PRGX®

Achieving 9-10x Performance Improvement and Driving Innovation

Overview

PRGX Global, Inc. is the world's leading provider of accounts payable recovery audit services. With over 1,400 employees, PRGX operates and serves clients in more than 30 countries and provides its services to over 75 percent of the top 20 global retailers.

The company's goal is to help its clients detect, find, and fix leakage in their procurement and payment processes. To do so, PRGX auditors must analyze purchasing, receiving, and payment transactions, along with buyer/supplier contracts, agreements, and emails, to find and recover overpayments. It's a huge task with more than 2.3 petabytes (PB) of data processed annually for its clients.

Working with Cloudera and Talend, PRGX created a high-performance computing platform for data analytics and discovery that could more rapidly process, discover, model, and serve this massive amount of structured and unstructured data. This new platform delivers on average 9-10x performance improvements—with a 45x performance improvement in one case.

Faster performance translates into more auditing time. The more auditing time PRGX staff has, the more payment errors they can identify. The result is greater profitability for both PRGX's clients and the company itself.

Additionally, greater scalability and flexibility to incorporate new data types is expected to help PRGX innovate and offer new products and services.

The Challenge

"Our legacy environment was a RDBMS [relational database management system], and it served us well for a number of years," said Tushar Sachdev, CIO and Senior Vice President, PRGX. "But as we moved into larger data volumes and into the unstructured space, our RDBMS showed some strains at times. It still could serve the purpose, but processing a billion rows took longer than we wanted and we can't afford that. The more time our auditors get to audit, the more money we can deliver to our clients."

PRGX typically mines 2.3 PB of "live" data annually—which comes from various systems and sources including purchasing, payment, receiving, deals, point of sale, and more than 150 million emails.

cloudera® talend

Key Highlights

Industries

- Business Services
- Technology
- Retail

Locations

- Headquarters: Atlanta, Georgia, USA
- Operates in 30+ countries

Business Application Supported

- High-performance computing platform for data analytics and discovery

Impact

- 9-10x performance improvement on average, with a 45x increase in one case
- 25% decrease in storage costs
- Greater innovation via the delivery of new products and services

Technologies in Use

- Hadoop Platform: Cloudera Enterprise, Data Hub Edition
- Hadoop Components: Apache HBase, Apache Hive, Apache Pig, Apache Spark, Cloudera Navigator, Cloudera Search, Impala
- Data Integration Tool: Talend

Big Data Scale

- 2.3 PB, including 150 million emails and 2 million files annually

"Our process goes from procure-to-pay, so we look at everything from when a deal was made, to when an order was placed, to how it was invoiced, shipped, paid, and sold," said Jonathon Whitton, Director, Data Services, PRGX. "We get a lot of different data types and files, with data arriving at a variety of intervals—from daily, weekly, and monthly, to quarterly and annually."

The lead time to deliver data to auditors was measured in weeks in certain situations and re-running processes as new data arrived required significant time and effort. While the company upgraded its legacy platform over the years to improve performance, executives realized a fundamentally new architecture was now required to keep the company competitive.

"The business has changed a lot, and we saw that Hadoop would enable us to stay with current technology," explained Whitton. "When PRGX first started, auditors went through boxes of paper, one box at a time. As the business grew, we moved to structured data, and then to emails. Now, we're looking at bringing other types of communications between buyers and sellers."

The Solution

PRGX executives sought to create a high-performance computing platform (called HiPer) based on Hadoop that would provide the scalability, flexibility, and performance it needed.

The IT team conducted a high-level assessment of 19 vendors, reviewing each company's strengths, solution scalability, and technology feasibility within PRGX's environment. This list was cut to just six vendors, each of which completed a Request for Information (RFI) regarding technical capabilities, performance, and information security. Ultimately, five vendors were invited to participate in a Proof of Concept (POC) demonstration.

"We looked at a number of options for our architecture, and at the end of the day, we zoomed in on Cloudera and Talend," said Sachdev. "We not only assessed the strength of the technology, which was great, but also looked at each company's stability, responsiveness, and pricing, and we selected Cloudera and Talend as our final choice."

Working with Cloudera, the company is implementing a data discovery and analytics environment that enables auditors to more quickly find relevant emails, confirm agreements regarding deals or promotions, and correlate the relevant data. Auditors are often given hundreds of documents to review to identify a single issue, and machine-learning techniques can reduce that number significantly to enhance their productivity. Additionally, the ability to re-run and experiment on large data sets will enable staff to identify new service opportunities to deliver more value to clients.

Talend's big data integration technology is used to bring structured and unstructured data into the Cloudera platform and prepare that data for analytics by using the power of Hadoop to cleanse and transform it. Using Talend's visual design environment, PRGX developers can easily create data integration "jobs" that run inside Hadoop. Talend generates native Hadoop MapReduce and Apache Spark code, saving developer time while taking full advantage of the scalability and parallel processing power of Hadoop.

Apache Hive and Pig are used to process the data in batch format and perform "known" queries from auditors. Impala enables data managers to conduct ad-hoc queries for quality assurance.

Critical to the company's success is the ability to find communications about a specific deal or promotion across millions of emails. To facilitate this work, the company uses Cloudera Search. Explained Billy Detweiler, Senior Software Developer, PRGX, "With our legacy system, when we searched for specific words, such as 'Super Bowl' and 'Sunday,' the search results provided any email with those phrases. We then had to manually review a large group of emails to

find the ones that were relevant to a 'Super Bowl Sunday' sale. With Cloudera Search, we can search for words that are near each other. This has allowed us to hone in on the right emails even faster, so we only have to look at a few emails in all."

Detweiler also called Spark "a saving grace" when it comes to correlating emails for each client. "One vendor may have different email identifiers, such as .com and .uk.com, for its different entities worldwide," said Detweiler. "Spark's distributed processing capabilities allow us to enrich searchable metadata fields associated with client emails with one (or more) custom value(s) very quickly, which makes searches quicker and more effective."

According to Whitton, finding the source of any data issue is also easier with Cloudera tools. "Our auditors identified a data delivery issue and we needed to find where the problem was," he said. "With Cloudera Navigator, we were able to trace the issue back to the source data set from the client in a matter of hours as opposed to weeks. It was amazing."

Impact: Achieving 9-10x Performance Improvement

According to Sachdev, the performance benefits of the company's new platform have been phenomenal. "We anticipated 2-3x improvement in performance from our legacy process, and instead we are averaging 9-10x improvement," he said. "In one case, we saw a 45x improvement. This is significant in terms of what it can do for our business and for the future of our company."

Faster data processing translates into more audit time, which directly increases revenue. Additionally, the ability to find the right email in 150 million emails much more quickly drives greater staff productivity and, ultimately, greater profitability for PRGX and its clients.

"Certain processes that took 140 hours to complete previously, now take six hours," said Tony Cupid, Vice President, Data Services, PRGX. "That means we're delivering data much faster to the business units, so they are able to spend more time auditing, which allows them to find more money for clients."

Impact: Reducing IT Costs

IT organizations continue to face significant budgetary pressures, and, as the amount of data grows, many IT teams are seeking ways to control storage costs. At PRGX, Sachdev's team saw Hadoop as a way to cost-effectively manage its increasing data volumes.

"From an IT perspective, we knew the reduction in storage costs that we'd get from Hadoop," Sachdev said. "We will be operating at one-fourth the footprint in terms of storage costs, which is a big benefit."

Impact: Enabling the Delivery of New Products and Services

The ability to mine large volumes of unstructured data and experiment on large data sets is expected to open new business opportunities for PRGX.

"The platform is enabling innovation in our core business, the recovery audit business, but it also allows us to develop new products and new services for our clients that we were not able to do with our legacy platform," said Cupid.

In fact, executives believe the platform will help PRGX achieve the "holy grail" in the recovery audit industry. "The holy grail in our industry is to prevent the erroneous transaction from happening in the first place," said Sachdev. "With what we are doing on the data processing side, and what we can do with the emails and the unstructured data, I think we can get there."

Sachdev concluded, "This project has frankly changed the way people view IT in our organization. We were people who fixed things. Now, we're looked at to take the company to the next level. It's a paradigm shift."

> "We are averaging 9-10x improvement [in performance]. In one case, we saw a 45x improvement. This is significant in terms of what it can do for our business and for the future of our company."
>
> — Tushar Sachdev, CIO and Senior Vice President, PRGX

About Talend:

Leading organizations use Talend's integration solutions to gain instant value from all their data with the goal of becoming data-driven. Talend's open and unified solutions take the complexity out of any integration project and equip IT to be more responsive to the demands of the business, at a predictable cost. By design, Talend integration software simplifies the development process, reduces the learning curve, and decreases total cost of ownership with a unified, open, and predictable platform. Only Talend runs natively in Hadoop using the latest innovations from the Apache ecosystem. Talend is now the first data integration platform built on Spark, giving customers up to 100X better performance than any other solution on the market. Talend makes it easy to build big data integration jobs leveraging Spark, Spark Streaming and machine learning, all with an easy to use, visual, eclipse-based designer. More than 1700 enterprise customers worldwide rely on Talend's solutions and services. www.talend.com

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, access, process, secure, 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. Cloudera's open source big data platform is the most widely adopted in the world, and Cloudera is the most prolific contributor to the open source Hadoop ecosystem. As the leading educator of Hadoop professionals, Cloudera has trained over 40,000 individuals worldwide. Over 1,700 partners and a seasoned professional services team help deliver greater time to value. 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

© 2016 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

Cloudera_CaseStudy_PRGX_106