

EBOOK

Unlock the Potential Inside Your Data Lake

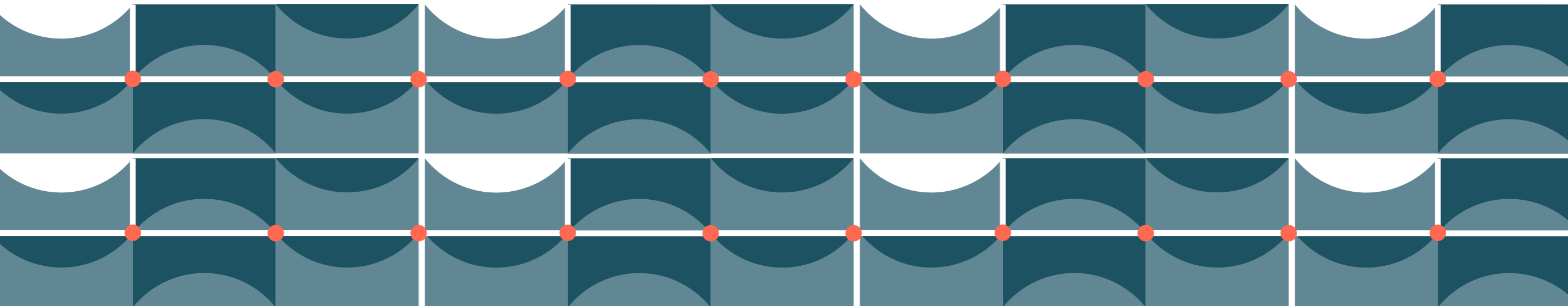
Building a Lakehouse with Databricks on AWS at Comcast, Condé Nast, and Showtime

Contents

- Executive Summary3
- The Value of Data and AI Initiatives4
- The Data Warehouse vs. Data Lake Challenge.....5
- Introducing Delta Lake and Lakehouse 6
- Collaboration and Streamlined Workflows for Data Scientists and Data Engineers at Scale7
- Delta Lake on Databricks: The Latest Innovation on the Lakehouse Platform8
- Photon: The Next-Generation Query Engine for the Lakehouse8
- Business-Impact Use Cases 9
- Customer Stories
 - Comcast10
 - Condé Nast..... 11
 - Showtime12
- Summary13

Executive Summary

Data analytics and artificial intelligence power most of the significant innovation and success across industries today. When it comes to leveraging the capabilities of data science, artificial intelligence, and machine learning (ML), FAANG companies (Facebook, Apple, Amazon, Netflix and Google) are leading the way, but traditional industry organizations are also capable of achieving demonstrative innovation and accelerated growth by prioritizing the adoption of AI and data science tools and initiatives.



The Value of Data and AI Initiatives

Over the last few years, FAANG companies and other organizations that accelerate their paths to differentiation and success practice four common trends with their data:

01 They embrace machine learning and artificial intelligence as guides for the future of their data initiatives

Fully embracing ML and AI capabilities allows organizations to gain unprecedented and predictive insights into customer behavior, based on real-time data from their environment. These insights also enable organizations to develop strategies for real-time responses to events for an increasingly streamlined customer experience and more operational dexterity.

To simplify the implementation of ML initiatives, organizations should consider building data platforms from the ground up. Attempting to add data science and machine learning capabilities to an existing data platform can be particularly challenging. Most of what powers data science and machine learning initiatives tends to be unstructured data, which operates at a level lower than the structured, clean data powering most data warehousing applications. Moving lower in the stack is difficult so choosing data platforms that can handle unstructured data and structured data with agility from the start is crucial.

02 They embrace open source and open formats

Leveraging open source formats accelerates access to the latest innovations available. Open source spaces serve as powerful databases when organizations seek to hire teams that are already well versed in the implementation and development of the most innovative applications. Companies that have prioritized open source and open formats also tend to attract talent with a stronger connection to and more knowledge of the open source world and its benefits.

In contrast with proprietary formats, open standards also enable faster and more streamlined migrations and integration of new technology. The ability to swiftly and seamlessly migrate from proprietary format platforms liberates teams to achieve faster innovation and drive more impactful results.

03 They prioritize cloud migration

As a pioneer in cloud computing, many organizations are migrating to the AWS cloud for greater growth and agility. For the best results, many are choosing to adopt cloud native tools at the start of their journeys rather than lift and shift strategies that take what was once on premises and replicate it in the cloud.

All the data does not necessarily live in a single cloud, and AWS has the capabilities to ensure that all users can absorb, manage and gain visibility into their data everywhere it lives.

04 They simplify data architecture

To optimize agility and cost efficiency, successful companies are seeking a single model for governance, security and lineage over all data sources. They also seek to reduce the copies of their data and to develop a single source of truth.

The Data Warehouse vs. Data Lake Challenge

There are key distinctions and incompatibilities across the various tools, governance models, and languages that are integral to the implementation of data science and machine learning initiatives. These differences can make it challenging for organizations to efficiently collaborate, integrate tasks and achieve their goals. Moving and copying data across these silos is often time consuming and tedious, leading to the protraction or outright abandonment of many data and AI initiatives.

In most instances, these challenges result from the reality that there are two distinct data paradigms: data lakes and data warehouses. This paradigm sets up a binary system that fragments the ways that we think about, use and manage our data.

This paradigm also requires organizations to create two copies of their data. Data lakes tend to be based in open standards and open formats, while data warehouses tend to be exclusively proprietary formats that only take structured data. As a result, organizations are left with two sources of their data that are not necessarily compatible.

Effectively operating with these systems requires distinct approaches and skill sets. Most users of data lakes understand Python or Java, while most users of data warehouses understand SQL. Consequently, the ecosystem gets built up around this binary. There also tends to be incompatibilities when it comes to security and governance models.



Introducing Delta Lake and Lakehouse

Databricks has introduced Delta Lake, a highly secure and integrated data management framework. As an open source format layer, Delta Lake replaces data silos, providing a single source of truth for all types of data including structured, semi-structured and unstructured. By centralizing and integrating data, Delta Lake optimizes security, functionality and accelerates performance for users.

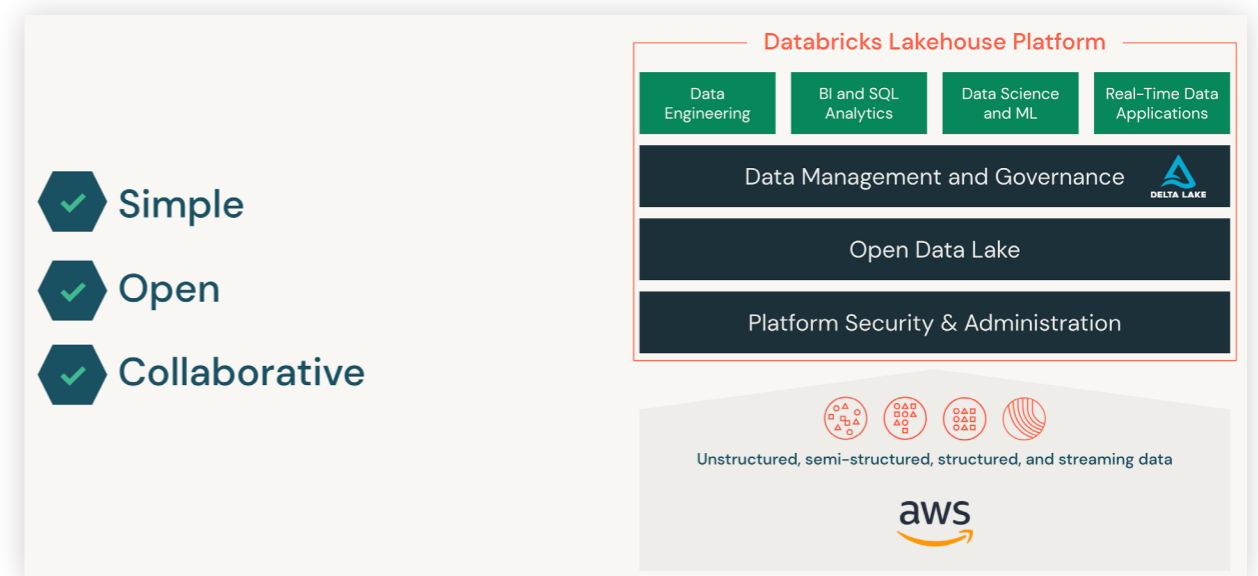
With Delta Lake, we are seeing the rise of the lakehouse, the ability for a data lake to handle the key use cases of a data warehouse, enabling one source of data for all the audiences and use cases of an organization.

Delta Lake has addressed the key shortcomings of data lakes: reliability and performance, while retaining the cost model of the data lake, and the ability to handle unstructured information.

As the ideal foundation for a highly scalable and cost-effective lakehouse, Delta Lake also supports real-time streams, enabling data teams to operate with the most up-to-date and accurate information.

With Delta Lake on AWS, data teams can now operate with more ease and dexterity by working from a single copy of their data. This data can also be efficiently and securely shared with all authenticated users across all major internal data and analytics workloads as well as with external organizations. With the lakehouse built completely on open standards and open source, an organization maintains full control over its data as it travels. With Delta Lake, users have the flexibility to bring in new tools and new systems into their organization's lakehouse without having to worry about whether or not they're compatible with a proprietary format.

Through a central data governance interface, users can access a fine grained table, row or column-based approach inside of the data lake. This means that the high level of governance that one might typically attribute to a data warehouse is now available for a data lake. This supports the generation and management of a single common security model and governance model, across all of an organization's workloads. Users can now complete data warehousing workloads and any additional ones on their data lake as a single source of truth.



Collaboration and Streamlined Workflows for Data Scientists and Data Engineers at Scale

The Unity Catalog UI in Databricks provides organizations with a central point of access for data security and governance, auditing, sharing and discovery across all clouds.

Data teams have access to a notebook-based interface, from which they can centrally operate in a diverse array of programming languages including Python, R and Scala. From this centralized interface, users can quickly and securely build, train, deploy and manage those models.

The Auto ML feature in Databricks enables users to rapidly scale machine learning throughout an environment without extensive knowledge of the science behind ML operations. This allows teams to focus on fine tuning models and maximizing accuracy and performance instead of deep diving on ML.

The Databricks interface also supports collaboration, allowing data scientists and other authenticated stakeholders across an organization to join in on workflows as a team in real time. The collaborative capabilities available through Databricks help teams identify problems and arrive at solutions more quickly by democratizing their collective access to insights.

Open standard settings enable data engineers to access powerful native job orchestration capabilities available through Databricks, as well as plug in their preferred tools like dbt and Airflow. This ease of flow between native and plug-in tools remains seamless at scale.

Delta Lake on Databricks: The Latest Innovation on the Lakehouse Platform

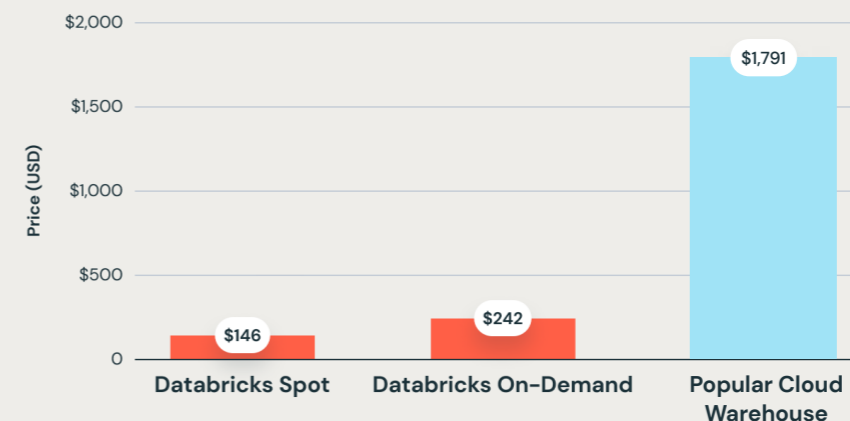
There are other innovations in Delta Lake, including its capacity to support a number of performance enhancements for analytics workloads on an organization's data lake. As a platform that launches 9 million VMs every single day and processes 9 exabytes of data, Databricks is calibrated to handle immense data processing jobs with reliable ease and speed.

Photon: The Next-Generation Query Engine for the Lakehouse

As the latest native execution engine at Databricks, Photon accelerates the time for queries to finish on the existing compute mechanisms. For data warehousing, users can expect exceptional query performance. With Photon, Databricks recently set the world record for the TPC DS 100 terabyte benchmark.

[Read the blog →](#)


Barcelona Supercomputing Center Price/Performance for test derived from TPC-DS 100TB Power run (lower is better)



Databricks was also independently verified as the highest performing data warehouse to participate in that test at a 12 times better price/performance than the other cloud data warehouses available on the market.


Customer Use Cases Across Industries

The Databricks Lakehouse Platform can address any machine learning use case. These use cases can increase revenue, decrease costs, or decrease risk. Many of these use cases are impactful across many different industries, and Databricks provides these building blocks for customers to modify for their specific data and purpose.




Media and Entertainment

- Quality of Service
- Advertising Effectiveness
- Customer Attrition
- Churn Prediction
- Customer Lifetime Value
- Recommendation
- ...



Retail and Manufacturing

- Demand Forecasting
- Time-Series Forecasting
- Safety Stock
- Customer Lifetime Value
- Customer Attrition
- Customer Churn
- ...



Financial Services

- Market Risk
- Reputation Risk
- Fraud Detection
- Financial Time Series
- ESG Analytics
- Operationalizing Sustainability
- ...



Healthcare and Life Sciences

- Genomics
- Detecting At-Risk Patients
- HL7 Streaming
- Clinical Delta Lake
- Digital Pathology
- ...

Customer Story:

Comcast

Comcast required a new infrastructure to effectively support its expanding data and ML needs. The company struggled with massive data, fragile data pipelines, and poor data science collaboration. Databricks, including Delta Lake and MLflow, enabled the company to build performant data pipelines for petabytes of data and efficiently manage the lifecycle of 100s of models to deliver an award winning viewer experience using voice recognition and machine learning.

With Databricks Delta Lake and MLflow, the company completed a successful overhaul of their approach to analytics. Through modernization, the company was able to successfully and seamlessly deploy new machine learning features that enhanced the customer experience. Delta Lake also enabled reliable ETL at scale, as well as the optimizing of files for fast and reliable ingestion.

While providing a personalized, Emmy-Award winning viewing experience for customers, Comcast experienced a 10x reduction in its overall compute costs to process data. The company also experienced a 90% reduction in the devops resources needed to manage its infrastructure.

Customer Story: Condé Nast

As Condé Nast delivers iconic content to over 1 billion customers through its print, online, video, and social media platforms, the company needed a way to enhance its data science productivity and infrastructure management. The company faced challenges with data complexity, silos, and datasets that were outgrowing existing data lakes.

With Databricks delivering an end-to-end solution, Condé Nast has enabled cluster automation that has eliminated unnecessary DevOps efforts. Delta Lake has also supported Comcast in building data pipelines that scale to 1 trillion data points per month. Data science innovation has been unlocked with a collaborative environment with MLflow to manage the entire ML lifecycle. This has allowed them to deliver personalized content across their brands to engage and retain customers.

Databricks has supported Condé Nast in facilitating better collaboration between data engineering and data science teams to drive more innovative solutions and experiences for customers. Databricks has also supported Condé Nast in significantly improving the data pipeline, reducing process times by 60%. Databricks has also enabled a 50% reduction in Condé Nast's IT operational costs and supported business growth through a faster time-to-insight.

Customer Story: Showtime

To drive new innovations that could enhance the customer experience, Showtime needed to get more value from its data, but the company needed to overcome the scaling limitations of its legacy systems and inefficient data pipelines before that would be possible.

With the Databricks Lakehouse Platform, Showtime has now acquired an actionable view into the consumer journey to inform programming and content with the goal of increasing engagement while lowering churn.

With Databricks, Showtime can fully optimize its fully managed, serverless cloud infrastructure and leverage a simplified and streamlined ML lifecycle with MLflow. The company is now experiencing 6x faster data pipelines, and a significant reduction in cloud infrastructure complexity. Showtime has been able to enhance the customer experience. With data science collaboration and increased productivity, Showtime has significantly reduced the time-to-market for new features, giving subscribers access to a more personalized and faster viewing experience.

Summary

Comcast, Condé Nast, and Showtime are all creating a lakehouse to address the use cases of their many different audiences – from analysts to data scientists. These companies have created automated data pipelines that combine streaming and batch data to provide insights that are shaping their businesses.

Delta Lake on Databricks helps organizations across industries get the most from their data and analytics initiatives to drive innovation, growth and differentiation. This unified lakehouse approach provides a secure and reliable open storage format layer. Delta Lake also supports stronger collaboration, performance and efficiency for teams by providing a single source of truth for structured, semi-structured, and unstructured data. [Try the Databricks Lakehouse Platform](#) and see how it can impact your organization.

Databricks is the data and AI company. More than 7,000 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](#).

START YOUR FREE TRIAL

