INDUSTRY
Technology

TECHNOLOGIES
Cloudera; Apache Hadoop and Impala; DataFox; Google BigTable, BigQuery, Machine Learning Engine, and Dataproc; AWS EMR and Simple Storage Service (S3).

BUSINESS NEED
The client's iteration of on-premise Cloudera Express was due for retirement, a development that would have seriously disrupted its machine learning processes for algorithm training within its own software product – a vital cog for a service underpinned by AI.

SOLUTION
Pythian advised the client of its two best options toward achieving continuity with its machine learning program while improving scalability: moving to AWS S3 with Athena, or Google BigTable combined with BigQuery.

RESULT
Pythian's recommendation confirmed the client's hunch that moving its data collection and ingestion process for machine learning to the cloud was the best way to continue operations while improving scalability and cost-effectiveness.

PYTHIAN PROVIDES EXPERT DATA COLLECTION AND INGESTION CONSULTING FOR A MACHINE LEARNING COMPANY FACING END-OF-LIFE CLOUDERA EXPRESS

BUSINESS NEED

A large software-as-a-service (SaaS) company depended on its on-premise, proprietary neural network (advanced machine learning) application in order to continually refine its product. Because the product is AI-based, properly training its underlying algorithms was of the utmost importance to ensure a well-functioning service that could quickly and accurately adapt to the needs of customers.

The client's on-prem Hadoop cluster (running on Cloudera Express) was used to collect and store tens of thousands of online audio files and metadata per day, stored in Hbase and processed using Apache Spark and the Apache Impala SQL query engine. These were then fed into a proprietary neural network (advanced machine learning) application and DataFox with the goal of continually refining and improving the system's understanding and ability to respond to these interactions. However, because their version of Cloudera Express was no longer supported, the company needed to re-evaluate its data preparation and ingestion processes to find the most cost-effective and scalable alternative with the least possible disruption.

SOLUTION

Pythian's experience in data science, machine learning and neural networks – including our Machine Learning Partner Specialization from Google, and wealth of expertise working with other machine learning tools like AWS SageMaker, Apache Spark MLlib, TensorFlow, and Apache MXNet – meant we were well-positioned to provide advice on their best possible options. Pythian advised the client that its best bet to achieve continuity with its machine learning program while improving scalability was to replace its Hadoop cluster with a cloud-native solution such as AWS combined with Athena or Google Cloud Platform and Google BigQuery.

RESULT

Pythian's recommendation confirmed the client's hunch that moving its machine learning data collection and ingestion processes to the cloud was the best way to continue its machine learning operations with the least disruption – ensuring the company's software could continue improving in near-real-time – while also improving scalability and cost-effectiveness by using cloud-native ephemeral tools.

ABOUT PYTHIAN

Pythian excels at helping businesses around the world use data and the cloud to transform how they compete and win in the data economy. From cloud automation to machine learning, Pythian leads the industry with proven innovative technologies and deep data expertise. For more than 20 years Pythian has built its reputation by delivering solutions to the toughest data challenges faster and better than anyone else.

WORLDWIDE OFFICES

Ottawa, Canada
New York City, USA
London, England
Hyderabad, India

Pythian
love your data®