eBook

Empowering success through data and innovation.

Real stories and real success... Discover how Cirata transformed Data management goals into reality for global industry leaders.

cirata

Contents:

Top 3
US Bank

Transforming financial security: Powering disaster recovery solutions for a top 3 US bank.

About the top 3 US bank

The top 3 US bank is an American multinational investment bank and financial services holding company headquartered in North Carolina. The organization is a global leader in wealth management, corporate and investment banking, and trading across a broad range of asset classes, serving corporations, governments, institutions, and individuals around the world. The company serves clients through operations across the United States, in U.S. territories, and in approximately 35 countries.

The organization is focused on

Disaster recovery

Objective and challenges

The organization uses an on-premises Cloudera CDP cluster for one of their analytics platforms and a secondary CDP environment for Disaster Recovery (DR). The company has established a 15-minute service level agreement (SLA) for the recovery time objective (RTO) and recovery point objective (RPO) of the DR environment. Their current implementation is not able to meet the SLA requirements.

To achieve the SLAs, they need a solution that can replicate and keep synchronized more than one million transactions per minute. This high volume of transactions means that manual or batch-based tools would not be a viable option, and this has opened the opportunity for IBM Big Replicate, which utilizes Cirata LiveData Migrator to support real-time data replication. The initial process is replicating data between two on-premises data centers and there is future opportunities to also replicate the data to a public cloud environment.

Solution

Cirata LiveData Migrator (LDM) supports complete and continuous replication of data sets at any scale. With zero disruption or impact to the existing production system, LDM migrates the initial data sets with a single pass through the source storage, eliminating the overhead of repeated scans while also supporting continuous replication of any ongoing changes as they occur.

The organization reviewed multiple approaches before selecting IBM Big Replicate / LDM. The other approaches that were evaluated included DistCp, Replication Manager (which leverages DistCp), and other open-source tools such as Apache NiFi. None of the alternatives were able to meet their SLA requirements because the alternatives operate in a batch/scheduled fashion and do not replicate ongoing changes in real time. This means the alternatives cannot guarantee that changes made within the last 15 minutes have been successfully replicated to the DR environment.

Other critical factors in selecting IBM Big Replicate / LDM were as follows:

- Ease of use of the solution (simple deployment, configuration, and management).
- Ability to replicate Hadoop data as well as Hive metadata.
- Ability to set up and begin replicating data very quickly.

NatWest

NatWest re-architects their Hadoop data lake for the cloud.

About NatWest

National Westminster Bank trading as NatWest is a major retail and commercial bank in the United Kingdom based in London, England. It is one of the Big Four clearing banks in the UK and has more than 7.5 million personal customers and 850,000 small business accounts.

Objectives

To be able to take advantage of the Sage analytics within the AWS cloud, NatWest planned to migrate their current on-premise Hadoop based data, the Central Customer DNA Database to the Amazon cloud.

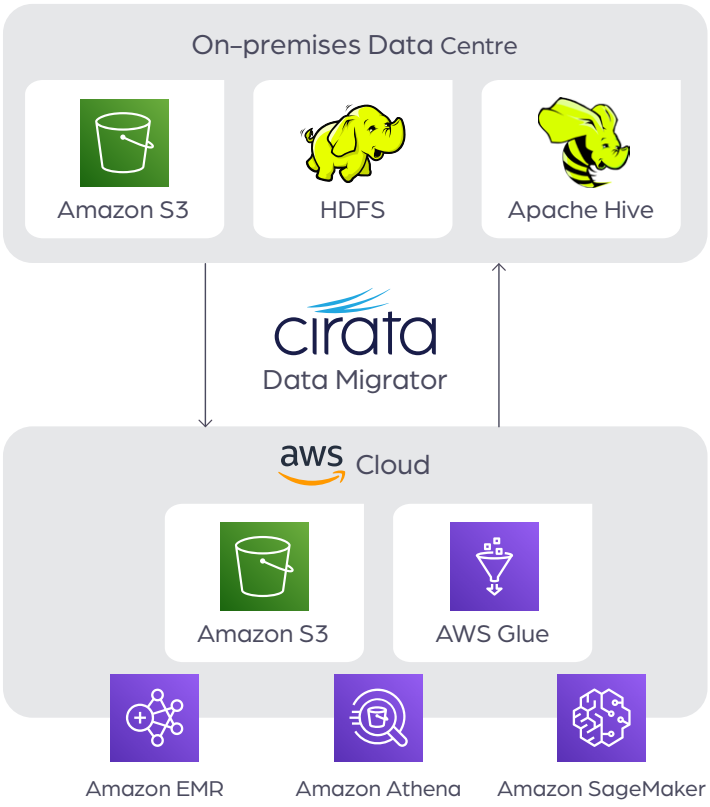
Challenges

NatWest on-premises data lake used HIVE metadata that they wanted to consolidate into the Amazon Glue repository in the cloud. They also needed the ability to move the results of the Amazon analysis back onto on-premises storage to support regulatory reporting applications, that had not yet themselves been adapted for cloud use.

Following a proof-of-concept (PoC) NatWest selected Cirata for their on-premises data lake to AWS 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 NatWest's requirements and address their data transfer challenges.

NatWest are focused on

- Re-Pipeline
- Re-platform
- Redevelop applications



 |

NatWest's original solution for moving data to amazon involved relying on the Cloudera BDR utility and scripted functions in AWS Lambda. BDR uses the Distributed copy functionality of Hadoop to move the data, and this has its own inherent problems.

- It is labor-intensive to manually reconcile differences because of data changes made since the last DistCp run. This results in higher costs and often leads to delayed and failed projects.
- Multiple scans are required to capture ongoing data changes made between DistCp runs. Depending on the size of the dataset and the number of changes occurring, it may be impossible to ever catch up with all changes.
- DistCp runs as a standard MapReduce job competing for resources with other processes and requires you to have open firewalls across all nodes in the cluster, posing security issues.

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.

Results

Data Migrator enabled NatWest to:

- Data flow optimization - Automate and manage the data transfer far easier than their previously script driven solution
- Provided a standard mechanism for moving data whether it be a cloud target or on-premise.
- Future proofed the solution to take advantage of newer analytics and Metadata storage.
- To date 1.4PB of data moved
- Increased the number of data science and innovation lab experiments to develop machine learning models across the bank. This has increased the use of AI models in production across the bank.

Leading global retailer achieves data placement flexibility to satisfy the needs of all information consumers.

Global Retailer

 |

About the global retailer

This leading global retailer is a British multinational groceries and general merchandise retailer headquartered in the UK. They have revenues of over £65 Billion from a portfolio of 3,700 retail outlets and employs over 330,000 individuals.

The organization is focused on

- Data and compute segmentation
- Retail demand forecasting
- Disaster recovery
- Financial analytics
- Data orchestration

Initial objectives

The retailer has been a Cirata customer for over 5 years. The initial use of Cirata technology was to replicate data between their production and disaster recovery (DR) environments. The current solution enables them to support near-zero recovery time objective (RTO) and recovery point objective (RPO) in the case of an issue that results in their production environment failing. To extend the use of Cirata technology, the retailer wanted to segment a subset of their data into a separate on-premises environment to provide a smaller, more relevant and succinct data set to perform analysis and other machine learning and analytical workloads, whilst the larger data set was focused on retail demand forecasting (RDF).

Challenges

Due to the volumes of data gathered through the retail operations of the company, there was a requirement to provide cost effective on-premises storage to manage this data, and to that end the customer chose to use IBM® Storage Ceph. This software-defined storage platform delivers a cloud-like experience while retaining the cost benefits and data sovereignty advantages of on-premises IT. To take advantage of this environment, the customer needed to be able to populate this new storage platform from their existing Hadoop clusters whilst still conducting their day-to-day business.

Solution

The Ceph storage is Amazon Simple Storage Service (S3) compatible, and Cirata Data Migrator (CDM) moves the HDFS data into this storage platform. CDM provided the flexibility to configure migrations on a granular level to meet the retailer's specific needs for this new environment. Once the migration is started, CDM transfers the initial data sets with a single pass through the source storage, eliminating the overhead of repeated scans while also supporting continuous replication of any ongoing changes to the defined data sets. This automated process ensured the data is kept up to date, so the retailer does not need to manually refresh the new environment. This approach eliminates the risk of data loss, costly downtime and repeated attempts of data copy.

Outcome

Data Migrator enabled the organization to:

- Continuously move data to where and when it is needed.
- Provide appropriate compute resources to the many different types of users whilst sharing the same data lake.
- Provided a standard mechanism for moving data whether it be a cloud target or on-premises.
- Future proofed the solution to take advantage of newer analytics and Metadata storage.
- To date 4PB of data moved.

The future roadmap

The retailer continues to separate their compute away from their storage infrastructure to provide a more agile analytics capability to the business. There is also a longer-term plan to migrate fully from the legacy Hadoop infrastructure to the Ceph based architecture and Cirata technology has a large part to play in this transition.

For ultimate flexibility they intend to adopt an open data format. The first project to take advantage of this will be a new finance data hub specifically set up for the analytical requirements of the Finance business unit. The finance data hub will hold ERP data augmented with reference data from their existing clusters. This is a critical data set providing valuable information to the senior leadership team.

The open data platform will be based on the Iceberg format which is quickly becoming the de-facto standard for data storage as it allows a much wider variety of analytics and AI tools to access data without any additional data conversion. Cirata is working with the customer to develop the functionality to support their iceberg requirements.

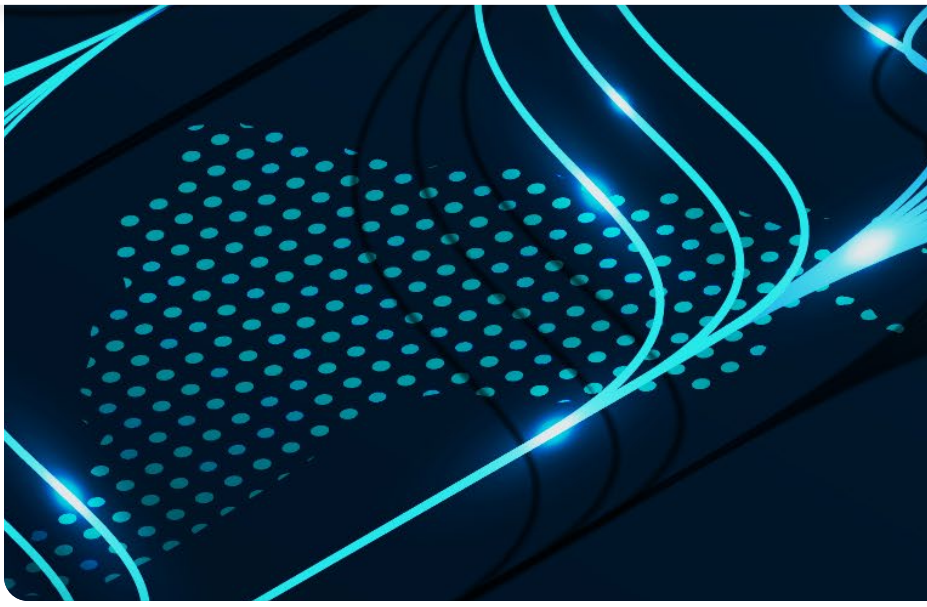
To support the future roadmap, Cirata is working with the customer to provide a ubiquitous data movement solution to satisfy all their data movement needs, the intent is to replace other legacy technologies to establish a data orchestration center of excellence whilst providing federated autonomy across the business.

This orchestration platform will deliver:

- A single skill set for all data movement needs.
- Iceberg to Iceberg data movement.
- Live data migration for Ceph storage.
- Greater flexibility in backup, disaster recovery and test data strategies.
- Observability across all data movement.

Westpac

Westpac begin to exploit their digital and hybrid multi–cloud strategy.



About Westpac

Westpac is Australia's first bank and oldest company, one of four major banking organizations in Australia and one of the largest banks in New Zealand. The bank is halfway through their 5 year program to modernize its technology environment and expand use of cloud-based systems. To achieve this Westpac, have partnerships with both Microsoft Azure and Accenture.

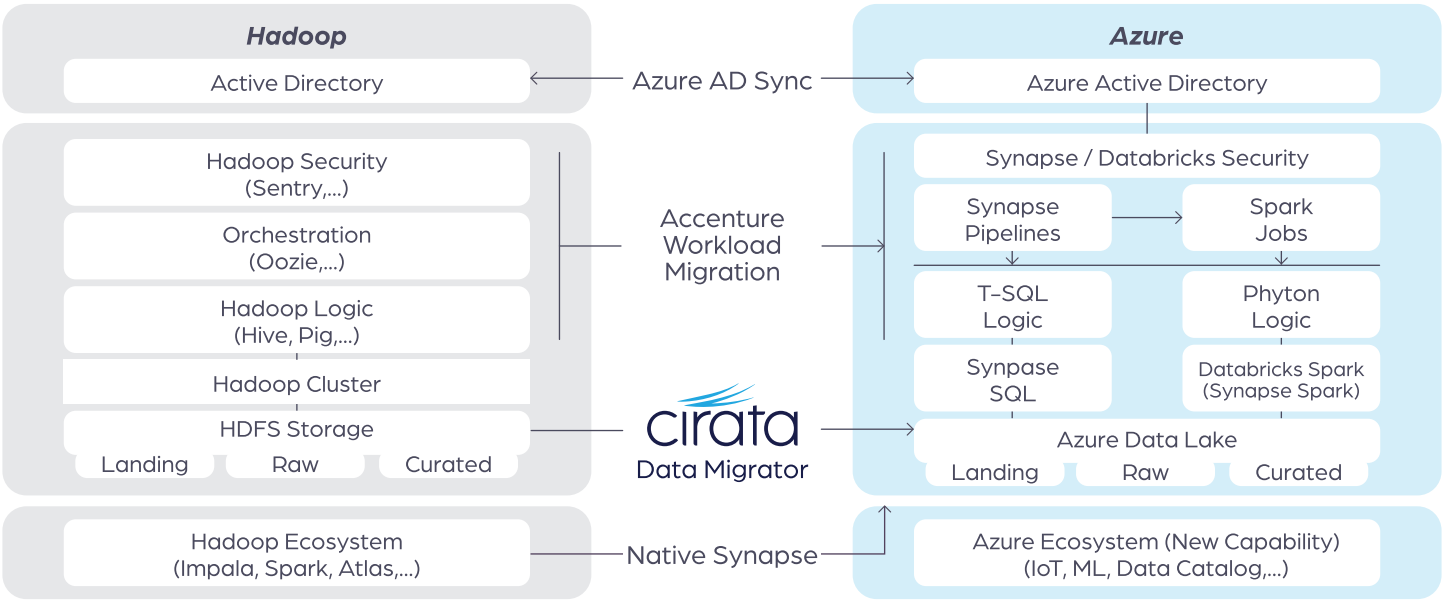
Westpac are focused on

- Reduce technology duplication
- Making it easy to discover customer insights
- Provide unique customer interactions

Objectives

To embellish the banks capabilities in customer insight, the bank wanted to utilize the analytics capabilities of Azure Synapse. To create a quick win for this business case Accenture were commissioned to help provide an ingestion process to move data currently stored on the banks Hadoop data lake environment into Azure data lake storage. With a significant investment in Microsoft , Westpac needed to show some early returns on that investment , the key drivers of the project where:

- Provide early access to Microsoft Analytics tools
- Introduce the new capabilities without interrupting existing business processes.



Solution

Cirata data migrator was used to successfully move the Hadoop data into the Azure Data Lake, and once migrated also move this data to other domains within the Azure infrastructure. The data migration was a live migration in that any updates that occurred to the Hadoop cluster during data migration were also captured as part of the migration process, ensuring that the cloud version of the data was in sync with the original.

Results

- Automate the data transfer from the on-premise Hadoop cluster to Microsoft Azure
- Optimize data transfer performance and network bandwidth usage.
- Provide the data to enable upstream applications access to the data in the cloud. This required Data movement between ADLS domains
- 2 PB of data moved to the cloud

Enabling a global telecom leader to optimize the use of their greatest asset: their data.

Global Telecommunication organization

Global Telecommunication organization

Objective

A global telecommunications leader started their journey to cloud computing by adopting Microsoft Azure. The company wanted to move away from costly on-premises data centers and into the cloud, while at the same time taking advantage of cloud-scale analytics, where possible.

Challenge

The company's cloud program was initially very application-focused, and the data science team quickly realized the cloud migration needed to account for — and leverage — the value of their data being available in the cloud. The challenge centered around migrating tens of petabytes of live production data without disrupting mission-critical business applications and operations.

Solution

The company adopted a data-first approach and used Cirata Data Migrator to move over 20 petabytes of data without requiring any business disruption. The data was immediately available for data scientists to use Azure services for AI and machine learning. Data teams went from blocking nearly 138 million robocalls per month to blocking over 1 billion robocalls per month, as a result of using cloud-scale analytics.

Business Impact

- Original data migration timeline objective was cut by over 50%
- Data-first approach enabled data scientists to immediately begin new AI & ML development
- Early development of high-value AI models enabled enhanced fraud detection
- Able to block 1 billion robocalls per month, over 7.2 times more per year than before
- No disruption to existing production environment during entire migration process
- Able to save millions of dollars by decommissioning on-premises disaster recovery data center

a great leap forward

50%

faster data migration

Customer Profile

This case study is based on a diversified global leader in telecommunications, media and entertainment, and technology.

Objectives

The company had started their journey to cloud computing by adopting Microsoft Azure. They wanted to move to the cloud for several reasons:

- To reduce costs and gain operational efficiency
- To gain the benefits of cloud elasticity
- To modernize and transform their technology environment

Challenges

Initially, the company's cloud program was very application-focused. This is a large company with thousands of applications, and as they started to look at their application journey, one of their big realizations was that it wasn't so much about the applications as it was about the data. This was especially true for the data science office, so the organization started to think about a "datafirst" approach and finding a way to get the data to the cloud to start getting benefits early. They wanted to be able to quickly demonstrate the business benefits of accelerating their business and improving customer experiences, revenue, and customer retention, as well as achieving cost savings.

The challenge was regarding the volume of existing data (tens of petabytes) and the fact that new data was continuously being ingested. How could the company migrate this volume of data efficiently and without impacting current business operations?

The company looked at a variety of solutions, including data transfer devices, ETL tools, and open source tools such as DistCp based software. Each of these solutions weren't fit for purpose. With data transfer devices, the data needs to be copied onto multiple devices, the devices must be transported by truck to the cloud data center, and the data then must be copied from the devices onto the cloud storage. After realizing multiple trucks would be needed, the security issues associated with shipping the data via trucks became clear. They would either need to bring their production systems down to prevent changes from occurring during the transfer process or undertake custom code development to identify and migrate any new or changed data. They quickly discarded transfer devices as a viable approach. Using ETL or open-source software tools also had similar issues in handling changing data, and the company estimated it would be too costly to develop and maintain the custom solutions they would need to develop with those tools. A better alternative was required.

> "We cut our entire cloud data migration timeline for moving 13 petabytes in half."
>
> *Vice President of Data and Analytics, Global Telecommunications Company*

Solution

The company ended up selecting Cirata Data Migrator to enable the data-first approach they were looking for and to automate the Hadoop-to-Azure data migration process without requiring any business disruption.

Initially, a short production pilot was conducted using Data Migrator to transfer 100 TB of Hadoop data directly from the company's on-premises production environment to ADLS Gen2 storage. The pilot was performed over a weekend without the need for any custom development and without any impact to their production systems. And, the data was immediately available for use by Azure services.

The pilot was very successful. It showed the company that they could achieve their data-first strategy with Data Migrator. To perform the migration even faster, they decided to put in an order for additional network bandwidth. However, nothing prevented them from proceeding with the current available bandwidth. Their goal was to migrate about 1 PB per month and get the initial set of data migrated from their on-premises Hadoop cluster into Azure within 12 months.

Results

As they began the migration, they found that Data Migrator could enable them to achieve their 1 PB goal even faster than a month. When the additional network bandwidth arrived, the migration was already ahead of schedule. Data Migrator was able to leverage the additional bandwidth that was now available, and the overall timeline for migrating the initial 13 petabytes of data was cut by over 50%—and would have been cut by even more over the other legacy approaches. The company has since migrated over 22 PB in less than one year and it continues to use Data Migrator in a hybrid environment to transfer ongoing data changes from on-premises to the cloud. The company's Vice President of Data Analytics noted that "We cut our entire cloud data migration timeline for moving 13 petabytes in half."

By following a data-first approach, data became immediately available to the data scientists to not only begin modernizing existing workloads, but to start building new AI and machine learning models that were able to provide new business benefits very quickly. One area they were able to quickly enhance was with regard to fraud detection. Leveraging the data, cloud elasticity, and new AI capabilities, they are now able to quickly achieve savings by identifying robocalls in seconds where it previously took them days, resulting in 7.2 times more blocked robocalls per year than before.

The migration was achieved without requiring any production system downtime or disruption to their business. The existing production environment remained in use during the entire process, and any ongoing changes were replicated to the new cloud environment as they occurred. The company was able to avoid any big bang approaches where all workloads are moved at the same time, and the migration strategy provided the organization with time to modernize and optimize the workloads for the new cloud environment. This allowed the company to make sure applications ran optimally and to take advantage of new capabilities available to them.

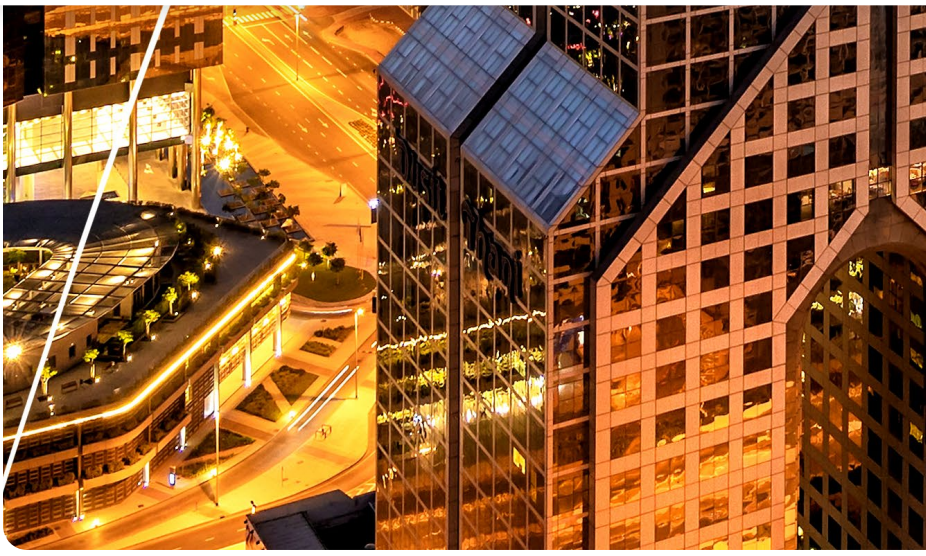
The company was also able to achieve their objective of reducing costs and gaining operational efficiency by decommissioning their on-premises disaster recovery center, since the data residing in the cloud could now be used for this purpose. This decommissioning alone more than paid for the migration project.

Overall, the company was able to achieve a very fast ROI on the project by leveraging a data-first approach.

Results Summary

- Original data migration timeline objective was cut by over 50%
- Data-first approach with Data Migrator enabled data to become immediately available to data scientists
- Faster development of high-value AI models enabled enhanced fraud detection
  - Now able to identify robocalls in seconds, where it previously took days
  - More than 7.2 times more blocked robocalls per year
- Existing production environment remained in use during entire process (no business disruption)
- Able to take time to optimize existing workloads for new cloud environment and avoid big bang cutover
- Able to save millions of dollars by decommissioning the on-premises disaster recovery data center
- Achieved very fast ROI on the project by leveraging a datafirst approach

du ensures their data lake remains resilient in supporting the Dubai Smart City Initiative.



Company overview

Emirates Integrated Telecommunications Company P.J.S.C., commercially rebranded as du in February 2007, is one of the two main telecom operators in the United Arab Emirates. du offers fixed line, mobile telephony, internet, and digital television services across the UAE. It also provides carrier services, a data hub, internet exchange facilities and satellite service for broadcasters.

du is focused on

Provisioning for disaster recovery

Objectives

As a major service provider to Dubai's smart city initiative, du needed to ensure that their on-premise data lake was always available to the Smart Dubai Platform. This platform facilitates the exchange of open and shared data between the public sector, private sector and individuals, creating new opportunities for engagement, contributing to the smart city future of Dubai, and therefore data availability was key. du had tried using other data replication capabilities such as Apache NiFi but the manual intervention required in coding data migration and the time spent repairing data through erroneous data transfer proved prohibitive.

Solution

du has successfully implemented Cirata Data migration technology to effectively support the disaster recovery needs of their expansive data lake. This advanced technology offers an automated platform for data replication that is inherently resilient to network outages, ensuring uninterrupted operations. By requiring minimal manual intervention, it streamlines processes and reduces the likelihood of human error. Additionally, it guarantees that any live data changes are captured in real time at the disaster recovery site, thereby enhancing data integrity and reliability. Over the past 8 years, du has relied on this robust technology to support its 54 node Hadoop environment, ensuring a remarkable 24/7 availability. This consistent uptime is crucial for handling their extensive data processing needs and maintaining operational efficiency, ultimately leading to improved decision-making and business continuity challenging given the growth in the amount of data being ingested daily into the production data lake.

Results

- du successfully supports Dubai's Smart City initiative by ensuring consistent data availability.
- Implemented resilient and efficient disaster recovery solutions for their on-premise data lake.
- Automated data replication using Cirata technology, negating the need for manual intervention.
- Proved resiliency to network outages to assure consistent operation.
- Enabled live data change capture at the disaster recovery site which in turn assured data integrity.
- Has effectively utilized this technology to support its 54 node Hadoop environment for the last 8 years, ensuring 24/7 availability.

Manulife

Manulife repositions its enterprise data lake strategy with Microsoft Azure.

Company overview

Manulife Financial Corporation is a Canadian multinational insurance company and financial services provider headquartered in Toronto, Ontario. The company operates in Canada and Asia as “Manulife” and in the United States primarily through its John Hancock Financial Division. Manulife is the largest insurance company in Canada and the 28th largest fund manager in the world based on worldwide institutional assets under management.

Manulife is focused on:

- Re-Pipeline
- Re-platform
- Improved Analytics

Challenges

Manulife needed to deploy 9 local instances across 3 Azure regions to support regional analytics for Japan, Hong Kong, China & Singapore whilst ensuring no business interruption in all geographies.

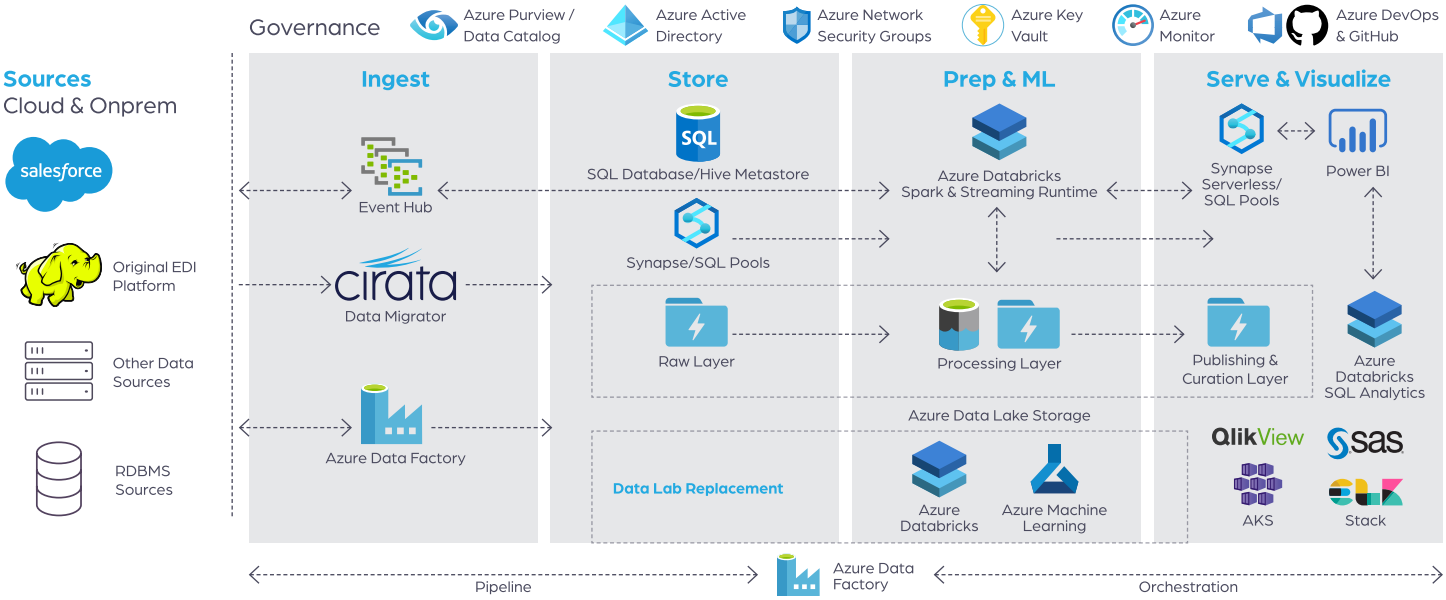
Solution

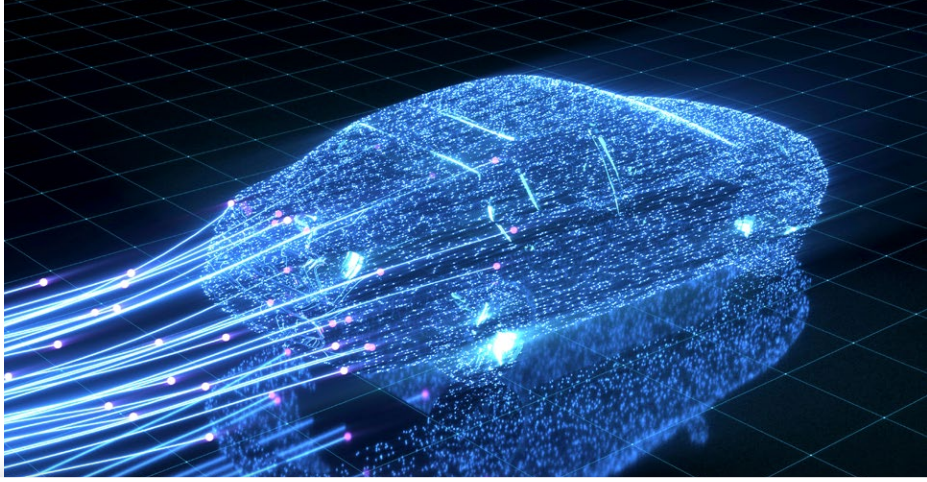
Following a proof-of-concept (PoC) at Manulife’s headquarters in Canada, Cirata’s Data migration technology was chosen to support the Asia EDL migration. The PoC demonstrated that Data Migrator would meet all of Manulife’s requirements and address their data transfer challenges.

Results

Data Migrator enabled Manulife to:

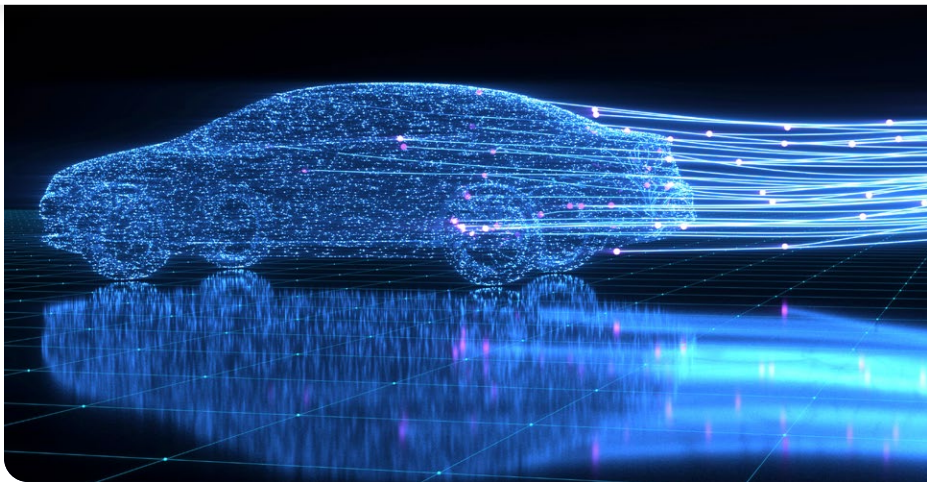
- Migrate 300 TB of data across the geographies.
- Provided rapid access to the Azure analytics platform for earlier insight gain.
- Reduced expenditure on Hadoop infrastructure.





A global automotive company enhances lead generation and AI optimization through Hadoop data migration to Databricks.

Global Automotive Organization



About the global automotive company.

Headquartered in the northern mid-eastern United States, they supply digital marketing and advertising solutions exclusively for the main automotive organizations dealers and retailers. The organization has a mission of helping the dealerships sell more cars and trucks, and have already begun to take advantage of AI capabilities in the cloud and want to continue improving lead generation in this area.

The organization is focused on

- Reducing technology debt
- Provide more accurate insights and buying signals
- Capitalize on their investment in Databricks

Objectives

The global automotive company wanted to embellish their use of cloud analytics by feeding additional customer data into their customer journey platform. This platform was hosted on the Microsoft Azure cloud and utilized Databricks analytics to provide AI powered optimizations, recommendations and direct buying signals to the dealerships and product partners.

Solution

Following a successful proof-of-concept (PoC) they selected Cirata Data Migrator to move their Hadoop customer data to the Azure cloud. This consisted of moving 1500 tables over 30 databases in the parquet file format.

They also wanted the data to remain in parquet format in the Azure cloud to enable other upstream applications access. For the customer journey platform, Databricks notebooks were used to create the required Delta format structures for Databricks analytics.

Results

- All tables migrated successfully to Azure with no business interruption.
- Reduction in the on-premise Hadoop footprint.
- Increase in Lead quality – The highest scored leads are now 18 times more likely to purchase a vehicle.
- The organization is exploring other business areas that can utilize Cirata Data Migrator to move useful data into their cloud infrastructure.

Global Financial Services Organization

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

cirata

Global Financial Services Organization

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

About:

Trusted by global brands and industry leaders for more than 15 years, Cirata specializes in the migration of Hadoop data lakes into leading cloud platforms to enable game-changing Artificial Intelligence ("AI") and analytics. With Cirata, data leaders can leverage the power of AI and analytics across their entire enterprise data estate to freely choose analytics technologies, avoid vendor, platform, or cloud lock-in while making AI and analytics faster, cheaper, and more flexible. Cirata's portfolio of products and technology solutions make strategic adoption of modern data analytics efficient, automated, and risk-free. In addition, leveraging our patented technologies, including the Distributed Coordination Engine ("DConE®"), our DevOps solutions integrate effortlessly with your existing source code management to increase security, minimize risk, reduce latency, and improve collaboration across globally distributed development teams.

For more information on Cirata visit www.cirata.com

For more information on Data management Solutions www.cirata.com/data-management

6101 Bollinger Canyon Road, Suite 361, San Ramon, CA 94583
US +1 925 380 1728
EMEA +44 (0) 114 3039985
APAC +61 2 8211 0620
All other +1 925 380 1728
www.cirata.com

