. Data teams are more efficient because Databricks Notebooks enable collaboration. Typically teams would correspond via email or Jira tickets, now they just comment 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 address more critical use cases. These organizations find that the power of a modern cloud-based analytics platform quickly outstrips the cost of the migration due to the ability to address more advanced use cases.

- **Delivering data to business users** faster for better and more timely business decisions
- **Delivering consistent data** from a shared data lake properly governed to ensure the entire organization is working off the same data
- **Greater scalability** to take on the largest ML and AI use cases — such as curing diseases and intrusion detection — to enable analytics with greater impact through finding new markets and increasing revenue, reducing costs, lowering risk
- **Larger historical data sets** are available to business decision makers — providing the ability to visualize your entire data lake, while keeping costs low through a pay-for-what-you-use architecture

INFRASTRUCTURE
PRODUCTIVITY
BUSINESS IMPACT

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 automation that manages clusters more efficiently and cost effectively
2. Databricks makes data teams more productive through a unified, collaborative workspace that reduces the complexity of multiple tools and handoffs that plague many data science and data engineering teams
3. Databricks scales to handle the most impactful use cases, which often require huge amounts of data, and through collaborative tools that enable the 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 have a much bigger impact as an analytics-driven organization. We have found that in many cases we can automate portions of the migration process, reducing migration costs and duration significantly.

Part of the way we help customers identify the impact of migration is a thorough assessment of customer costs as well as a forecast of their future costs with their current platform versus a new 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 over a group of customers (hence, numbers like 5.2 employees).

	UNITS	VALUE
How many **nodes** in your Hadoop cluster?	*# nodes*	156
How many **people** supporting 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

A typical customer result looks like Figure 2. In this scenario, we show the net savings each year by migrating to Databricks. 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

We can show this impact for many customers, as in Figure 3. This figure shows dollars of cumulative present value over three years. Most organizations see payback within two quarters.

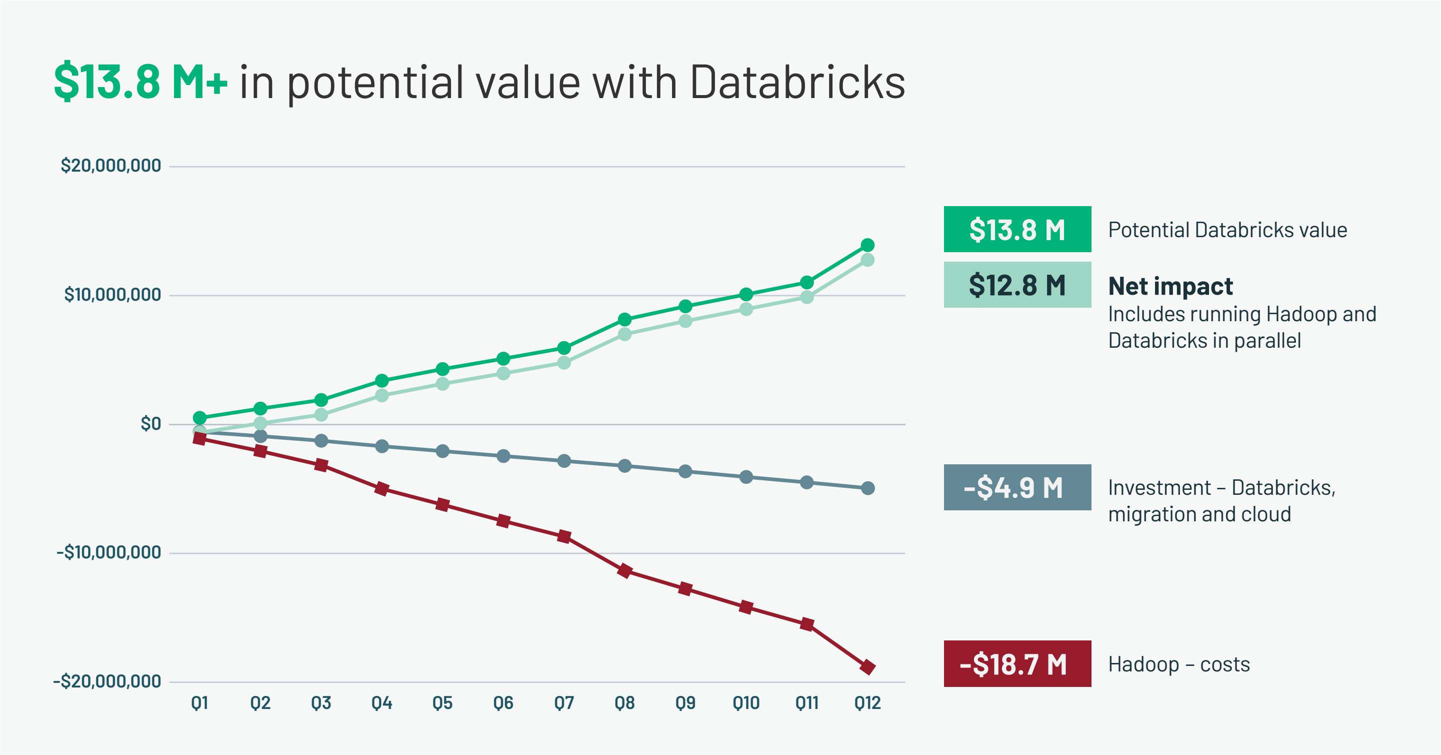
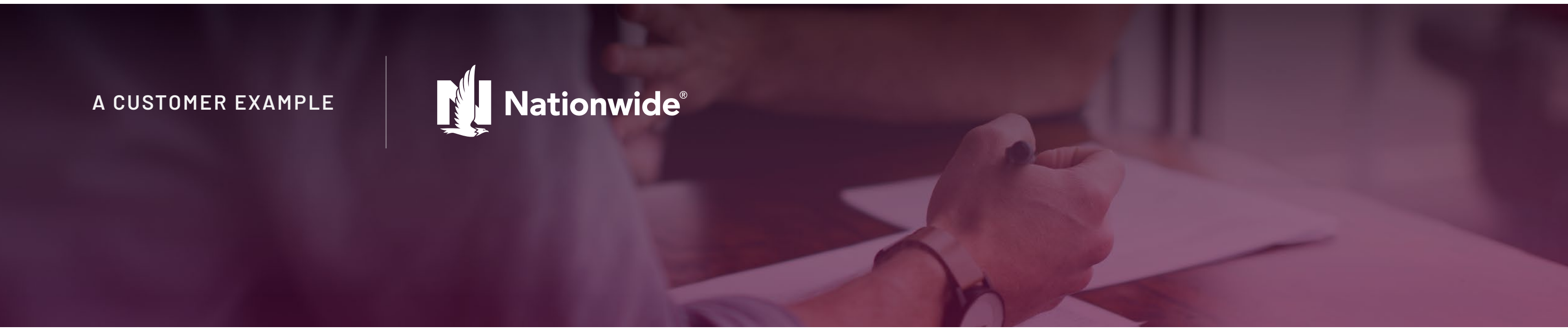


Figure 3: Hadoop costs vs. Databricks value and investment



A CUSTOMER EXAMPLE

Nationwide®

Nationwide chose Databricks for actuarial modeling. They are able to optimize insurance pricing by leveraging data and machine learning. Nationwide chose Databricks because it provides:

1. A unified 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 seen positive impact 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

A CUSTOMER EXAMPLE

SCRIBD

Scribd is an American eBook and audiobook subscription service that includes 1 million titles and hosts 60 million documents on its 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, Databricks, 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 onto Databricks while continuing the migration. Scribd automated migration for 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 AT SCRIBD

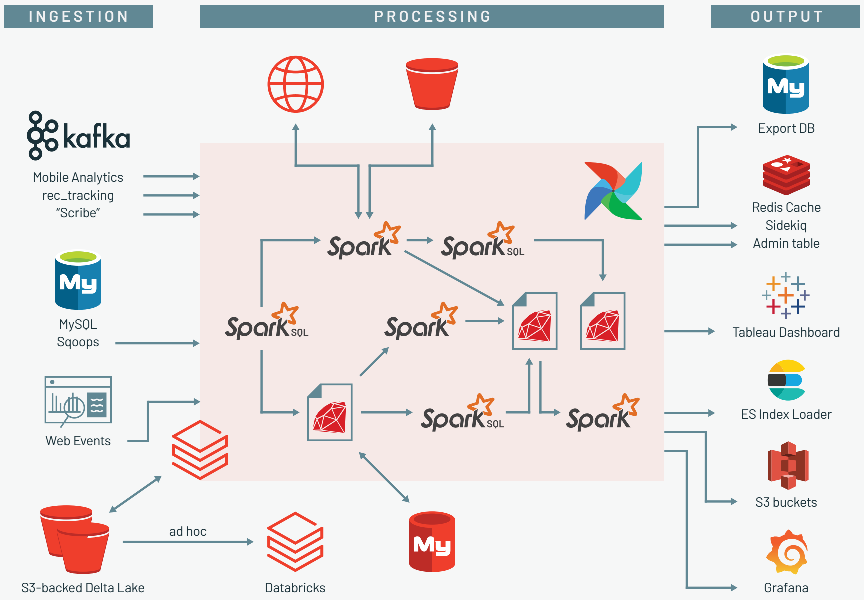


Figure 4: The Scribd data platform on AWS

A CUSTOMER EXAMPLE | PicsArt

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. Build a scalable data infrastructure to support the company's rapid growth
2. Provide faster insights to add critical new features to **increase engagement** and **improve conversion**:
   a. Exploratory analysis and A/B tests
   b. Fraud detection ("copy cats")
   c. 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 to **1–3 day analytics delays**. 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 leverage expertise to **get it right the first time**.

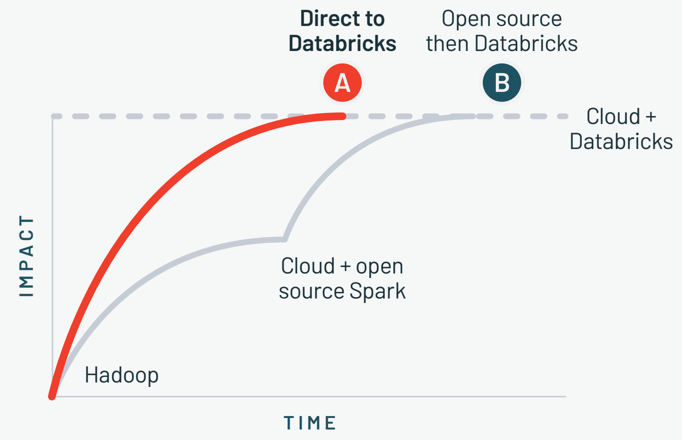
Databricks brought value to PicsArt in these ways:

1. A scalable, reliable, managed architecture built natively in the cloud (Databricks runtime, Delta Lake, etc.)
2. Collaborative workspaces that enable their data scientists, engineers and analysts to collaborate, accelerating their rate of innovation
3. 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. They bring the same problems with them. The faster way to maximize business value is going directly to Databricks.

Databricks value framework



APPROACH	IMPACT
A Direct to Databricks	• Partner with Databricks experts to accelerate and de-risk project delivery • Extract the greatest value from the models that are produced
B Open source then Databricks	• Slower time to value • Greater risk of slower project execution

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 Summits annually to find out how companies are moving forward with a modern cloud-based data and analytics architecture. The Apache Spark community has over 500,000 members, much larger compared to other tools — making it much easier to build teams. Databricks also provides free training, including this series.

databricks

EVALUATE DATABRICKS FOR YOURSELF

START YOUR FREE TRIAL

Contact us for a personalized demo databricks.com/contact

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