NatWest re-architect 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

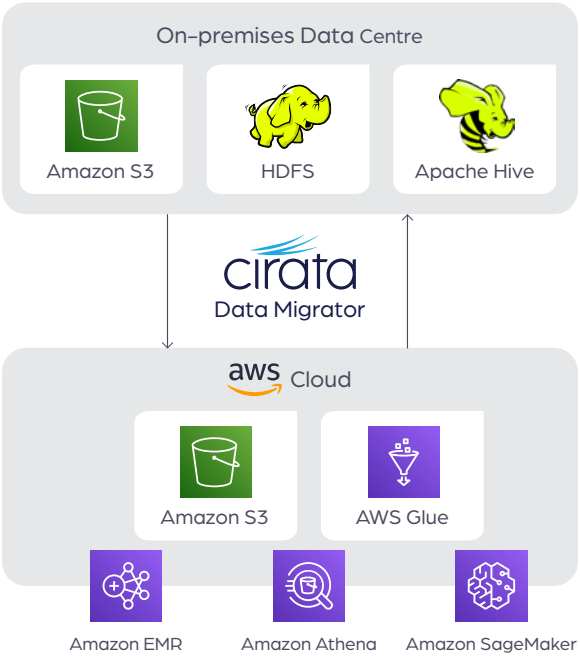
Natwest on-premises data lake used HIVE metadata that they wanted to consolidate into the Amazon Glue repository in the cloud. They also needed the ability to move the results of the Amazon analysis back onto on-premises storage to support regulatory reporting applications, that had not yet themselves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

NatWest are focused on

- Re-Pipeline
- Re-platform
- Redevelop applications



- 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.