connexity

Complements the EDW with Cloudera to Improve Retail Insights

Overview

As an online retail pioneer for over 15 years, Connexity, Inc.—formerly Shopzilla, Inc.—operates a portfolio of shopping websites and is recognized as a leading source for connecting buyers and sellers in the digital world. Through its destination websites and affiliate networks, Connexity touches a global audience of more than 50 million shoppers each month, linking them with over 100 million products from tens of thousands of retailers. The company equips shoppers to make informed decisions on the best products and deals, and it partners with retailers and advertisers to connect them with consumers, gain insight into their business, and boost their marketing ROI.

Connexity's portfolio of comparison shopping, merchant ratings, and review sites includes Bizrate, Beso, Shopzilla, Retrevo, TaDa, PrixMoinsCher, and SparDeinGeld, as well as B2B businesses, including the Connexity Publisher Program and Insights powered by Bizrate. Headquartered in Los Angeles, California, the company operates sites and business services in the United States, the United Kingdom, France, Germany, and other countries.

With more than two billion page views a month, Connexity is recognized among the leading providers of search engine marketing (SEM) and search engine optimization (SEO) services. To accommodate its requirement to process and deliver insights on millions of page views or ten billion ad bid requests daily, reaching over 100 million unique visitors, the company has deployed Cloudera to complement its Oracle enterprise data warehouse (EDW) in a hybrid Big Data environment that meets the needs of a wide range of users.

The Challenge

Throughout its history, this retail pioneer has enjoyed steady growth. But in 2011, the company – then known as Shopzilla – underwent rapid expansion and its 10-year-old legacy system had reached capacity. Shopzilla was attempting to process billions of data items and the system simply could not handle it.

Rony Sawdayi, vice president of engineering at Connexity, commented, "It was taking hours to process 100 million products per day using our incumbent system. We needed to decrease the processing time because the data was not feeding our downstream systems—particularly data science and business analysis—in a timely manner."

cloudera®

KEY HIGHLIGHTS

Industry

- Retail
- Business Services

Location

- Los Angeles, California, USA

Business Applications Supported

- Online price comparison services
- Search engine optimization
- Search engine marketing
- Merchandising
- Audience scoring
- Data science

Impact

- Data processing reduced from days to hours or minutes
- Real-time reporting on 10 billion ad requests daily expands and improves insights into buyers, sellers, and advertisers
- Improved monetization of data science and business analysis processes

Technologies in Use

- Hadoop Platform: Cloudera Enterprise
- Hadoop Components: Apache HBase, Apache Hive, Apache Mahout, Apache Pig, Apache Spark, Apache Sqoop, Cloudera Impala, Cloudera Manager
- Servers: Dell
- EDW: Oracle
- BI & Analytic Tools: Oracle BI Enterprise Edition (OBIEE); R

Big Data Scale

- Processing 15,000 feeds and 100 million retail products per day
- Scoring and bidding on 10 million keywords each day

He continued, "Latency is always an important consideration for us. Every day we process over 15,000 feeds from retailers. The process of aligning their products to our own product categories, enriching the data, classifying it, and then building a search index was taking as much as two days, and then another ten hours to reformat and provide feedback to our partners. By the time data was delivered it was frequently stale or possibly even inaccurate."

Meanwhile, the company's 500-terabyte EDW was growing by five terabytes every day. "To process and aggregate the data, and pull reports via business intelligence (BI) tools, was taking far too long. We knew we had to find an alternative solution," commented Sawdayi.

Executives considered numerous options with three key criteria in mind: performance, reliability, and flexibility.

Of paramount importance was the ability to expedite data processing and manipulation performance, which led to evaluations of distributed parallel processing systems such as Apache Hadoop and other NoSQL offerings. "We knew that providing data to our data scientists and business analysts in a shorter time period would improve monetization," noted Sawdayi.

Several internal groups had started prototyping systems in Hadoop. Sawdayi recalled, "Hadoop was clearly proving itself to be reliable, flexible, and fast, but we were concerned about our ability to manage the compatibility of individual components; the environment evolves so rapidly that keeping track of everything on our own was a daunting proposition. We started looking for a distribution that met all of our needs and a mature partner who could coordinate everything, and help us migrate data and optimize the clusters."

The Solution

"Cloudera had the expertise, the experience, and the community support we needed to be successful," said Sawdayi. "We initially worked with the company in 2011 when we started migrating some of our test systems to CDH. Questions were answered quickly, the Cloudera University training we took was very professional, and their support was terrific. It was only natural that we partner with Cloudera on this initiative."

Connexity has augmented its Oracle EDW with a multi-tenant Cloudera Enterprise system to create a hybrid environment. A large amount of processing, cleansing, transformation, and crunching is done in the Hadoop environment and then aggregated data is pushed into the Oracle data warehouse via Apache Sqoop for reporting. Connexity has written a custom tool, known internally as Forklift, to move data from Oracle into CDH in an optimized fashion.

Connexity's use of Cloudera with Apache HBase supports real-time reporting across the huge number of variables that the company tracks. Sawdayi noted, "HBase supports the use of 'atomic counters' and we use over 500 billion of these to provide real-time reporting in a very scalable, distributed manner. This information is not only used by our business analysts to manage and optimize our business systems but also to populate a portal for our clients to see how their own advertising is performing."

Oracle Business Intelligence Enterprise Edition (OBIEE) is being configured to allow users to pull reports from Oracle as well as from Cloudera. Users have also been trained to use Apache Pig and Apache Hive for direct access to Cloudera. And now the team is upgrading its clusters to take advantage of Cloudera Impala and Apache Spark, which will offer even better performance for analytics on data in Hadoop.

Paramjit Singh, director of data for Connexity, elaborated, "We currently utilize three main clusters: One that runs all of the company's optimized media solutions, another that handles mission-critical applications, and an analytics and research cluster that is tasked with ad-hoc processing."

Cloudera Manager is used to provision, monitor, and manage all of Connexity's clusters. The tool also performs automatic installations, alerting, and trending. "We're running about 3,000 jobs per day," said Singh, "and we rely on Cloudera Manager to give us insights as to what is happening on the cluster, how to administer it, which jobs are taking the most resources, and which jobs are not getting enough resources so we can optimize performance."

Impact: Faster Processing

For its merchandising process, Connexity takes over 15,000 feeds and 100 million products from retailers and processes them with Cloudera each day. Sawdayi pointed out, "What once took several days has been reduced to just a few hours, and a new approach is being tested that will slash that to minutes. Having this processing capability available has also allowed us to provide full service offerings to our retailers, enabling them to target audiences on a massive scale."

The improved processing performance also benefits the company's SEM activities. "Our SEM system is one of the best in the world," said Singh. "With the processing power of our Cloudera platform, Connexity is able to score and bid on ten million keywords each day. For our audience solutions business, we're in a position to touch 100 million unique visitors over the web, and collect billions of data points to feed our data science and create immensely rich shopping intent data."

Sawdayi added, "Many of the things we do as a business would not be possible without this platform."

Impact: Detailed Insight, Relevant Results

Connexity uses a combination of Apache Mahout and R running on Cloudera to perform classifications and user segmentation on its analytics and research cluster. "We are able to answer complex questions, such as how a user is behaving on a particular site and what ads would be most effective, as well as execute other sophisticated data mining queries. It improves Connexity's ability to provide relevant results to users, and this is a core tenet of our business," Singh explained.

He observed, "Data scientists don't typically need to consume data warehouse resources now because all of the most recent data is available in Cloudera via R or Mahout. They have tremendous insight into the data, which would be virtually impossible to obtain otherwise. Connexity has billions and billions of rows, which simply cannot be handled by our existing relational data warehouses. We needed enormous processing capabilities, scalability, full redundancy, and extensive storage–all at a cost-effective price. Our Cloudera platform provides all that and more."

Sawdayi concluded, "Cloudera is and continues to be an ideal partner. The company provides everything we need from a full-service vendor, including comprehensive training, prompt support, and professional consulting. Cloudera is there for us on a daily basis while helping us to determine next steps and evaluate new technologies."

"We needed enormous processing capabilities, scalability, full redundancy, and extensive storage–all at a cost-effective price. Our Cloudera platform provides all that and more."

Paramjit Singh, Director of Data, Connexity, Inc.

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 800 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