GoDaddy migrates very highly used Hadoop cluster to Amazon S3 with zero business disruption

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.

Objective

While the on-premises data lake is business critical, it is aging and running on an old version of Apache Hadoop (2.8). GoDaddy wanted to modernize the implementation by migrating the data to Amazon Web Services (AWS) to take advantage of the modern tooling and analytics capabilities available on AWS, and mitigating the risks and costs associated with maintaining the on-premises Hadoop cluster and the underlying hardware.

Challenge

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

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.

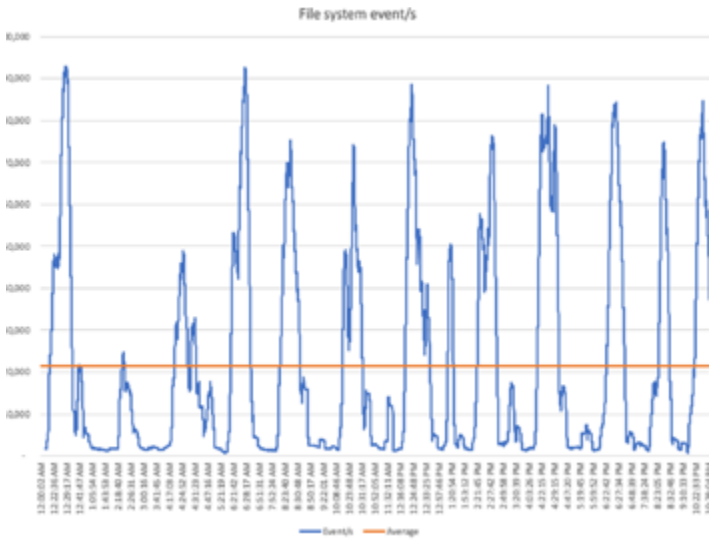
GoDaddy is empowering everyday entrepreneurs around the world by providing all of the help and tools to succeed online. With 20 million customers worldwide, GoDaddy is the place people go to name their idea, build a professional website, attract customers and manage their work. GoDaddy's mission is to give their customers the tools, insights and the people to transform their ideas and personal initiative into success.

Current deployment

Data is critical to GoDaddy. It is used continuously to guide business operations, configure products, and provide client services to its millions of customers. GoDaddy utilizes an 800-node Apache Hadoop cluster to hold over 2.5 petabytes of customer-related activity and behavior data, growing with more than 4TB of change every day. The system is in operation 24x7 and very actively utilized.

One measure of this activity is the rate of change in the file system. The graph below shows a moving average of file system change operations per second in this cluster over 12 hours. It performed an average of 21,388 file system change operations every second, and had activity peaks around five times that rate.

> "Cirata provided a solution that addressed our cloud migration challenges. This includes managing the technical debt inherited by our aging software and hardware, handling the scale of data that we accumulated over the years, and doing so without impacting our business continuity. We couldn't afford any downtime during the migration process. With Cirata and AWS we were able to get up and running in weeks so we could begin experimenting with the new cloud environment very quickly."
>
> *Jeremy Zogg, Senior Director of Engineering, Godaddy*



Data migration requirements

Although this cluster has been operating successfully for years, the organization wanted to modernize the implementation by migrating the data to AWS S3, and perform analytics using AWS EMR, Athena, and Redshift Spectrum. GoDaddy wanted to take advantage of the scale, depth of services, demand-driven operational cost effectiveness, proximity to other data and services, resilience, the community of users and of vendor solutions, security, compliance options, the pace of innovation, and other benefits that come with Amazon Web Services.

The CTO had stated that analytics applications must be running on AWS by the end of 2020. In order to achieve this goal, GoDaddy wanted to move its critical business data in Apache Hadoop to AWS EMR as quickly as possible. The scope of the initial project was to migrate a selected subset—totalling 500TB over 8.6 million files—of their 2.5 PB to AWS S3, allowing GoDaddy to run production workloads at scale in the cloud to validate operation and performance before migrating the majority of their cluster data.

This was complicated by the additional requirement to maintain operations at all times during migration, and to constrain the bandwidth used to a subset of their available 10 Gb/s capacity: no more than 3 Gb/s. There is no room for downtime or service disruption for the cluster or other network users.

Requirements summary

- Migrate 500TB of actively changing data from Hadoop to AWS S3
- Complete the data migration in as short a time as possible
- Maintain current operations at all times (zero system downtime)
- Ensure any changes made during the migration process are replicated appropriately

Results

The project resulted in a successful migration with the following outcomes:

Successful migration of initial 500TB data subset with zero business disruption
With Data Migrator GoDaddy was able to complete the 500TB migration, while maintaining business continuity and ensuring all data and changes were migrated successfully.

Reduced cost and risk of custom data migration development
The automated data migration reduced the cost and risks associated with manual and custom data migration initiatives, and enabled GoDaddy's engineers to focus on high value tasks, such as analytics, AI and machine learning development.

Faster time-to-value for AWS services
Data Migrator enabled GoDaddy to establish the new AWS environment much more quickly than would have otherwise been possible. This allows GoDaddy to focus on the new cloud solution and ensure they meet their objectives of strengthening their platform, increasing their pace of experimentation, and accelerating delivery of their product to provide increased value to customers and financial outcomes to their shareholders.

Cirata solution

GoDaddy, being a technically oriented company with deep software development skills, often builds their own solutions. As such, they investigated building their own custom migration solution leveraging open source tools. However, it was deemed that performing the initial migration and ongoing synchronization manually is a complex, error-prone task, and not the core competency on which they wanted their highly skilled engineers to spend their time.

Instead, following a quick demonstration of a 2TB migration, and a subsequent 10TB proof-of-concept GoDaddy selected Cirata Data Migrator to automate the migration. Data Migrator combines a single scan of the source datasets with processing of the ongoing changes that occur to achieve a complete and continuous data migration. It does not impose any cluster downtime or disruption, and requires no changes to cluster operation or application behavior.

> "Data Migrator's uniqueness lies in how it packages Hadoop data migration as a fully hands-off service. Moving data under active change is delicate, and organizations don't want to use their best IT people on it. Data Migrator handles everything in the background and doesn't require expertise from the customer. It's as close to a silver bullet as you can find for this type of project."
>
> *Merv Adrian, Research Vice President of Data and Analytics, Gartner*

> "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 Dirata'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 dnd Analytics Officer, Godaddy*

