DATA MIGRATION TO THE CLOUD LEADS TO MORE FLEXIBILITY AND CONSISTENCY AT A REDUCED COST

IMPACT

- Utilizing autoscaling to scale up from 150 static nodes to 250-500 nodes as needed
- Runs at 25% of full scale during weekdays, then scaled out 4x for weekends
- Spark jobs now running faster on CDP
- Combined workflows, enabled more flexibility and consistency with data

Challenge

One of the biggest examples of a Cloudera customer undergoing a migration from on-premises to the cloud is this American technology company. The company's core business is listening to advertising networks and member information, including cell network usage, to create clusters of groups to advertise to, and is effectively implemented as a giant ETL engine. The business correlates all the data and creates identity profiles for people and groups them together based on similarities. By doing this, the company can run analytics and gain insights on how to personalize ads, improve sales and marketing efforts, and gain a 360-degree view of its customers, including what products and services they would be most interested in. The data is also used to create graphs for analysts to use to help inform decision making.

In order to do this effectively, the company needed to be able to access, store, and process vast amounts of data in a cost effective manner. Previously, the technology company was running CDH on-premises in two data centers, one in the East and West regions. It became important to combine workflows for both regions, and the company also wanted to adopt the latest and greatest technology to leverage new features.

Solutions

The company decided that it wanted to vacate the data centers and move everything to Cloudera Data Platform (CDP) Public Cloud on Microsoft Azure, utilizing four different production CDP Data Hubs across two different environments. The company leverages a fully-HA medium duty data lake for one of its production environments. Two of the data hubs are for Apache Kafka and the other two run Apache Hive and Apache Spark services. Leveraging the cloud enables the company to easily set up its infrastructure for compute, storage, network and cluster management.

Right now the clusters act as a log aggregator. Apache Hive and Apache Spark are being used for certain workloads, while Apache Kafka is utilized for ingesting web services and logs into HDFS. The advertisement ID's are ingested from both the East and West regions. These are replicated between each other and then copied over, so that everything can be viewed in one place. There is also recipe and template incorporation for the DevOps team to start deploying.

Moving everything to the cloud combines workflows, enables more flexibility and consistency with data. It also meant reducing operational and resource costs through the ability to leverage CDP's autoscaling and automation. Additionally, by leveraging SDX within CDP, the company was able to set granular security and governance controls while migrating to the public cloud.

Results

By undergoing the migration from on-premises to the public cloud, this technology company has been able to take full advantage of the benefits of CDP, including autoscaling, and gain additional automation with more of a PaaS model. This frees up the company's platform management team from having to worry about the day-to-day cluster management.

100

Percentage of data migrated to CDP Public Cloud

About Cloudera

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises.
Learn more at cloudera.com

Cloudera's Professional Services team initially set up the cluster and everything since gets taken care of with CDP, so it's no longer necessary to have someone constantly monitoring the health of the cluster.

All data has been fully migrated from legacy on-premises onto CDP Public Cloud in Azure. Both production environments (East and West regions) are up and running stably. The company's data and analytics demand isn't consistent all the time, so by utilizing CDP's autoscaling feature, the cluster can resize itself based on how many resources it needs for the jobs being done. The company has 150 static nodes and can scale up to 250-500 nodes when workloads burst as needed. From Monday through Friday, the cluster is being run at 25% of its full scale. On weekends it is scaled out four times for main production needs, generating reports and graphs for Monday morning. Spark jobs are also running much faster in CDP than they were running in CDH.

CLOUDERA

Cloudera, Inc. 5470 Great America Pkwy, Santa Clara, CA 95054 USA. cloudera.com