cloudera
Ask Bigger Questions

CUSTOMER CASE STUDY

Experian: Transforming the Marketing Landscape with Cloudera

KEY HIGHLIGHTS

INDUSTRY
Digital Marketing

LOCATION
Schaumburg, IL, USA

BUSINESS APPLICATIONS SUPPORTED
- Matching engine for CCIR engine, facilitating a holistic and current view of consumers

IMPACT
- 50x performance gains
- Processing 500% more matches per day
- Deployment in < 6 months

Company Overview

With 15,000+ employees and annual revenues exceeding $4 billion (USD), Experian is a global leader in credit reporting and marketing services. The company is comprised of four main business units: Credit Information Services, Decision Analytics, Business Information Services, and Marketing Services.

Experian Marketing Services (EMS) helps marketers connect with customers through relevant communications across a variety of channels, driven by advanced analytics on an extensive database of geographic, demographic, and lifestyle data.

Business Challenges Before Cloudera

EMS has built its business on the effective collection, analysis and use of data. As Jeff Hassemer, VP of product strategy for EMS explained, "Experian has handled large amounts of data for a very long time: who consumers are, how they're connected, how they interact. We've done this over billions and quadrillions of records over time. But with the proliferation of channels and information that are now flowing into client organizations — social media likes, web interactions, email responses — that data has gotten so large that it's maxed the capacity of older systems. We needed to leap forward in our processing ability. We wanted to process data orders of magnitude faster so we could react to tomorrow's consumer."

In the past, it was normal to send customer database updates to clients once monthly for campaign adjustments, allowing Experian to process large volumes of data through a number of diverse platforms, mostly mainframe based. "We weren't required to provide data in real time. We weren't required to provide the level of volume in terms of the growth rates we've seen from our storage and our data. It's been a total paradigm shift that compelled us to look at other solutions," explained Emad Georgy, CTO for Experian Marketing Services.

Today's consumers leave a digital trail of behaviors and preferences for marketers to leverage so they can enhance the customer experience. Experian's clients have started asking for more frequent updates on consumers' latest purchasing behaviors, online browsing patterns and social media activity so they can respond in real time. "We serve many of the top retail companies in the world, and they're increasingly looking for a single, integrated view of their customer," noted Georgy. "We're looking for a holistic view of who that person is so we can determine how to message them in the right way."

KEY HIGHLIGHTS

TECHNOLOGIES IN USE

- Hadoop Platform: Cloudera Enterprise
- Hadoop Components: HBase, Hive, Hue, MapReduce, Pig
- Servers: HP DL380
- Data Warehouse: IBM DB2
- Data Marts: Oracle, SQL Server, Sybase IQ

BIG DATA SCALE

- 5B rows of data, growing tenfold
- Processing 100M records per hour
- 35 CDH nodes today

ADVICE TO NEW HADOOP USERS

- Involve expert architects early on
- Have patience & get trained on Hadoop/HBase before building applications

But the data exhaust from these digital channels is massive and requires a technological infrastructure that can accommodate rapid processing, large-scale storage, and flexible analysis of multi-structured data. Experian's mainframes were hitting the tipping point in terms of performance, flexibility and scalability. Given the need for immediacy of information and customization of data in real time for clients, EMS set an internal goal to process more than 100 million records of data per hour. That translates to 28,000 records per second.

"Instead of trying to fit a square peg in a round hole, we went out and decided to look for new architectures that can handle the new volumes of data that we manage," said Joe McCullough, IT business analyst at EMS. The team identified about 30 criteria for the new platform, ranging from depth and breadth of offering to support capabilities to price to unique distribution features. They prioritized two criteria above the rest:

- Both batch and real-time data processing capabilities
- Scalability to accommodate large and growing data volumes

"We compared Hadoop as well as HBase to a number of other options in the industry," said Georgy. "The North America Experian Marketing Services group has organically led the evaluation of NoSQL technologies within Experian." Hadoop and HBase quickly surfaced as a natural fit for Experian's needs. EMS engineers downloaded raw Apache Hadoop, but quickly saw the gaps that could be filled by a commercial distribution. EMS critiqued several distributions and "found that, by and far, Cloudera was in the lead. We went with Cloudera for a number of reasons, primarily being the strength of the distribution and the features that CDH gives us."

EMS' enterprise-level Hadoop needs, such as meeting client SLAs and having 24x7 reliability, led the organization to invest in Cloudera Enterprise, which is comprised of three things: Cloudera's open source Hadoop stack (CDH), a powerful management toolkit (Cloudera Manager), and expert technical support.

Use Case

A few months of exploring Hadoop translated into a production version of Experian's Cross-Channel Identity Resolution (CCIR) engine: a linkage engine that is used to keep a persistent repository of client touch points. CCIR runs on HBase, resolving needs for persistency, redundancy, and the ability to automatically redistribute data. HBase offers a shared architecture that is distributed, fault tolerant, and optimized for storage. And most importantly, HBase enables both batch and real-time data processing.

Experian feeds data into the CDH-powered CCIR engine using custom extract, transform, load (ETL) scripts from in-house mainframes and relational databases including IBM DB2, Oracle, SQL Server, and Sybase IQ. EMS' HBase system currently spans five billion rows of data, "and we expect that number to grow tenfold in the near future," said Paul Perry, EMS' director of software. Experian also uses Hive and Pig, primarily for Q&A and development purposes. EMS currently has 35 Hadoop nodes across its production and development clusters.

"Deploying Cloudera allows us to process orders of magnitude more information through our systems, and that technological capability in combination with Experian's expertise in bringing together data assets is driving new, real insights into tomorrow's marketing environments."

JEFF HASSEMER, VP PRODUCT STRATEGY, EXPERIAN MARKETING SERVICES

Impact: Operational Efficiency

Hadoop is delivering operational efficiency to Experian by accelerating processing performance by 50x. And the cost of their new infrastructure is only a fraction of the legacy environment. Georgy said, "We've been very happy with the implementation of Hadoop. We're less than six months in and are already closing the gap on our 100 million record per hour goal." In comparison, EMS used to process 50 million matches per day. That translates to a 500% improvement.

Further, Cloudera Enterprise allows Experian to get maximum operational efficiency out of their Hadoop clusters. "Cloudera Enterprise gives us an easy way to manage multiple clusters, and because the use cases for clients vary so much, we have to do a lot of tweaking on the platform to get the performance we need. The ability to store different configuration settings and actually version those settings is huge for us," said Perry.

McCullough added, "Not only has Cloudera Manager simplified our process, but it's made it possible at all. Without a Linux background, I would not have been able to deploy Hadoop across a cluster and configure it and have anything up and running in nearly the timeframe that we had."

Cloudera Manager delivers the following operational benefits to Experian:

- Monitors services running on cluster
- Reports when servers are unhealthy, services have stopped, and/or nodes are bad
- Automates distribution across the cluster easily
- Monitors CPU usage across various applications and data storage availability
- Provides a single portal to see into all cluster details

Perry summarized the project by saying, "We haven't done anything this cool for a long time. Our developers are excited about it. I'm excited about it. Senior management is excited about it. And the experience we've had with Cloudera — I don't think it could be better. It's been great working with those guys and the architects that they've given us access to are extremely knowledgeable. The only regret I have is that we didn't bring them in sooner."

Impact: Driving Competitive Advantage

"Our Hadoop infrastructure has become a real transformational change. Deploying Cloudera allows us to process orders of magnitude more information through our systems, and that technological capability in combination with Experian's expertise in bringing together data assets is driving new, real insights into tomorrow's marketing environments," stated Hassemer. "Nobody is doing what we're doing with Hadoop today, especially at this order of magnitude. Ours is the first data management platform of its kind that accepts data, links information together across an entire marketing ecosystem, and puts it into a usable format for a solid customer experience."

"One of the key enablers that investing in Hadoop with Cloudera will empower is allowing our clients to become obsessed with creating the perfect customer experience. This tool allows us to operate at the speed of business, at the speed where those interactions occur, so our clients can meet their customers with the right message at the right time."

JEFF HASSEMER, VP PRODUCT STRATEGY, EXPERIAN MARKETING SERVICES

The performance gains offered by Cloudera give Experian new insights and more flexible ways of understanding consumers. EMS no longer relies on a postal address to identify a consumer; they can now match social media IDs, email addresses, web cookies, phone numbers and more. With the broader match set that Hadoop enables, EMS' clients have a more accurate, current view of who their customers are across multiple channels so they can have better, more informed interactions.

"A consumer might give their email and sign up for a program with one particular brand," explained Hassemer. "The brand doesn't know much about the consumer at that point, thus information sent to the consumer is very bland, not very enticing. The consumer's time is wasted because the emails they receive aren't relevant. By offering a holistic view of the customer in real time, we can increase the relevance of offers that go to the consumer so they say, 'This is why I signed up.' And they're going to interact with that brand a lot more. One of the key enablers that investing in Hadoop with Cloudera will empower is allowing our clients to become obsessed with creating the perfect customer experience. This tool allows us to operate at the speed of business, at the speed where those interactions occur, so our clients can meet their customers with the right message at the right time."