LiveRamp

Work Smarter with Cloudera

Overview

Founded in 2006, LiveRamp's mission is to make it incredibly easy for marketers to access the data they need so they can personalize content for their customers. LiveRamp helps clients "fill in the blanks" about their customers by taking contact lists and, in real time, providing supplemental data points, statistics and aggregate charts and graphs that are guaranteed to have greater than 90 percent accuracy. LiveRamp is powered by Cloudera.

The Challenge

LiveRamp established itself as a data driven business early on, collecting feeds from numerous sources to create a single, accurate view of each customer. By 2008, "we were processing data in a complex pipeline that involved an organic structure of many MySQL instances and queues," explained LiveRamp's vice president of engineering, Jeremy Lizt. "As data volumes increased, that structure became unmanageable and expensive. It started getting difficult to perform the kinds of operations that we wanted to be able to do. It was no secret that this wasn't going to scale."

As part of its data synthesis, LiveRamp performs numerous processes and analytics to get the data it collects into a refined state. If they want to change an algorithm, that requires re-processing of all their data. Because of this, LiveRamp didn't even consider migrating to a relational data warehouse. "If we had a stable set of data and had to do incremental updates, a traditional RDBMS might have been appropriate," commented Lizt, who had read Google's papers on MapReduce and learned about Hadoop as an open source implementation of the Google paradigm. "It was pretty straightforward: Hadoop was the clear opportunity and seemed to be the only option for us."

The company made the move to Apache Hadoop and was running in production after nine to twelve months. They managed the environment independently for a year before learning about Cloudera. "We were already invested in the promise of Hadoop and our feeling was that Cloudera would be good for the community and good for Hadoop," said Lizt. LiveRamp decided to migrate to Cloudera's open source Hadoop distribution (CDH) due to its quality and stability, and soon signed on as one of the first Cloudera Enterprise customers.

Key Highlights

Industry

- Digital Media

Location

- San Francisco, CA, USA

Business Application Supported

- Complex data processing to generate a single, accurate customer view

Impact

- Data processing pipeline delivers > 90% accuracy
- 24x7 reliability
- Smarter allocation of engineering efforts

Technologies in Use

- Hadoop Platform: Cloudera Enterprise
- Hadoop Components: Cascading, Cascalog, HDFS, MapReduce, Scribe, Zookeeper
- Servers: Dell
- Data Mart: MySQL

Big Data Scale

- 250 TB on 280 CDH nodes; capacity for 400 TB
- 1-2 TB log data ingested daily
- Unlimited data retention

Advice to New Hadoop Users

- Manage your cluster through a centralized system
- Make sure you have good monitoring tools

The Solution

LiveRamp processes the many feeds of data that it collects and synthesizes all of that data into a single, accurate view using Hadoop. Log messages are sent through Scribe and loaded into the Hadoop Distributed File System (HDFS). Log data is loaded into Hadoop every ten minutes, amounting to one to two TB each day. Other data sources load hourly or daily. LiveRamp has other jobs that run periodically on the logs to compute stats and make sure everything is running correctly.

After processing their data in CDH, LiveRamp puts it into a distributed hash table — an open source key value store built by LiveRamp called Hank — that can be queried with very predictable, fast response times. The company has 250 TB on 280 CDH nodes today, with capacity for up to 400 TB of raw, unreplicated data. MySQL is also still in use at LiveRamp, mostly to store Hadoop's output.

Impact: Business-critical Reliability at Scale

Hadoop provides the foundation for LiveRamp's business. "If something is critical to your infrastructure, it's hard to articulate value. It's like asking what your heart or your liver is worth," commented Lizt. With that being said, the reliability and stability of LiveRamp's Hadoop platform is imperative. This is the biggest benefit that Cloudera Enterprise offers to LiveRamp.

As an early Hadoop adopter, LiveRamp doesn't rely heavily on the support provided by Cloudera but finds peace of mind in knowing it's there. "Every time we've had an issue, we've had very fast support," said software engineer Andre Rodriguez.

Lizt added, "Cloudera's engineers are a talented bunch of people — they're really intelligent, and we have confidence in their abilities. Whether it's a glitch or just a question, it's really helpful for us to be able to get a quick answer from someone at Cloudera who knows what they're talking about."

Impact: Operational Efficiency

Hadoop delivers a massively scalable data processing and storage platform that costs, on average, ten times less than traditional relational systems. But deploying Hadoop and keeping the cluster running at peak performance is no easy task. LiveRamp has found value in Cloudera's ability to simplify the deployment, management and monitoring of Hadoop through support, services and the Cloudera Manager tool.

Rodriguez reported three main advantages offered by Cloudera Manager:

- **Job-level statistics**: "A lot of the things that I used to do manually before, I can just do through Cloudera Manager now," said Rodriguez. "We can go back and see specifically what happened with jobs, get statistics from each job, and — perhaps most importantly — it keeps that information in the database. We actually use that data to compute other metrics so we can decide where to spend our engineering time."
- **Configuration management:** Cloudera Manager provides explanations for what every configuration parameter means. "We could get that information before, but it wasn't easy," noted Rodriguez.
- **Visibility into long-term trends**: While LiveRamp had a very mature configuration management system before implementing Cloudera Manager, they had less visibility into long-term trends such as how things were performing over time. This is another key benefit that helps LiveRamp identify focus areas for their engineering efforts.

About Cloudera

Cloudera is revolutionizing enterprise data management by offering the first unified Platform for Big Data, an enterprise data hub built on Apache Hadoop. Cloudera offers enterprises one place to store, process and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data. Only Cloudera offers everything needed on a journey to an enterprise data hub, including software for business critical data challenges such as storage, access, management, analysis, security and search. As the leading educator of Hadoop professionals, Cloudera has trained over 40,000 individuals worldwide. Over 1,400 partners and a seasoned professional services team help deliver greater time to value. Finally, only Cloudera provides proactive and predictive support to run an enterprise data hub with confidence. Leading organizations in every industry plus top public sector organizations globally run Cloudera in production. www.cloudera.com.

cloudera.com

1-888-789-1488 or 1-650-362-0488
Cloudera, Inc. 1001 Page Mill Road, Palo Alto, CA 94304, USA

cloudera®