

# The Delta Lake Series Customer Use Cases

---

See how customers are using  
Delta Lake to rapidly innovate



# What's inside?

The Delta Lake Series of eBooks is published by Databricks to help leaders and practitioners understand the full capabilities of Delta Lake as well as the landscape it resides in. This eBook, **The Delta Lake Series – Customer Use Cases**, shares examples of how real-life customers are using Delta Lake to solve challenging problems with data.

# What's next?

After reading this eBook, you'll understand how customers are using Delta Lake's capabilities to create best-in-class solutions for a variety of business challenges.

## Here's what you'll find inside

---

### Introduction

#### What is Delta Lake?

#### Chapter 01

**USE CASE #1: Healthdirect Australia**  
Provides Personalized and Secure Online Patient Care With Databricks

#### Chapter 02

**USE CASE #2: Comcast**  
Uses Delta Lake and MLflow to Transform the Viewer Experience

#### Chapter 03

**USE CASE #3: Viacom18**  
Migrates From Hadoop to Databricks to Deliver More Engaging Experiences

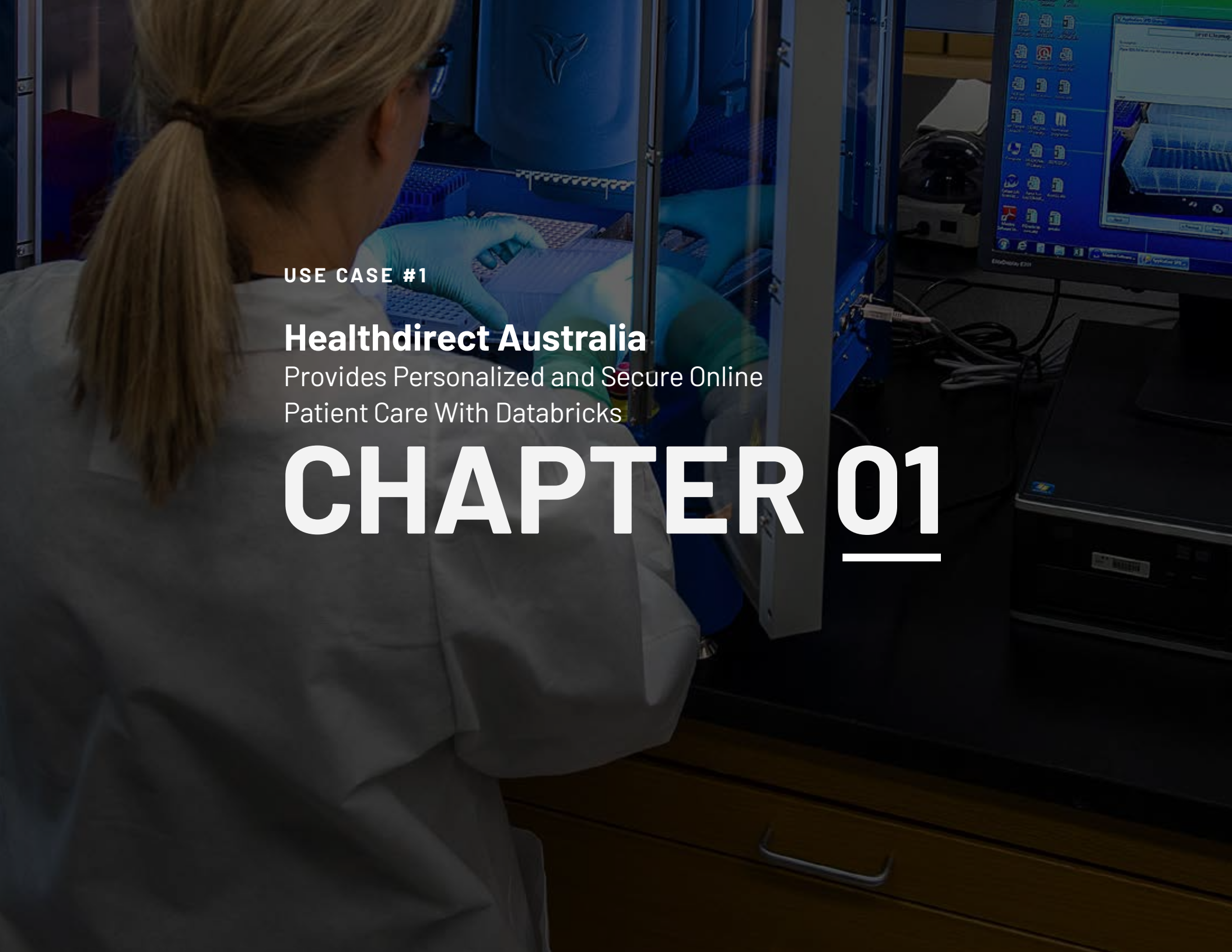


# What is Delta Lake?

[Delta Lake](#) is a unified data management system that brings data reliability and fast analytics to cloud data lakes. Delta Lake runs on top of existing data lakes and is fully compatible with Apache Spark™ APIs.

At Databricks, we've seen how Delta Lake can bring reliability, performance and lifecycle management to data lakes. With Delta Lake, there will be no more malformed data ingestion, difficulties deleting data for compliance or issues modifying data for data capture.

With Delta Lake, you can accelerate the velocity that high-quality data can get into your data lake, and the rate that teams can leverage that data with a secure and scalable cloud service.

A scientist with blonde hair in a ponytail, wearing a white lab coat and blue gloves, is working in a laboratory. She is positioned in front of a biosafety cabinet, handling a multi-well plate. To her right, a computer monitor displays a software interface with various icons and a window showing a 3D model of a tray. The scene is dimly lit, with a blueish tint from the lab equipment.

USE CASE #1

## Healthdirect Australia

Provides Personalized and Secure Online  
Patient Care With Databricks

# CHAPTER 01

# 01

## Healthdirect Australia

Provides Personalized and Secure Online Patient Care With Databricks

As the shepherds of the National Health Services Directory (NHSD), Healthdirect is focused on leveraging terabytes of data covering time-driven, activity-based healthcare transactions to improve health care services and support. With governance requirements, siloed teams and a legacy system that was difficult to scale, they moved to Databricks. This boosted data processing for downstream machine learning while improving data security to meet HIPAA requirements.

### Spotlight on Healthdirect

**INDUSTRY:** Healthcare and life sciences

**6x**

Improvement in data processing

**20M**

Records ingested in minutes

### Data quality and governance issues, silos, and the inability to scale

Due to regulatory pressures, Healthdirect Australia set forth to improve overall data quality and ensure a level of governance on top of that, but they ran into challenges when it came to data storage and access. On top of that, data silos were blocking the

team from efficiently preparing data for downstream analytics. These disjointed data sources impacted the consistency of data reads, as data was oftentimes out-of-sync between the various systems in their stack. The low-quality data also led to higher error rates and processing inefficiencies. This fragmented architecture created significant operational overhead and limited their ability to have a comprehensive view of the patient.

Further, they needed to ingest over 1 billion data points due to a changing landscape of customer demand such as bookings, appointments, pricing, eHealth transaction activity, etc. — estimated at over 1TB of data.

“We had a lot of data challenges. We just couldn’t process efficiently enough. We were starting to get batch overruns. We were starting to see that a 24-hour window isn’t the most optimum time in which we want to be able to deliver healthcare data and services,” explained Peter James, Chief Architect at Healthdirect Australia.

Ultimately, Healthdirect realized they needed to modernize their end-to-end process and tech stack to properly support the business.

## Modernizing analytics with Databricks and Delta Lake

Databricks provides Healthdirect Australia with a Unified Data Platform that simplifies data engineering and accelerates data science innovation. The notebook environment enables them to make content changes in a controlled fashion rather than having to run bespoke jobs each time.

“Databricks has provided a big uplift for our teams and our data operations,” said James. “The analysts were working directly with the data operations teams. They are able to achieve the same pieces of work together within the same time frames that used to take twice as long. They’re working together, and we’re seeing just a massive acceleration in the speed at which we can deliver service.”





## Faster data pipelines result in better patient-driven healthcare

With Delta Lake, they've created logical data zones: Landing, Raw, Staging and Gold. Within these zones, they store their data "as is," in their structured or unstructured state, in Delta Lake tables. From there, they use a metadata-driven schema and hold the data within a nested structure within that table. What this allows them to do is handle data consistently from every source and simplifies the mapping of data to the various applications pulling the data.

Meanwhile, through Structured Streaming, they were able to convert all of their ETL batch jobs into streaming ETL jobs that could serve multiple applications consistently. Overall, the advent of Spark Structured Streaming, Delta Lake and the Databricks Unified Data Platform provides significant architectural improvements that have boosted performance, reduced operational overheads and increased process efficiencies.

As a result of the performance gains delivered by Databricks and the improved data reliability through Delta Lake, Healthdirect Australia realized improved accuracy of their fuzzy name match algorithm from less than 80% with manual verification to 95% and no manual intervention.

The processing improvements with Delta Lake and Structured Streaming allowed them to process more than 30,000 automated updates per month. Prior to Databricks, they had to use unreliable batch jobs that were highly manual to process the same number of updates over a span of 6 months – a 6x improvement in data processing.

"Databricks delivered the time to market as well as the analytics and operational uplift that we needed in order to be able to meet the new demands of the healthcare sector."

– Peter James, Chief Architect, Healthdirect Australia

They were also able to increase their data load rate to 1 million records per minute, loading their entire 20 million record data set in 20 minutes. Before the adoption of Databricks, this used to take more than 24 hours to process the same 1 million transactions, blocking analysts from making swift decisions to drive results.

Last, data security, which was critical to meet compliance requirements, was greatly improved. Databricks provides standard security accreditations like HIPAA, and Healthdirect was able to use Databricks to meet Australia's security requirements. This yielded significant cost reductions and gave them continuous data assurance by monitoring changes to access privileges like changes in roles, metadata-level security changes, data leakage, etc.

"Databricks delivered the time to market as well as the analytics and operational uplift that we needed in order to be able to meet the new demands of the healthcare sector," said James.

With the help of Databricks, they have proven the value of data and analytics and how it can impact their business vision. With transparent access to data that boasts well-documented lineage and quality, participation across various business and analyst groups has increased – empowering teams to collaborate and more easily and quickly extract value from their data with the goal of improving healthcare for everyone.📍





A person is operating a professional camera on a set. The camera has a monitor attached to the top, which displays a video feed of two men in a meeting. The person is wearing a head-mounted display (HMD) and is looking at the camera. The background is a blurred office setting with other people.

USE CASE #2

## Comcast

Uses Delta Lake and MLflow to  
Transform the Viewer Experience

# CHAPTER 02

# 02 Comcast

## Uses Delta Lake and MLflow to Transform the Viewer Experience

### Spotlight on Comcast

Industry: Media and entertainment

**10x**

Reduction in overall compute costs to process data

**90%**

Reduction in required DevOps resources to manage infrastructure

**Reduced**

Deployment times from weeks to minutes

As a global technology and media company connecting millions of customers to personalized experiences, Comcast struggled with massive data, fragile data pipelines and poor data science collaboration. With Databricks – leveraging Delta Lake and MLflow – they can build performant data pipelines for petabytes of data and easily manage the lifecycle of hundreds of models to create a highly innovative, unique and award-winning viewer experience using voice recognition and machine learning.

## Infrastructure unable to support data and ML needs

Instantly answering a customer's voice request for a particular program while turning billions of individual interactions into actionable insights, strained Comcast's IT infrastructure and data analytics and data science teams. To make matters more complicated, Comcast needed to deploy models to a disjointed and disparate range of environments: cloud, on-premises and even directly to devices in some instances.

- **Massive data:** Billions of events generated by the entertainment system and 20+ million voice remotes, resulting in petabytes of data that need to be sessionized for analysis.
- **Fragile pipelines:** Complicated data pipelines that frequently failed and were hard to recover. Small files were difficult to manage, slowing data ingestion for downstream machine learning.
- **Poor collaboration:** Globally dispersed data scientists working in different scripting languages struggled to share and reuse code.
- **Management of ML models:** Developing, training and deploying hundreds of models was highly manual, slow and hard to replicate, making it difficult to scale.
- **Friction between dev and deployment:** Dev teams wanted to use the latest tools and models while ops wanted to deploy on proven infrastructure.





## Automated infrastructure, faster data pipelines with Delta Lake

Comcast realized they needed to modernize their entire approach to analytics from data ingest to the deployment of machine learning models to delivering new features that delight their customers. Today, the Databricks Unified Data Platform enables Comcast to build rich data sets and optimize machine learning at scale, streamline workflows across teams, foster collaboration, reduce infrastructure complexity, and deliver superior customer experiences.

- **Simplified infrastructure management:** Reduced operational costs through automated cluster management and cost management features such as autoscaling and spot instances.

- **Performant data pipelines:** Delta Lake is used for the ingest, data enrichment and initial processing of the raw telemetry from video and voice applications and devices.
- **Reliably manage small files:** Delta Lake enabled them to optimize files for rapid and reliable ingestion at scale.
- **Collaborative workspaces:** Interactive notebooks improve cross-team collaboration and data science creativity, allowing Comcast to greatly accelerate model prototyping for faster iteration.
- **Simplified ML lifecycle:** Managed MLflow simplifies the machine learning lifecycle and model serving via the Kubeflow environment, allowing them to track and manage hundreds of models with ease.
- **Reliable ETL at scale:** Delta Lake provides efficient analytics pipelines at scale that can reliably join historic and streaming data for richer insights.

## Delivering personalized experiences with ML

In the intensely competitive entertainment industry, there is no time to press the Pause button. Armed with a unified approach to analytics, Comcast can now fast-forward into the future of AI-powered entertainment — keeping viewers engaged and delighted with competition-beating customer experiences.

- **Emmy-winning viewer experience:** Databricks helps enable Comcast to create a highly innovative and award-winning viewer experience with intelligent voice commands that boosts engagement.
- **Reduced compute costs by 10x:** Delta Lake has enabled Comcast to optimize data ingestion, replacing 640 machines with 64 while improving performance. Teams can spend more time on analytics and less time on infrastructure management.
- **Less DevOps:** Reduced the number of DevOps full-time employees required for onboarding 200 users from 5 to 0.5.
- **Higher data science productivity:** Fostered collaboration between global data scientists by enabling different programming languages through a single interactive workspace. Also, Delta Lake has enabled the data team to use data at any point within the data pipeline, allowing them to act more quickly in building and training new models.
- **Faster model deployment:** Reduced deployment times from weeks to minutes as operations teams deployed models on disparate platforms.🔗





USE CASE #3

## **Viacom18**

Migrates From Hadoop to Databricks  
to Deliver More Engaging Experiences

# CHAPTER 03

## USE CASE #3

# 03

## Viacom18

### Migrates From Hadoop to Databricks to Deliver More Engaging Experiences

Viacom18 Media Pvt. Ltd. is one of India's fastest-growing entertainment networks with 40x growth over the past decade. They offer multi-platform, multigenerational and multicultural brand experiences to 600+ million monthly viewers.

In order to deliver more engaging experiences for their millions of viewers, Viacom18 migrated from their Hadoop environment due to its inability to process data at scale efficiently. With Databricks, they have streamlined their infrastructure management, increased data pipeline speeds and increased productivity among their data teams.

Today, Viacom18 is able to deliver more relevant viewing experiences to their subscribers, while identifying opportunities to optimize the business and drive greater ROI.

#### Spotlight on Viacom18

Industry: Media and entertainment

**26%**

Increase in operational efficiency lowers overall costs

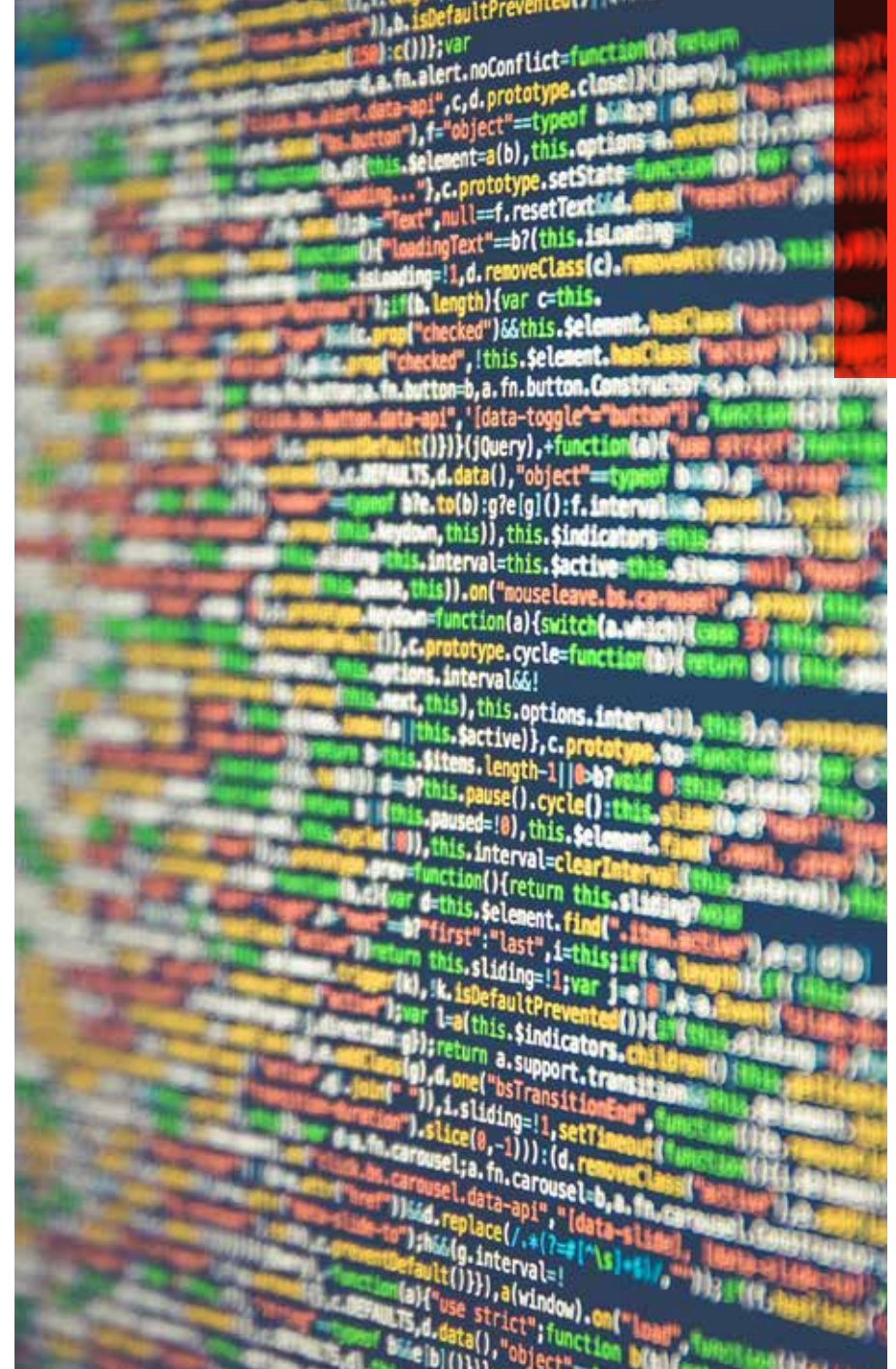
## Growth in subscribers and terabytes of viewing data push Hadoop to its limits

Viacom18, a joint venture between Network18 and ViacomCBS, is focused on providing its audiences with highly personalized viewing experiences. The core of this strategy requires implementing an enterprise data architecture that enables the building of powerful customer analytics on daily viewer data. But with millions of consumers across India, the sheer amount of data was tough to wrangle: They were tasked with ingesting and processing over 45,000 hours of daily content on VOOT (Viacom18's on-demand video subscription platform), which easily generated 700GB to 1TB of data per day.

"Content is at the heart of what we do," explained Parijat Dey, Viacom18's Assistant Vice President of Digital Transformation and Technology. "We deliver personalized content recommendations across our audiences around the world based on individual viewing history and preferences in order to increase viewership and customer loyalty."

Viacom18's data lake, which was leveraging on-premises Hadoop for operations, wasn't able to optimally process 90 days of rolling data within their management's defined SLAs, limiting their ability to deliver on their analytics needs, which impacted not only the customer experience but also overall costs.

To meet this challenge head-on, Viacom18 needed a modern data warehouse with the ability to analyze data trends for a longer period of time instead of daily snapshots. They also needed a platform that simplified infrastructure by allowing their team to easily provision clusters with features like auto-scaling to help reduce compute costs.







## Rapid data processing for analytics and ML with Databricks

To enable the processing power and data science capabilities they required, Viacom18 partnered with Celebal Technologies, a premier Salesforce, data analytics and big data consulting organization based in India. The team at Celebal leveraged Azure Databricks to provide Viacom18 with a Unified Data Platform that modernizes its data warehousing capabilities and accelerates data processing at scale.

The ability to cache data within Delta Lake resulted in the much-needed acceleration of queries, while cluster management with auto-scaling and the decoupling of

storage and compute simplified Viacom18's infrastructure management and optimized operational costs. "Delta Lake has created a streamlined approach to the management of data pipelines," explained Dey. "This has led to a decrease in operational costs while speeding up time-to-insight for downstream analytics and data science."

The notebooks feature was an unexpected bonus for Viacom18, as a common workspace gave data teams a way to collaborate and increase productivity on everything from model training to ad hoc analysis, dashboarding and reporting via PowerBI.

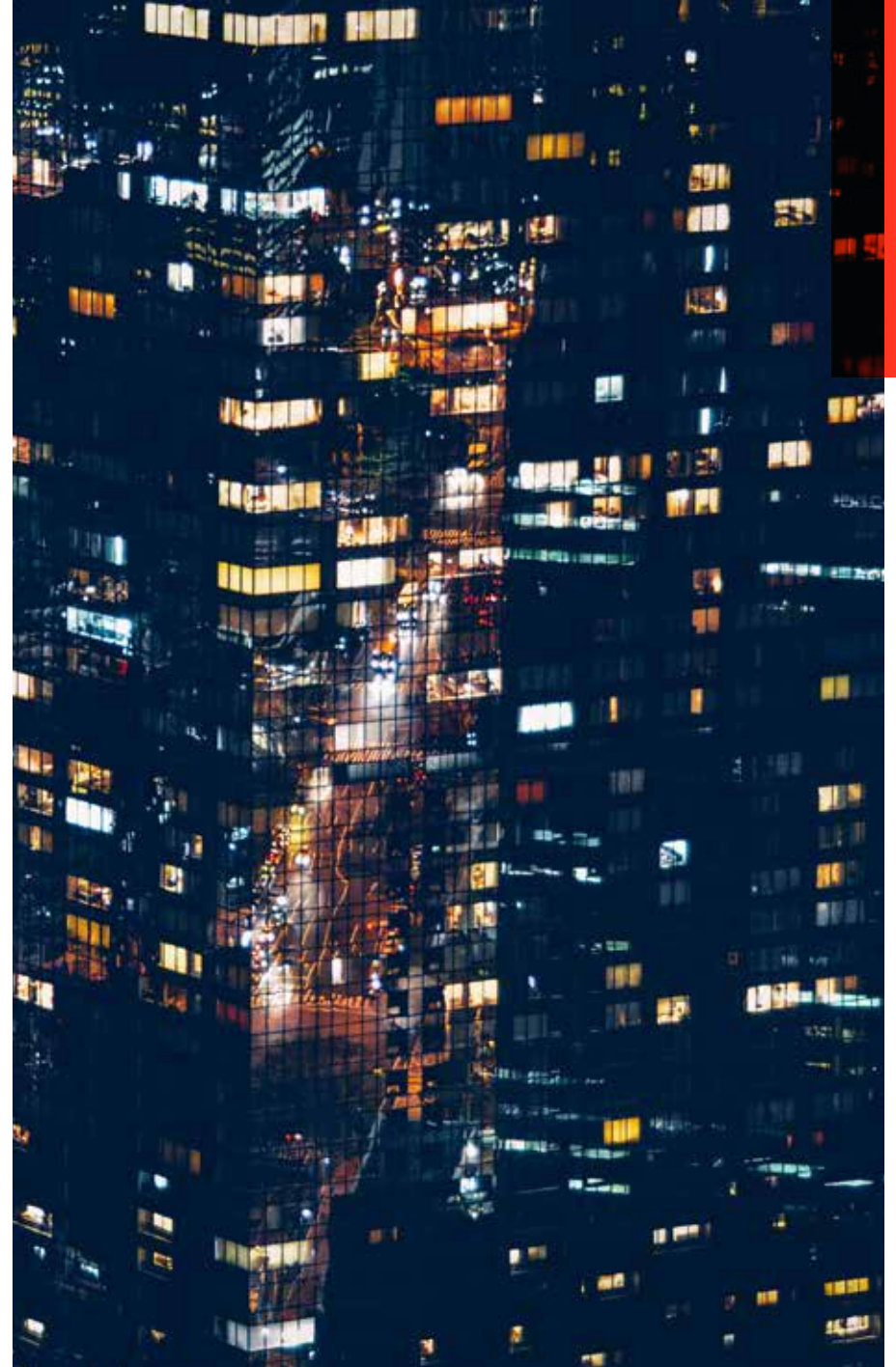
## Leveraging viewer data to power personalized viewing experiences

Celebal Technologies and Databricks have enabled Viacom18 to deliver innovative customer solutions and insights with increased cross-team collaboration and productivity. With Databricks, Viacom18's data team is now able to seamlessly navigate their data while better serving their customers.

"With Databricks, Viacom18's engineers can now slice and dice large volumes of data and deliver customer behavioral and engagement insights to the analysts and data scientists," said Dey.

In addition to performance gains, the faster query times have also lowered the overall cost of ownership, even with daily increases in data volumes. "Azure Databricks has greatly streamlined processes and improved productivity by an estimated 26%," concluded Dey.

Overall, Dey cites the migration from Hadoop to Databricks has delivered significant business value — reducing the cost of failure, accelerating processing speeds at scale, and simplifying ad hoc analysis for easier data exploration and innovations that deliver highly engaging customer experiences.📍



# What's next?

Now that you understand Delta Lake and how its features can improve performance, it may be time to take a look at some additional resources.

## **Explore subsequent eBooks in the collection >**

- [The Delta Lake Series – Fundamentals and Performance](#)
- [The Delta Lake Series – Features](#)
- [The Delta Lake Series – Lakehouse](#)
- [The Delta Lake Series – Streaming](#)

## **Do a deep dive into Delta Lake >**

- [Getting Started With Delta Lake Tech Talk Series](#)
- [Diving Into Delta Lake Tech Talk Series](#)
- [Visit the site for additional resources](#)

## **Try Databricks for free >**

## **Learn more >**