

Digital Turbine streams 180M events per day with Delta Lake running on Databricks with AWS



Databricks is an Advanced Technology Partner in the Amazon Partner Network.

Executive Summary

Digital Turbine's On-Device Media Platform enables end users to discover new apps, OEMs to promote their content, and advertisers to reach more consumers. The company runs on constant streams of real-time data to make sure the right app reaches the right user at exactly the right time. But after a major growth spurt—where Digital Turbine doubled its staff and data—it needed a new platform that could scale and still deliver accurate, precise, reliable, and fast data query responses. Using Delta Lake running on Databricks, Digital Turbine now makes petabytes of data available in minutes, delivering 180 million events per day.

Streaming platform reaches limit at 180M events per day

Digital Turbine relies on several streaming data sources to determine how best to serve apps to end users. The 180 million events that flow through the data system, include unified mobile events derived from installed app activities on mobile devices, as well as legacy mobile events, and campaign events from app installs and opens. All were brought together into a data sink where unified mobile events were used for BI analysis, legacy events were sent to Amazon Redshift, and campaign events were stored in a data lake then aggregated by Apache Spark and written to Redshift for revenue reports.

However, the Digital Turbine team was struggling to maintain an architecture that contained mostly self-managed clusters. They wanted to standardize their ETL approach and were only able to run batch ETLs for legacy events. The setup resulted in higher costs and the inability to deliver on service level agreements (SLA).

About Digital Turbine



Digital Turbine simplifies content discovery and delivers it directly to the device. Its on device media platform powers frictionless app and content discovery, user acquisition and engagement, operational efficiency and monetization opportunities.

CHALLENGE

Digital Turbine outgrew its event streaming platform and needed a solution that could scale to accommodate 180 million events per day.

SOLUTION

Using Delta Lake running on Databricks, Digital Turbine built a scalable event streaming platform in just one month that accelerated query responses.

RESULTS

- Built new event streaming platform in one month
- Created solution that makes data available in minutes
- Delivered 180 million events per day
- Achieved 99.99% SLA

“We needed to be able to deliver fast query responses to enable internal business analysts. Now they are using Databricks Notebooks to query Delta Lake directly.”

Daniel Ferrante
Director of Platform Engineering, Digital Turbine

Databricks on AWS delivers scale for faster queries

Using Spark and Delta Lake running on Databricks, Digital Turbine created a new architecture in just one month that improved data streaming and achieved a 99.99 percent SLA. The platform sends logging events to a microservice that publishes list events to Amazon Managed Streaming for Apache Kafka (Amazon MKS). From there, Databricks consumes the stream and writes the raw data to Delta Lake as a regionalized table in Amazon Simple Storage Service (Amazon S3). The new architecture enables real-time dashboards to track key performance indicators, ad-hoc queries via Notebooks and, fast transformations using Databricks fully-managed clusters. “We needed to be able to deliver fast query responses to enable internal business analysts,” said Daniel Ferrante, director of platform engineering. “Now they are using Notebooks to query Delta Lake directly.”

Accelerated data query responses

With auto-optimize on writes, Digital Turbine developers can continuously optimize the size of Amazon S3 parquet files, speeding up the time it takes to pull data. The highest load partner handles between 300 and 500 events per second. In addition, the team can easily join tables using Delta Lake, Amazon S3, and Redshift dimensions with Databricks. Through Notebook templates, developers can quickly access commonly used queries, instead of starting from scratch.

Improved data access and visibility

By integrating Delta Lake running on Databricks into its architecture, Digital Turbine made its unified mobile events data much more accessible to non-technical users. Now, business analysts can create reports in Tableau that are connected to Delta Lake, the operations team can more easily isolate production issues, and technical account managers can detect business issues earlier on.

About Databricks

Databricks is the data and AI company. Thousands of organizations worldwide—including Comcast, Condé Nast, Nationwide and H&M—rely on Databricks’ open and unified platform for data engineering, machine learning and analytics. Founded by the original creators of Apache Spark™, Delta Lake and MLflow, Databricks was first launched on AWS. For more information, visit www.databricks.com/aws.

