Yellowbrick

CASE STUDY

Catalina Marketing Enables Data Scientists and Power Users with up to 182x Better Query Performance

BUSINESS CHALLENGE

Catalina Marketing is the market leader in shopper intelligence and targeted in-store and digital media. The company delivers $6.1 billion in consumer value annually, combining the richest buyer-history database in the world with its own deep analytics and insights to help retailers, CPG brands, and agencies optimize every stage of media planning, execution, and measurement.

Catalina understands there's a science behind every buy and a unique buyer behind the data. To uncover those insights, the company must ingest terabytes of data, process and transform it, and then consume and analyze the results to help customers mobilize meaningful, real-time engagement and results.

Until recently, the company's enterprise data warehouse ran entirely on an IBM Netezza system. It supported two main workloads:

- Data processing, consisting of the complex ETL processes required to convert nightly data feeds from customers into a normalized set of databases for querying and reporting
- Consumption of processed data by the company's analytics team, consisting of about 100 data scientists and other power users who use advanced analytics and data mining tools such as SAS, R, and Python to run large, ad hoc queries as they support customers across all of the company's lines of business

However, the Netezza system lacked

OVERVIEW

CATALINA®

Company
Catalina Marketing

Website
www.catalina.com

Country or region
United States

Industry
Retail/CPG

CUSTOMER PROFILE

Catalina Marketing helps retailers, CPG brands, and agencies optimize media planning, execution, and measurement.

SITUATION

The company's enterprise data warehouse ran entirely on an IBM Netezza system. It lacked the compute capacity to support both daily ETL processes and advanced analysis of that processed data by the company's analytics team.

SOLUTION

Catalina moved its analytics workload onto Yellowbrick, enabling the company's 100-member analytics team to run deep, complex queries on three years of POS data using tools such as SAS, R, and Python.

"Now that we no longer need to triage resource constraints day in and day out, we've been able to collaborate on the development of two new products—something that in the past we were unable to do."

Luis Velez, Data Engineering Manager, Catalina Marketing

the compute capacity to support both workloads. "It was an unsustainable environment, in which we weren't able to finish our data loads because we had 15 to 20 queries running at any given time," says Luis Velez, data engineering manager at Catalina. "Every day, it was getting a little bit worse."

The company's analytics team—the source of those queries—was also hamstrung by the lack of compute capacity, having to wait 20 minutes or more for their queries to run. "Sometimes queries took hours, and other times they were simply killed so ETL processes could run," says Aaron Augustine, executive director of data science at Catalina. "We had only small windows of time for heavy analysis—like maybe 25 percent of the day. Given that we have analytics teams overseas and in Japan, it was a 24x7 problem."

ENABLING POWER USERS WITH YELLOWBRICK

Catalina decided to augment its existing data warehouse with a second system, dividing the compute workload into two parts. Data processing would continue to run on the Netezza system, while consumption of the processed data—including queries by the analytics team—would be supported by a Yellowbrick Data Warehouse. "The Netezza box was running some pretty advanced ETL processes, which were working fine when there weren't a lot of big queries hitting the system," explains Augustine. "At the same time, we needed to give data scientists and other power users the means to ask some really big questions of the data, as needed, to discover entirely new insights into buyer behavior. Moving that ad hoc advanced analytics workload onto Yellowbrick was an incremental, low-risk approach that made a lot of sense."

Catalina discovered Yellowbrick after a failed proof-of-technology (POT) exercise with a vendor that couldn't deliver the needed performance or compatibility. During a successful three-

BENEFIT

- Up to 182x faster query performance
- Data load speeds of up to 10TB/hour
- More engineering resources for new product development
- Fast, low-risk migration—completed in four months with minimal code rewrites
- 10U form factor, enabling deployment in an already full data center
- High reliability, with no moving parts or single point of failure
- A partner that anticipates Catalina's needs and is committed to its success

week POT on Yellowbrick, a single 10U, 30-node Yellowbrick system delivered up to 182x better performance than an 8-rack, 56-node Netezza Mako system. The company immediately purchased the Yellowbrick system and spent the next four to six weeks migrating its 260TB of data. Over the following two months, Catalina moved the analytics team onto Yellowbrick. "All in all, our migration onto Yellowbrick took four months—far faster than we expected," says Augustine.

Today, the company's analytics team can work unimpeded. Queries that once took 30 minutes—if they weren't killed first—are now completed within a few seconds or minutes. "Our Yellowbrick system has made our analytics team a lot more productive," says Augustine. "These are power users doing deep and complex analytics—using tools like SAS, R, and Python to query three years of point-of-sale data. We're continuously hammering on Yellowbrick with some really big queries, and it's handling them very well."

Now that it's no longer hamstrung by a lack of compute capacity, the analytics team can fully contribute to all parts of the business:

- Data mining and predictive analytics as needed to fuel Catalina's targeting solutions
- Retail analytics to help retailers build effective multichannel campaigns
- Manufacturing analytics to help CPG brands optimize engagement and targeting
- Digital analytics to help agencies drive effective advertising and promotions for CPG clients

"Yellowbrick is being used across all of Catalina's work streams—data science and data mining—serving our retail partners, serving our brand partners, and serving our digital analytics teams," says Augustine. "And it's performing wonderfully across all those use cases. We were so happy with our first Yellowbrick system that we purchased a second one, which we plan to use to modernize some of our legacy applications."

Catalina is also considering how it can take advantage of Yellowbrick's hybrid cloud architecture to help the company move its computing workloads to the cloud. "Everything we're doing on premises is about keeping the lights on and making sure existing applications are running effectively," says Velez. "Our long-term vision is to retire our on-premises footprint and create a new platform in the cloud. We're excited to explore how we can partner with Yellowbrick on those efforts."

BENEFITS/RESULTS

Catalina' use of Yellowbrick is benefiting the company in several ways:

- **Uncompromised query performance.** Catalina's

“Our Yellowbrick system has made our analytics team a lot more productive. These are power users doing deep and complex analytics—using tools like SAS, R, and Python to query three years of point-of-sale data.”

Aaron Augustine, Executive Director, Data Science
Catalina MarketingI

100-member analytics team now enjoys superior query performance at all times, with 24x7 access to all the compute resources it needs to work at top speed. This is due to Yellowbrick’s unique massively parallel processing (MPP) architecture, which eliminates the traditional bottlenecks between CPU and storage to make all of the 260TB stored in Yellowbrick “hot” at all times. “The biggest benefit we’ve seen from our move to Yellowbrick is performance,” says Augustine. “Our queries actually run to completion instead of getting killed. And we get full utilization of the box 24x7. Even when they’re loading new data into Yellowbrick, it doesn’t really slow the system down.”

- **Faster data loads.** Yellowbrick’s PostgreSQL interface lets Catalina use familiar tools such as Informatica to load data at rates in excess of 100,000 rows per second. ybload, Yellowbrick’s native data-loading utility, delivers even more performance, bypassing the system’s PostgreSQL layer to bulk-load data directly into NVMe flash memory at rates of up to 10TB/hour. “Bulk loads using ybload have been very effective,” says Velez. “During the POT, we loaded a fact table with 303 billion rows in just two days.”
- Increased focus on product development. The analytics team isn’t the only beneficiary of Catalina’s new Yellowbrick system. The data engineering team is also more productive. “In the past, we frequently had to act as Level-3 support for the analytics team, which prevented us from focusing on more strategic initiatives,” says Velez. “Now that we no longer need to triage resource constraints day in and day out, we’ve been able to collaborate on the development of two new products—something that in the past we were unable to do.”
- **Rapid, low-risk migration.** With both Netezza and Yellowbrick based on PostgreSQL, Catalina was able to complete its migration onto Yellowbrick quickly, with minimal risk. “Our transition onto Yellowbrick was easy, because the underlying SQL technologies were the same,” says Augustine. “We didn’t have to change much code, which was another key selling point for Yellowbrick.”
- **Minimal form factor.** Yellowbrick’s minimal form factor—a single 10U rack-mount system—also helped facilitate the company’s migration. “Coming into all this, our data center was already full,” explains Velez. “To facilitate the preceding POT with the other vendor, we had to clear six floor tiles to install the test system, only to find that it

couldn't deliver the performance or compatibility we needed. Being able to satisfy both of those requirements with a single 10U Yellowbrick system was another big win."

- **High reliability.** Compared to the old Netezza environment, which had hundreds of spinning disks, Yellowbrick is delivering better reliability because it has a fault-tolerant design and no moving parts. "From a reliability perspective, Yellowbrick has been extremely stable," says Velez. "Today I don't hear a lot about things that aren't working correctly—and it's a very good sign when my team doesn't need to escalate such issues to my level."

Although the aforementioned technical benefits are impressive on their own, to Augustine, the largest benefit of the company's decision to partner with Yellowbrick has been Yellowbrick's commitment to Catalina's success. "To me, Yellowbrick is a lot more than just great hardware," he says. "What really impressed me was the great service and support. From the very beginning of the project, our partnership with Yellowbrick gave us confidence that what we were doing would work out well, and it has."

Velez echoes Augustine's thoughts on how the two companies have worked together. "Hands-on help from Yellowbrick is a big reason why we completed our migration so quickly," he says. "Unlike some other vendors, who tend to sit back and wait for us to request their assistance, Yellowbrick proactively predicts what we need and then helps make it happen—like how they're now socializing with our users to help fine-tune their queries and optimize resource usage. Our relationship with Yellowbrick has been very effective. It was a big win for us internally, and I'm excited to see what else we can do through our continued strategic partnership."

About Yellowbrick Data

Yellowbrick Data provides the world's fastest data warehouse for hybrid and multi-cloud environments. Enterprises rely on Yellowbrick to power critical business outcomes and get answers to the hardest business questions for improved profitability, better customer loyalty, and faster innovation in near real time, and at a fraction of the cost of alternatives. Yellowbrick offers superior price/performance for thousands of concurrent users on petabytes of data, along with the unique ability to run analytic workloads on-premises, in a private cloud, and/or in any public cloud and manage them in a simple, consistent way—all with predictable pricing via annual subscription.

Learn more at **yellowbrick.com**.

250 Cambridge Avenue, Suite 300, Palo Alto, California 94131 | USA | 1.650.687.0896