EQUIFAX

Building Better Data Models with Cloudera

Overview

Equifax organizes, assimilates, and analyzes data on more than 600 million consumers and 80 million businesses worldwide to deliver insights that support a wide range of applications—from assessing a consumer's credit risk to helping businesses grow their customer base to helping government agencies fight fraud.

The company's success depends on its ability to perform risk analysis and build data models very quickly, with a high degree of statistical accuracy. Until recently, the company's data scientists used proprietary tools to prepare data for the development of new data models. By deploying an advanced analytics environment powered by Cloudera as part of its long-term enterprise data hub (EDH) strategy, Equifax data scientists more quickly analyze much larger data sets so their models can be built in less time than prior methods, with a high degree of accuracy. And the real-time insights delivered by Impala, an analytic database, enable data scientists to more quickly test new theories as they design new data models.

The Challenge

According to Yuvaraj Sankaran, vice president of Technology, Data Services, and Analytics at Equifax, as data volumes have increased, so has the company's need for accelerated turnaround time.

"The assembly of the data was very time-consuming and onerous. It would take weeks from the time a model was conceived to the time it was built and delivered," said Sankaran.

While the primary goal of building a new big data and analytics platform was to reduce the development time for new data models, Sankaran said the ability to perform deeper analytics on larger data sets was also critical.

"We constantly had to bring the data out of the proprietary data platforms into niche analytics environments to analyze and visualize small data samples," Sankaran explained. "We wanted a more open platform that would allow us to do advanced analytics on much larger data sets. And that led us to [Apache] Hadoop."

cloudera®

Key Highlights

Industry

- Financial Services

Locations

- Operates in 19 countries
- Headquarters: Atlanta, GA, USA

Business Application Supported

- Analytics sandbox

Impact

- Significant time reduction in delivery of new data models
- Deeper insights to help solve customer challenges
- Simplified management and reduced costs

Technologies in Use

- Hadoop Platform: Cloudera Enterprise
- Hadoop Components: Apache Hive, Apache Mahout, Apache Sentry, Apache Spark, Cloudera Impala
- ETL tool: Talend
- Analytics tools: Alpine Data Labs, R, SAS
- Security tool: Protegrity
- Servers: HP ProLiant DL360p Gen8 Servers using Intel® Xeon® E5-2600 v2 processors (management node), HP ProLiant SL4540 Gen8 Servers using Intel Xeon E5-2400 processors (edge node)

Big Data Scale

- Analysis of five years' data

"The open nature of the Cloudera platform will allow us to bring deeper insights to solve our customers' problems. We are looking forward to utilizing all this power on the massive datasets that we have at Equifax."

Yuvaraj Sankaran, Vice President, Technology, Data Services, and Analytics, Equifax

Solution

As Sankaran's team evaluated Hadoop offerings, they considered several important requirements, including each vendor's ability to:

- Deliver an integrated platform that supports a wide range of analytic and data integration tools—including those from Alpine Data Labs, SAS, and Talend
- Provide real-time querying capabilities
- Support a wide range of security systems

"We looked at different technologies in the marketplace," said Sankaran. "Cloudera is a leader in the Hadoop space and their team includes some of the leading thinkers in big data. They spent time discussing our problems with us, and they offered the experience, the integrated and secure platform, and the tools we needed to perform the low-latency data exploration work we want to do."

By implementing Cloudera as its advanced analytics sandbox, Equifax can process and analyze large volumes of data faster, allowing data scientists to build better models based on larger data sets. Equifax runs Cloudera on servers powered by the Intel® Xeon® processor E5 family to deliver the compute power and data throughput needed for rapid processing of these large data sets.

Data scientists have the flexibility to build models using the right tool for the job and for the user, whether that is Apache Mahout, Apache Spark, Impala, or R. Equifax has also implemented Apache Sentry (incubating), allowing sensitive data to be stored in the Cloudera environment via unified authorization and role-based access controls across all Hadoop access paths.

Impact: Time Savings for Deeper, Faster Insights

Equifax's primary goal was to reduce the time it took to develop new data models, and the deployment of the new analytics platform is expected to deliver dramatic time savings.

"With Cloudera, we are pre-assembling all the data and have installed various tools on top of Hadoop," said Sankaran.

Because data scientists have access to Mahout, R, and Spark in the open platform, they will be able to apply new techniques that offer greater understanding of much larger datasets.

"The open nature of the Cloudera platform running on infrastructure powered by Intel Architecture will allow us to apply deeper insights to solve our customers' challenges," said Sankaran.

Impact: Simplified Management Reduces Costs

Through the use of Impala, data scientists can perform real-time queries in their analytics sandbox, without having to move the data to another platform. This reduces the cost and effort of managing and synchronizing data across different environments.

"When Cloudera announced Impala, we saw there was no need to move data anywhere. The insights and data can sit side by side. We can do the analysis in one place and secure one cluster. It makes it much easier to deal with a single platform than multiple platforms. This was one of the reasons why we chose Cloudera," said Sankaran.

Based on the organization's early success with its Cloudera analytics sandbox, Equifax has a long-term vision of implementing an integrated enterprise data hub to maximize its use of big data.

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 1400 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®