INDUSTRY
Technology

BUSINESS NEED
With an exceedingly large, on-prem HBase database built on a soon-to-be-retired version of Cloudera, the client needed a cost-effective way to seamlessly migrate its database while beefing up its backup and disaster recovery capabilities.

SOLUTION
Pythian expert consultants presented three main cloud options: Hadoop in the cloud as a service, cloud-native options that most closely fit the client's current HBase column store, and cloud-native options centered around an object store and column store.

TECHNOLOGIES
Cloudera; Apache Hadoop, Impala, and HBase; AWS RedShift, DynamoDB and Simple Storage Service (S3); GCP BigTable, BigQuery and Google Cloud Storage; Ceph; ClickHouse.

RESULT
The client acquired cutting-edge insights and hard data on all their on-premises and cloud options from multiple vendors, allowing them to make the most informed decision possible to upgrade their system while dramatically improving backup recovery and cost savings.

PYTHIAN PROVIDES INVALUABLE CONSULTING TO A COMPANY FACING EXPENSIVE ON-PREM HADOOP UPGRADE

BUSINESS NEED

A large, private technology company had for years used an on-premises Hadoop cluster to collect and store tens of thousands of online audio interactions generated by the company's proprietary software, along with the associated metadata, for training purposes. These audio files and their related data were stored in a massive, 215-plus terabyte Apache HBase database, processed with the Apache Impala SQL query engine, and then fed into a proprietary neural network (advanced machine learning) application with the goal of continually refining and improving the system's understanding and ability to respond to these interactions. However, because their HBase database was so large, it didn't have enough capacity to support the necessary snapshots long enough to perform backups. This meant their disaster recovery capabilities were almost nonexistent, with the company facing a recovery process of several weeks if not months (along with possible significant data loss) should a disaster occur.

Because it was all being run on a version of Cloudera CDH that is no longer supported, the company was facing a bill of up to $233,000 per year in license fees and other costs to upgrade to a supported release. With the organization also needing to dramatically beef up its backup and disaster recovery capabilities, they were looking at costs of closer to half a million dollars per year to continue effectively using their on-prem Hadoop cluster.

SOLUTION

Thanks to our familiarity and experience in both supporting Hadoop clusters and migrating them to cloud-native systems, the company turned to Pythian for expert advice on the state of their current system along with suggestions on the best on-premises or public cloud

alternatives for their needs. However, it quickly became apparent that their on-premises options were limited to the expensive upgrade and licensing fee structure mentioned above, or going with an upgraded DIY Hadoop cluster that was likely to be even more expensive after staff, hardware and support costs were factored.

Three main cloud options were discussed during a two-day on-site workshop with company stakeholders and Pythian consultants: Hadoop in the cloud as a service, cloud-native options that most closely fit their current HBase column store, and cloud-native options centered around an object store and column store. Public cloud services from Google Cloud Platform (GCP) and Amazon Web Services (AWS) were considered, with each having the option of either significantly minimizing or completely replacing their current Hadoop cluster:

- **AWS DynamoDB combined with Redshift** was one option to replace their HBase use cases and Impala SQL query engine. This approach, while effective, was likely to be less cost-effective than other approaches.
- **GCP BigTable combined with BigQuery** was also considered, with BigTable performing the same function as DynamoDB and BigQuery doing the querying. Because BigTable uses the same API as HBase, this was the best fit for replicating their HBase use cases with the least disruption, and entailed costs estimated at the high end of the client's budget.
- **A completely new cloud object storage and column store option** was also presented (using either AWS Simple Storage Service (S3) or Google Cloud Storage), in conjunction with either the Athena S3 Query engine or BigQuery which would automatically support multi-region redundancy for backup and disaster recovery, while being the least expensive cloud option.
- **An onsite option was also discussed** using onsite object and column stores like the Ceph object store and Click House column store.

RESULT

The organization's IT leadership had already recognized the potential cost savings of an object store and column store avail over a Hadoop deployment but needed to verify that assumption–along with ensuring it had looked at all possible options on the table, their features, and their possible costs. Thanks to the company's engagement with Pythian consultants, they were able to come away with cutting-edge insights and hard data on all their on-premises and cloud options from multiple vendors, allowing them to make the most informed decision possible to upgrade their system, implement 12-hours-or-less backup and disaster recovery, and drive significant cost savings.

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®