Customer case studies

"Cirata's ability to move petabytes of data without interrupting production and without risk of losing the data midflight is something no other vendor does."

– Merv Adrian, Gartner Research Vice President of Data and Analytics

Table of contents

Use Cases
Sanlam
GoDaddy
Kobic
AMD
Daimler
Envestnet I Yodlee
HM Health Solutions
University of Sheffield

Business success now depends on how effectively organizations utilize their data. To do so companies are modernizing their data architecture to both cut costs and to get to better business insights powered by machine learning, which means bringing petabytes of constantly changing data to work in the cloud, and to ensure data consistency across multiple distributed environments where that data is used.

In order to minimize the risks and costs associated with these modernization and digital transformation efforts, many companies have selected to automate migration and replication with Cirata Data Activation Platform. Cirata ensures that critical data is continuously available in any environment, in every location, and at any scale — even when data is changing.

In this Customer Success eBook you will find a number of examples of customers that have successfully leveraged Cirata for their data migration, modernization and digital transformation initiates.

Use cases

Cloud data migration

Fully automated data migration with zero disruption allows your users and systems to continue operating while migration is underway. Free your systems from capital constraints when acquiring storage hardware by leveraging cloud storage systems, including Amazon S3, Azure Data Lake Storage, Google Cloud Storage, and others.

Hybrid cloud

Organizations can maintain critical on-premises applications in production for as long as necessary, while expanding their investment and innovation in the cloud.

Multi cloud

Set the foundation for a multi cloud strategy where your data remains accessible and active across multiple cloud environments. Experience the freedom of choice to move data to any public cloud and avoid vendor lock-in.

Cloud analytics

Bring your on-premises workloads to the analytic tooling available in the cloud faster. Use systems like Databricks, AWS EMR and Athena, Azure HDInsight and Synapse, Snowflake and others without affecting your on-premises activities.

Disaster recovery/backup

Bring your data to the cloud and leverage cloud platforms for your backup and Disaster Recovery (DR) requirements. Benefit from cloud computing platforms' use of automated virtual platforms to minimize recovery times, pay-as-you-go pricing models to lower costs and eliminate the need for on-premises DR environments that typically just sit idle.

Cloud bursting

Cloud bursting is a configuration that's set up between a private cloud (on-premises environment) and a public cloud to deal with peaks in IT demand. If an organization using a private cloud reaches 100 percent of its resource capacity, the overflow traffic is directed to the public cloud so there's no interruption of services.

Sanlam

Company overview

Sanlam was established as a life insurance company in South Africa but has since transformed into a diversified financial services group operating across Africa, India, and selected emerging and developed markets, with listings on the Johannesburg, A2X, and Namibian stock exchanges.

Sanlam has been creating value for stakeholders since 1918 — for more than 100 years. Digital capabilities are critical to transforming their business for client-centric growth. The Group focuses on innovation across their products and services, distribution channels, and back-office processes. Sanlam is on a journey to digital business transformation by digitally optimizing their traditional business to a more efficient tech enabled provider, and investing in new capabilities and assets to build a disruptive 3.0 platform.

"We were impressed with Cirata's unique capabilities for real-time replication without disrupting our production environments. It was critical to us that data in our disaster recovery cluster is kept in sync with our primary cluster as closely as possible to enable a near-zero recovery point objective and recovery time objective. Cirata and its partners provided Sanlam with great support during the product evaluation and initial project deployment. They addressed our questions and concerns in an efficient and professional manner and delivered results within the timeline they promised."

– Jacques Joubert, Big Data Manager and Architect, Sanlam

Sanlam

Challenge

Meeting the high availability objectives that Sanlam established for their CDP implementation meant that data in the DR environment needed to be kept consistent with that in production. Data is continuously being changed in production — either new data ingested, or existing data updated — so Sanlam required a solution that could replicate the data changes as they occurred in production to the DR environment.

Sanlam investigated several tools, such as Cloudera Replication Manager and DistCp (Distributed Copy) — an open-source tool provided with their Hadoop distribution. However, these tools do not support replication of changes as they occur. Instead, users must create schedules to periodically replicate data incrementally.

In the event of a failure, any changes made since the prior replication run will be lost, impacting both the RPO and RTO. Furthermore, these DistCp-based solutions run as standard MapReduce jobs competing for resources with other processes, which can impact production system performance.

Sanlam needed a solution that could replicate the actively-changing production data in near-real time without impacting performance so they could meet their high availability objectives.

Solution

Following a thorough technical evaluation, Sanlam selected Cirata Data Migrator to automate the migration and ongoing replication of data between their primary and secondary (DR) environments.

Data Migrator is a safe and reliable cloud migration solution that automates the migration and replication of Hadoop data and Hive metadata to the cloud or between data centers. Data Migrator deployment is performed in minutes and requires no changes to applications or business operations. Migrations and replication of any scale can begin immediately and be performed while the source data is under active change, without requiring any production system downtime or business disruption. Data Migrator is the ideal solution for Sanlam and enables them to support near-zero recovery time objective (RTO) and recovery point objective (RPO), which were critical objectives for their business intelligence platforms.

Results

- Original data migration of 70TB of data performed with no business disruption
- Ongoing replication of changes to production data as they occur (20TB per week)
- No disruption or impact to existing production environment
- Established DR environment that supports near-zero RTO and RPO
- DR environment provides additional capacity for more analytics, AI, and ML processing
- Data Migrator enables Sanlam to easily move data to other sources as well, such as for their future cloud requirements

Company overview

GoDaddy Inc. is an American publicly traded Internet domain registrar and web hosting company headquartered in Scottsdale, Arizona and incorporated in Delaware. As of August 2020, GoDaddy has approximately 20 million customers and over 7,000 employees worldwide. GoDaddy is empowering everyday entrepreneurs around the world by providing all of the help and tools needed to succeed online. With 20 million customers worldwide, GoDaddy is the place people go to name their idea, build a professional website, attract customers and manage their work. GoDaddy's mission is to give their customers the tools, insights and the people to transform their ideas and personal initiatives into success.

"At GoDaddy, deep technical knowledge is in our DNA, and we often build applications in-house to support growth. In the use case of a Hadoop to Amazon S3 data migration and replication, we found Cirata's Data Migrator to be the optimal approach to deliver the best time to value, rather than running a more time-consuming and costly manual migration project internally."

– Wayne Peacock, Chief Data and Analytics Officer, Godaddy

Challenge

GoDaddy utilizes an 800-node Apache Hadoop cluster to hold over 2.5 petabytes of customer-related activity and behavior data. This on-premises data lake is critical for guiding business operations and determining the company's investment strategies. The system is in operation 24x7. It can generate peak loads of more than 100,000 file system events per second, with sustained 12 hour periods processing an average of over 21,000 change operations every second.

The challenge for GoDaddy was how to migrate petabytes of actively changing, "live" data when the business depends on the continued operation of applications in the cluster and access to its data. Any disruption to business operations would be unacceptable and may have prevented a migration from even being attempted.

Solution

GoDaddy used Cirata's Data Migrator to migrate data from their actively used cluster to AWS S3. Data Migrator performs a single scan of the source datasets and processes the ongoing changes that occur to achieve a complete and continuous data migration. It does not impose any cluster downtime or disruption to production applications and requires no changes to cluster operation or application behavior. Data Migrator enabled GoDaddy to perform their migration without disrupting business operation, and ensured that datasets were transferred completely, even while under active change in a very large and busy Hadoop environment.

Results

- Using Data Migrator, GoDaddy achieved their initial migration goal—to migrate 500TB (over 8.6 million files) of the 2.5PB to AWS S3
- Completed the migration process while maintaining normal business operations at all times
- Reduced cost and risk of custom data migration development, enabling engineers to focus on other business-critical tasks
- Established a new environment using AWS where GoDaddy plans to leverage AWS S3, EMR, Athena and other AWS services to achieve the following:
  - Lower risk by moving off current aging hardware
  - Meet SLAs for critical ETL processing requirements
- Create a better experience for their users through faster queries
- Greater agility by putting more data and flexible compute in the hands of data consumers
- Improved operational efficiency by alleviating the burden of managing the large and complex on-premises hardware and software infrastructure

Company overview

The Korean Bioinformation Center (KOBIC) is the Korean national research center in bioinformatics, based in Daejeon, South Korea. KOBIC manages biological data from a number of different sources, with an emphasis on omics data. Research at KOBIC has an emphasis on next-generation sequencing methods, systems bioinformatics, biomedical informatics and structural informatics. In addition, they operate Bio-Express, a large-scale Genomic Data Analysis Cloud service, which is provided for free to bio-engineering researchers at hospitals, businesses, universities and research institutes in Korea.

In March 2020 KOBIC opened the COVID-19 research information portal (kobic.re.kr/covid19) to provide researchers with information by collecting COVID-19 related genomes and proteomic data scattered around the world for COVID-19 research.

"KOBIC uses Cirata to automate file transfer 13 times faster in both directions between Hadoop-based Big Data Analysis Program Execution Cluster (HDFS) and Linux-based Genomic Analysis Program Execution Cluster (Lustre). We were able to reduce the overall average time to analyze user genomic data of Bio-Express service by more than 30%."

– Kun-Hwan Ko, Researcher at Kobic's Computational Development team

Objective

The Korean Bioinformation Center (KOBIC) collects and manages bio-research resource information and genomic information. In 2020 they began to collect COVID-19 genome data produced worldwide. KOBIC also operates Bio-Express, a large-scale Genomic Data Analysis Cloud service, which they provide for free to bio-engineering researchers at hospitals, businesses, universities and research institutes in Korea.

Challenge

Bio-Express has high-performance data analysis requirements, but the time needed to replicate the large volume of data between their Hadoop Distributed File System (HDFS) and their Linux/Unix based Lustre file system was lengthy and impacts Bio-Express response times. KOBIC needed a way to reduce the time required to replicate data and improve the overall system performance.

Solution

In early 2020 Cirata announced free access to their suite of cloud migration and big data tools, for teams involved in developing potential treatments and cures for the COVID-19 pandemic. Cirata provided their software along with technical resources free of charge to KOBIC to assist the organization in enhancing its architecture, developing products, and introducing Cirata's automated replication technology into KOBIC's workflow.

Results

Leveraging Cirata the Bio-Express next generation service was able to:

- Replicate files 13 times faster
- Shorten average analysis time of Bio-Express services by greater than 30%
- Provide users with faster response times and the ability to perform their research with greater efficiency

> "Cirata provides AMD with consistent data in real time across our cloud and on-premises solutions, offering near-zero RPO and enabling hybrid cloud agility to drive the business forward."
>
> *– Ajay Prasad, Amd Big Data Leader*

Company overview

For more than 50 years AMD has driven innovation in high-performance computing, graphics and visualization technologies — the building blocks for gaming, immersive platforms and the datacenter. Hundreds of millions of consumers, leading Fortune 500 businesses and cutting-edge scientific research facilities around the world rely on AMD technology daily to improve how they live, work and play. AMD employees around the world are focused on building great products that push the boundaries of what is possible.

Challenge

In the age of cloud-enabled global business, it's easy to forget that even the most innovative companies still depend on systems based in a single location. If disaster strikes a data center, it can have a serious impact.

AMD faced an extreme version of this problem: the data center that supports its manufacturing processes is located in a tornado area, and any downtime threatened significant disruption to its finely tuned operations. In preparation for tornado season, the company looked for a geo-redundancy solution to help it to keep its systems online and reduce the risk to its manufacturing processes.

To increase geo-redundancy, AMD had begun to pursue a hybrid cloud strategy, which included Azure Data Lake Storage and Azure HDInsight. However, AMD's use of these solutions was limited, as taking production systems offline to transfer data to the cloud was not an option. With tornado season approaching, AMD looked for a way to replicate its critical data to the Azure cloud without interrupting day-to-day business operations.

Solution

As a first step, the company deployed Cirata to synchronize its critical business data to an Azure Data Box. Cirata replicated the data without requiring system downtime, enabling AMD to operate its business systems as usual during the process. AMD then shipped the Data Box to Microsoft to upload to the Azure cloud, and re-engaged Cirata to automatically synchronize data that had changed during the shipping process—again, without impacting production systems.

Cirata now replicates data automatically to the Azure environment, ensuring complete consistency between the on-premises data center and the Azure cloud systems. If the data center is unavailable because of planned or unplanned downtime, service can automatically continue from the Azure environment, without interruption. And because Cirata manages replication continuously, applications can access current data regardless of location, enabling seamless business operations.

Results

- Protects vital business and manufacturing processes from disruption in the event of a disaster at AMD's production data center
- Ensures no data will be lost in the event of a switchover from the on-premises site to the disaster recovery environment.
- Enables future transition to a hybrid cloud architecture for increased operational cost-efficiency.

> "The on-premises solution would have been nothing more than a dead end for us, as it would have been too inflexible as well as very expensive."
>
> *–Guido Vetter, Head of Corporate Center of Excellence Advanced Analytics and Big Data, Daimler AG*

Company overview

Daimler AG is one of the world's most successful automotive companies. With its Mercedes-Benz Cars & Vans, Daimler Trucks & Buses and Daimler Mobility divisions, the Group is one of the leading global suppliers of premium cars and one of the world's largest manufacturers of commercial vehicles.

Daimler Mobility offers financing, leasing, fleet management, investments, credit card and insurance brokerage as well as innovative mobility services.

Challenge

Over the years, as data was transforming their entire business, Daimler invested in its own on-premises data centers and big data platforms. However, their petabyte scale on-premises data lake soon became too expensive and too inflexible. The company needed to handle its data at a global scale, and it was not their core competency to run these big environments. Instead they decided to bring all of the company's activities around advanced analytics, big data and AI into the Azure cloud.

The challenge was how to migrate their large scale on-premises HDFS data and Hive metadata into Azure, and do so without blocking operations in their current production environment. They also needed to ensure zero data loss, and to maintain 100% data consistency between environments even as data continued to change in their on-premises environment.

Solution

Microsoft Gold partner, Cirata was selected for the migration. Cirata provided the only solution that could perform the migration without business disruption and ensure that all data was accurately replicated.

Cirata also enabled the migration link between Daimler's on-premises environment and the Azure cloud to be throttled so as to allow the connection to be shared. These two features (guaranteed data consistency even when data is changing, and bandwidth throttling) were critical in enabling Daimler to decommission their on-premises environment in a timely fashion.

Results

- Enabled Daimler's expansion to the cloud by facilitating seamless migration of 110 TB of critical data without downtime or disruption
- Guaranteed data consistency even with changing data
- Bandwidth throttling allowing the connection to be shared
- Azure cloud providing 5x storage capacity at 30% lower cost
- Azure's advanced AI services are now at their disposal
- Ability to scale development globally

> "Cirata was the only solution that made this possible when we upgraded to a new version of our Cloudera production system. Cirata works flawlessly on Cloudera. It was a very stress-free process."
>
> *– Jeff Schulte, Vice President of Operations at Envestnet | Yodlee*

Company overview

Envestnet | Yodlee is the leading data aggregation and data analytics platform powering dynamic, cloud-based innovation for digital financial services. The Envestnet | Yodlee platform has proudly fueled innovation for financial institutions (FIs) and FinTech for more than 20 years ultimately helping consumers get better lending rates, lower fees, higher returns, and more. Envestnet | Yodlee are the market leader in financial data aggregation, partnering with more than 1,200 financial institutions and FinTech innovators, including 15 of the top 20 U.S. banks, enabling a massive data network associated with tens of millions of consumers who use platform-related personalized apps and services. The power of Envestnet | Yodlee and their unique point of difference begins with the massive scale of financial data within the platform, which is utilized to fuel their data intelligence capabilities.

Challenge

Envestnet | Yodlee, was using Cloudera's CDH 4.4 for its Big Data environment. The environment handled over 2,000 jobs daily with over 1,000 customizations. They wanted to upgrade to CDH 5.7 but found an in-house upgrade was too risky due to the level of customizations and the risk of downtime. Envestnet | Yodlee evaluated several leading market solutions for an external upgrade service.

Solution

Envestnet | Yodlee found Cirata was the only solution that would enable an upgrade with no downtime and no business disruption. Cirata's patented technology ensured Envestnet | Yodlee's existing CDH 4.4 environment stayed fully operational during the upgrade to CDH 5.7 enabling the validation of all existing customizations so thorough testing could be done prior to their going live.

Results

- Simple and intuitive setup
- Support was readily available when required
- Over 200TB of data was replicated in a couple of days
- CDH 4.4 environment remained fully operational during the upgrade to CDH 5.7, enabling Envestnet | Yodlee to test applications in parallel in both environments
- Hive Metastore replication was performed at DB level, table level and for multiple tables
- Disaster recovery cluster is now infully active read/write mode enabling Envestnet | Yodlee to scale their deployment with existing hardware for major costs savings

> "One of the things that impressed us most about Cirata was its ability to resolve the data consistency challenges of replicating large amounts of data between active Hadoop clusters."
>
> *–Steven Swartzlander, Lead Architect at HM Health Solutions*

Company overview

HM Health Solutions (HMHS) delivers business solutions to health plan payers so they can run their organizations efficiently in a competitive and ever-changing market. By offering cutting-edge technology and unparalleled industry knowledge, HMHS meets the many operational needs of health plan payers. HMHS is partnered with 13 health plans serving 10 million members.

Challenge

In the United States, health plans and providers face tough challenges. The combination of an aging population and a constantly changing regulatory environment means that cost-effective delivery is more important than ever. And even though margins are tightening, delivering high-quality care experiences is essential to drive membership growth and encourage long-term member retention.

HM Health Solutions recognized that a new generation of machine learning (ML) and artificial intelligence (AI) capabilities could empower its healthcare affiliates to harness their data to solve these challenges.

To support informatics and data science teams across its affiliate network, HM Health Solutions decided to build a cutting-edge data lake, based on a Hortonworks Data Platform Hadoop cluster. The aim was to create a one-stop shop for analytics workloads, with the scalability to accommodate growing data volumes in the long-term.

Because the new Hadoop cluster was intended to become a critical data platform, HM Health Solutions needed a way to deliver round-the-clock availability with full protection for Hadoop Distributed File System [HDFS] and Apache Hive data. The question was how to most efficiently replicate its data to a secondary DR site.

Solution

To achieve its DR and high-availability goals, HM Health Solutions selected Cirata to continuously replicate HDFS and Hive data between both sites. By configuring Cirata for bidirectional replication, HM Health Solutions' DR environment could be used for production workloads, effectively doubling its compute capacity without compromising on its DR capabilities.

Results

- Enables seamless business continuity in the event of an outage, reducing operational disruption
- Ensures near-zero data loss in a recovery scenario, keeping valuable healthcare data protected
- Offers a scalable solution able to handle predicted 30% year-on-year increase in data volumes was performed at DB level, table level and for multiple tables
- Disaster recovery cluster is now in fully active read/write mode enabling Envestnet | Yodlee to scale their deployment with existing hardware for major costs savings

The University Of Sheffield.

> "This project wouldn't be possible without moving around large volumes of data that is heterogenous and changing over time."
>
> *–Alejandro Frangi, Professor of Biomedical Image Computing, The University of Sheffield*

Company overview

The Center for Computational imaging & Simulation Technologies in Biomedicine (CISTIB) group based at the University of Sheffield performs cutting-edge research in areas of fundamental and applied biomedical imaging & modeling with impact in personalized minimally invasive therapies and active and healthy ageing. Our team has an international and interdisciplinary profile and has a strong commitment to clinical and industrial translation with impact in future healthcare.

The University Of Sheffield.

Challenge

The University of Sheffield's CISTIB group wanted to learn more about the underlying pathology of dementia. They wanted to use a single Multix platform to analyze the rich library of unstructured biomedical data they had from over 6000 patients. The platform needed to be able to move the unstructured data easily and efficiently between 8 different cloud providers and 6 HPC Providers so it could be analyzed by over 950 different applications.

Solution

The CISTIB group looked at a number of different vendors to transfer and replicate data on an ad-hoc basis and during live synchronizations from on-premises to supporting cloud providers. They compared Cirata's technology to a number of alternatives and found Cirata was the only solution that could transfer continually changing data to the cloud but also do it at the speed and volume they required.

Results

- Cirata offered the most impressive data migration performance in the market and was the only solution offering active-active WAN replication.
- Cirata was found to be more resilient than other alternatives and was the only vendor solution which offered automatic recovery and guaranteed consistency.
- With Cirata's the CISTIB can analyze unstructured patient data, whether in terms of primary or original data (e.g. medical histories, radiological information, etc.); secondary or derivative data from analysis and exploitation; or even metadata associated with patient's genetic or therapeutic information.

About Cirata

Welcome to Cirata – a new company with over 45 patents and 15+ years of data science expertise in rapidly integrating high value datasets to leading cloud platforms for game changing AI activation and analytics insights.

We accelerate data-driven revenue growth by automating data transfer and integration to modern cloud analytics and AI platforms without downtime or disruption.

For more information on Cirata, visit www.cirata.com.

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

eb-ccs-230912