A global leader in financial technology modernizes their data lake in the cloud

About the global leader in payments and financial technology

As a global leader in payments and financial technology, they help clients achieve best–in–class results through a commitment to innovation and excellence in areas including account processing and digital banking solutions; card issuer processing and network services; payments; e–commerce; merchant acquiring and processing.

The organization is focused on

- Connecting data from various systems
- Making it easy to discover relevant insights
- Putting data at your fingertips when you need it

Objectives

Data and analytics are critical to the organizations aspiration, moving money and information to move the world. This includes their on–premises data lake, which consists of two on premise clusters running Cloudera Distributed Hadoop (CDH) 6.3. They have also started to use cloud analytics solutions from Databricks and one of their objectives is to build out a complete new set of analytics use cases and do so in the Azure cloud so they can:

- Leverage cloud elasticity and easily scale the environment as needed
- Utilize advanced analytic tooling available in the cloud (i.e. Databricks etc.)
- Modernize and clean their data architecture by selectively choosing the datasets to transfer to the cloud

"We selected Cirata Data Migrator to transfer data from our on–premises data lake to the cloud. Data Migrator provided superior performance and throughput over the alternatives we evaluated, and the organization delivered excellent support during the initial proof of concept, overall project, and continue to do so today."

— Senior Director Technical Operations, Global leader in payments and financial technology

Challenges

The organizations on–premises data lake includes data that they do not want to transfer, or for regulatory reasons has not been approved by their legal department to move to the cloud. They need to be able to easily select and control the data that is transferred and data that needs to remain on–prem. Since their on–premises data lake is business critical it needs to be available 24x7 for analytics as well as for data ingest and changes that occur daily and cannot afford any system downtime or business disruption. The organization also established throughput requirements that they need the data transfer process to achieve.

In summary, the key challenges and requirements include:

- Ability to easily select and manage what data is transferred
- No production downtime or business disruption
- Meet performance and throughput objectives

Azure Cloud

ADLS Gen2 Storage

cirata Data Migrator

On–premises Cluster 1 (with CDH 6.3)

On–premises Cluster 2 (with CDH 6.3)

Solution

Following a proof–of–concept (PoC) the organization selected Cirata for their on–premises data lake to Azure cloud data transfer process. Data Migrator is an automated, scalable, high performance, and cloud–agnostic data integration solution that simplifies making data available in and immediately usable across on–premises environments and with any cloud platform. The PoC demonstrated that Data Migrator would meet all of their requirements and address their data transfer challenges.

The organization also evaluated alternative solutions such as DistCp (distributed copy) and AZCopy (Microsoft Azure's DistCp based technology). They indicated that they were not able to reach their throughput requirements with AZCopy, and similarly saw a "performance lag" with DistCp. Furthermore, DistCp and AZCopy are designed to copy data based on a single point in time. Any data ingested or changed since the copy process started would not be picked up, and subsequent scans are
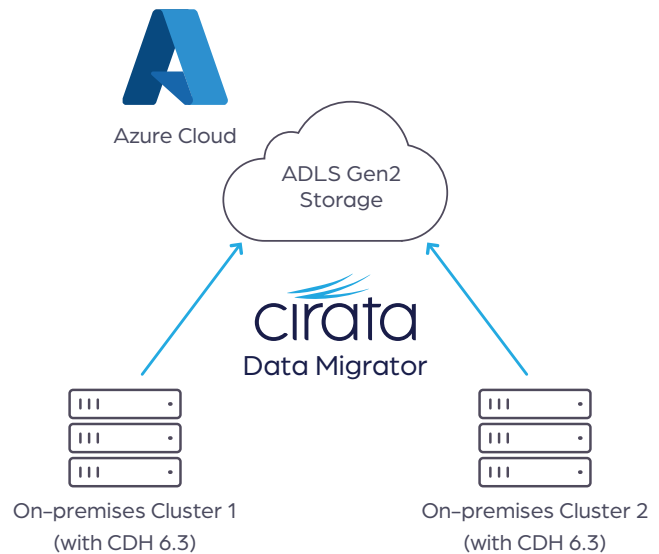
needed to capture ongoing data changes. To prevent this from happening requires the production system to be brought down, which was unacceptable.

Data Migrator performs the initial data transfer using a single scan of the source storage, while also supporting continuous replication of any ongoing changes from source to target with zero disruption to current production systems.

Data Migrator is installed on an edge node of the source cluster, and deployment can be performed in minutes and does not require any custom coding or changes to source applications. The organization was able to easily configure data transfer jobs to meet their specific requirements, such as data sets to transfer, exclusion rules, bandwidth management and more. Verification capabilities ensure all data is transferred, and the product user interface allows for management and to monitor the full data transfer process from a single console.

Results

Data Migrator enabled the organization to:

- Automate the data transfer from two on–premises CDH clusters to the Azure cloud with no production downtime or business disruption. This couldn't be done with all of the alternative methods they reviewed
- Optimize data transfer performance and network bandwidth usage.
- Configure, manage, and monitor data transfer jobs to meet specific business requirements.
- Modernize their data architecture in the Azure cloud enabling cloud elasticity.
- Provide the data to enable Databricks analytics platform for downstream applications
- 1.5 PB of data moved into Azure.

5000 Executive Parkway, Suite 270, San Ramon, CA 94583
US +1 877 (926–3472)   EMEA +44 (0) 114 3039985
APAC +61 2 8211 0620   All other +1 925 380 1728

www.cirata.com

giftmdlc–0924