databricks

Transitioning from Hadoop to Databricks Lakehouse Platform

More than 5,000 organizations worldwide and over 40% of the Fortune 500 rely on the Databricks Lakehouse Platform to unify their data, analytics, and AI.

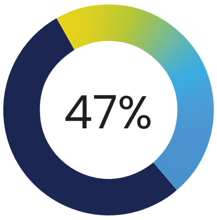
Founded by the original creators of the open-source Apache Spark™, Delta Lake, and MLFlow, the cloud-native Databricks Lakehouse Platform is being widely adopted across diverse industries for its ability to seamlessly integrate data, analytics, and AI workloads. The Lakehouse stands out by harmoniously combining the strengths of a data lake and a data warehouse, offering a unified solution that not only ensures high performance but also prioritizes cost-efficiency for businesses operating in various sectors.

Built on top of the popular Apache Spark framework, Databricks features a highly tuned processing engine that goes beyond conventional capabilities, delivering remarkable performance gains of up to 50 times faster than Spark. This exceptional speed and efficiency make Databricks a compelling choice for organizations seeking a robust and agile solution to propel their data, analytics, and AI initiatives to new heights in today's dynamic business landscape.

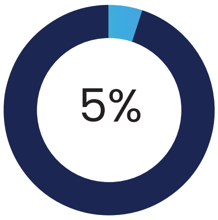
Why do organizations migrate to Databricks?

417%

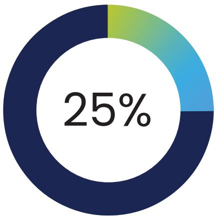
ROI is achieved when companies switch to Databricks.



Cost savings



Revenue growth



Increase in data team's productivity

Source: Forrester

Hadoop Vs Databricks

	hadoop	databricks
Infrastructure Management and Licensing Costs	Hadoop users grapple with high data center management costs, including servers, storage, and networking. Coupled storage and compute escalate hardware expenses. Nearly 75% of overall expenses relate to infrastructure, with 15% attributed to licensing costs.	Databricks offers flexibility, eliminating high data center and hardware costs. Autoscaling ensures organizations only pay for actual usage. Besides, compute and storage are billed separately. Users can opt for lower-performance processing for daily tasks and high-performance options to reduce overall expenses.
Productivity	Hadoop lacks a unified platform, impacting the data team's productivity. Siloed data management, analytics, and AI/ML processes hinder collaboration. Communication bottlenecks often arise, impacting overall productivity.	Databricks raises team productivity by up to 10x. The unified platform breaks down organizational silos, encourages cross-functional collaboration, and eliminates redundancies through tools like the Databricks Notebook, fostering direct connectivity and brainstorming among teams.

Innovation and Scalability	Hadoop faces challenges in efficiently scaling for advanced AI and ML use cases. Hadoop's complexities and batch processing model hinder agility and data monetization. Unlike Databricks, Hadoop may lack a unified view of the entire data lake, making holistic decision-making more cumbersome.	Databricks enables better scalability for large AI and ML use cases, streamlining business operations and facilitating the creation of data products for rapid monetization. The platform offers a single view of the entire data lake for holistic decision-making and robust data governance.
Security and Compliance	Hadoop's open-source framework requires companies to implement security features manually, such as setting up Access Control Lists (ACLs) for authorization and ensuring the secure transmission of data between nodes. Also, additional encryption, auditing, and access controls are required to ensure compliance with specific regulations such as GDPR, HIPAA, etc.	Databricks' built-in security features combined with the security features from the underlying cloud provider make a robust foundation. It allows users to offer role-based access controls, and encryption at rest and in transit, and integrate with cloud provider identity and access management services. This makes it easier to meet the various compliance certifications and attestations.

Hadoop's cost challenges and lack of a unified platform hinder productivity and innovation. Organizations find that shifting workloads from Hadoop to Databricks results in lowered costs and greater agility.

For businesses seeking a low-cost/low-risk migration approach, Databricks becomes the platform for increasing revenue, reducing costs, and lowering risk through the adoption of AI/ML use cases at scale. Besides, Databricks allows the automation of portions of the migration process, which brings down migration costs further. To optimize business value and future-proof their data and AI architecture, embracing the Databricks Lakehouse Platform holistically becomes more rewarding.

Migrating From Hadoop to Databricks

A Case in Point

How an Audio Streaming Services Company Builds ML and Real-time Data Processing Capabilities with Databricks

Business Challenge

An audio subscription services company hosting millions of titles and documents on its open-publishing Hadoop-based platform faced many limitations. Its conventional data platform, comprising HDFS and Hive, struggled to meet its machine learning and real-time processing requirements. It also hindered team collaboration to deliver new data products.

Solution

The company's IT consulting partners revamped their data platform by leveraging a combination of technologies including Airflow, Databricks Delta Lake, and AWS Glue Catalog. The data was seamlessly migrated from their data warehouse to the Databricks Delta Lake and the technology experts used advanced tools to automate the migration of about 80% of the company's Hive workloads.

Value Gained

This transformation significantly increased the company's development speed, scalability, and collaboration. The modernized and innovative infrastructure allowed the company to deploy new projects on Databricks and leverage AI/ML seamlessly in the future.

Mapping Hadoop Components to Databricks' Equivalents

Correctly mapping legacy Hadoop technologies to Databricks’ modern cloud capabilities is the key to a successful migration:

Hadoop Component	Databricks Equivalent
Data Storage	
HDFS atop block storage Kafka HBase	• Cloud object storage: S3, ADLS, Azure Blob • Cloud-native message bus: Kinesis, Azure Event Hubs, Azure IoT Hub • Cloud-native NoSQL: DynamoDB, Cosmos DB
Data Processing	
MapReduce Pig HiveQL Spark	• Databricks Delta Engine: Optimized Apache Spark for 10x-100x improvement • Databricks SQL: ANSI SQL 2023 compliant • Code-free ETL: Integrations with Azure Data Factory mapping flows, Prophecy, Talend, and more
Jobs	
Oozie (workflow automation)	• Databricks job scheduler • Native integration with Apache Airflow and Azure Data Factoryflow and Azure Data Factory • Use any scheduler via Databricks APIs
Code Development	
Apache Zeppelin notebook Jupyter notebook	• Databricks notebook • Support for Zeppelin, Jupyter, any notebook or IDE (Pycharm, IntelliJ, etc.) of your choice via Databricks APIs

About KANINI

KANINI is a digital transformation enabler, providing cutting-edge software services and solutions that help enterprises drive innovation and business growth. We create impeccable customer experiences through thoughtfully designed digital solutions that help improve our customer's efficiency, scale, and revenues.

We specialize in Cloud Modernization, Data Analytics & AI, Product Engineering, and ServiceNow Consultation and Implementation—all delivered through flexible engagement models.

We focus on empowering Banking and Financial Services, Healthcare, and Manufacturing, among other industries to harness the power of cloud technologies and solutions by implementing agile development practices and a global delivery framework. Find more about our Data Analytics & AI solutions and Consulting services here:
https://kanini.com/data-analytics-consulting/

USA | Colombia | Ukraine | UAE | Singapore | India
Contact us at transformations@kanini.com | 615-465-8287