DELL Technologies

intel

Powering AI for 30 billion recommendations a day

Taboola relies on powerful, scalable Dell EMC PowerEdge servers with Intel® Xeon® Scalable processors to make billions of AI-driven content recommendations daily, drawing upon continually trained machine learning models

Taboola | Content Marketing | Worldwide

Business needs

Taboola uses sophisticated AI to provide relevant content recommendations on up to four billion web pages with as many as 1.5 billion unique users monthly in 50 countries. Keeping up with the demands requires extraordinary performance and the scalability, automation and security of PowerEdge servers, Intel processors and Kubernetes containers.

Solutions at a glance

- Dell EMC PowerEdge FC640 servers
- Dell EMC PowerEdge R740xd servers
- Dell EMC Integrated Dell Remote Access Controller (iDRAC)
- Intel® Xeon® Silver 4214 and 4114, and Intel® Xeon® Gold 6140 processors
- Kubernetes open-source containers

Business results

- 30 billion personalized content recommendations delivered each day
- Real-time recommendations in as little as 50 milliseconds
- 6x improvement in AI-based inferencing over time
- Continual training of cutting-edge machine learning models

30B personalized content recommendations daily

50ms response time

6x performance increase

150K AI-driven requests processed each second

In Taboola's world, the content finds you—not the other way around.

As the world's largest content recommendation platform, Taboola provides the right recommendation 30 billion times daily across four billion web pages, processing up to 150,000 requests per second. The engine driving this consists of two components: front-end artificial intelligence (AI) for inferencing, which processes and delivers the real-time content recommendations to generate the desired clicks, views and shares; and back-end servers that host cutting-edge deep learning models, which are continually trained using sophisticated neural networks to infer user preferences.

With nine global data centers and only 12 site reliability engineers (SREs), meeting these challenges requires extraordinary computing power and simplified management to attain the maximum performance, agility, scalability and automation to serve clients and users worldwide. Taboola turned to Dell Technologies solutions powered by Intel processors.

"Dell EMC PowerEdge modular servers with Intel® Xeon® Scalable processors were not only a natural choice, but also almost a single contender for our front-end computing," says Ariel Pisetzky, vice president of information technology and cybersecurity.

150K requests a second—an edge in AI inferencing

Taboola knew that it couldn't just continue to add more servers as its global demands grew. The company relies on the PowerEdge modular architecture, using PowerEdge FC640 servers with Intel® Xeon® Silver 4214 and 4114 processors to run its sophisticated homegrown inferencing algorithms based on an open-source TensorFlow machine intelligence framework. This enables its recommendation engine to seamlessly respond to as many as 150,000 requests every second.

Taboola also leverages a Kubernetes Docker container environment that streamlines application development and deployment, and enhances the efficiency of Pisetzky's IT team as they manage more than 10,000 nodes around the globe.

> *"As part of our partnership with Dell, we have a real sharing of information that goes both ways and benefits both companies."*
>
> **Ariel Pisetzky,**
> VP of IT and Cybersecurity
> Taboola

"PowerEdge FX servers with 2nd Gen Intel® Xeon® Scalable processors let us serve more clients faster with better content recommendations," Pisetzky remarks. "To do that with the same investment in hardware is a great win."

The modular architecture of PowerEdge FX ensures Taboola highly scalable business performance to meet the computing needs of its front-end data centers. The modular design of the FC640 servers accommodates significant density of up to 64 servers per rack. Taboola also enjoys the versatility and simplicity necessary to support a "Lego block" approach, allowing Pisetzky to meet changing demands—cost-effectively using the same servers interchangeably as AI inferencing nodes, database servers or storage nodes with very simple configuration changes.

> *"With PE servers and Intel processors, we now get up to six times the performance on our AI-based inferencing...and we believe there a lot more to be gained over time."*
>
> **Ariel Pisetzky,**
> VP of IT and Cybersecurity
> Taboola

Taboola relies on Intel's latest processors on the PowerEdge FC640. Each request coming into a front-end data center runs the AI-driven inferencing algorithms in a unique, ultra-fast process that delivers a relevant recommendation within 50 milliseconds.

Intel® VTune™ Amplifier software helps analyze system performance and identify additional improvement opportunities. Taboola can profile and understand its applications in greater depth—going beyond the block level to review algorithm choices and quickly fix serial and parallel code bottlenecks.

Working with Dell and taking full advantage of the built-in performance acceleration of 2nd Gen Intel® Xeon® Scalable processors—together with the highly optimized Intel® Math Kernel Library for Deep Neural Networking (Intel® MKL-DNN)—Taboola was able to initially enhance its performance by a factor of 2.5x or more. Then, gaining the efficiencies of Kubernetes within the software layer—including the operating system, TCP/IP stack, load balancing and more—Pisetzky's team went much further.

"With PowerEdge servers and Intel processors, we now get up to six times the performance on our AI-based inferencing compared to when we started," states Pisetzky. "This helps reduce our costs, and we believe there's a lot more to be gained over time."

Machine learning models with deep neural networks

To support Taboola's powerful AI engine, Taboola's back-end data centers require servers that run deep learning-based models drawing upon neural networks to accurately and reliably train the Taboola models. Dell EMC PowerEdge R740xd servers with their lightning-fast accelerators were the answer.

"Training is much different from the real-time inferencing we do on the front end," Pisetzky explains. "The demands aren't in terms of response times, but rather the time it takes to process large volumes of data."

He continues, "PowerEdge R740xd servers provide the performance to access our massive data to train our models and push them back to our front-end data center for inferencing. We're using Vertica, Cassandra and MySQL databases across a variety of nodes."

Simplified, automated administration

The Dell EMC Integrated Dell Remote Access Controller (iDRAC) enables Taboola to remotely deploy, update and monitor its Dell servers across its global data centers.

"Automation is critical to our ability to run our business, with a small staff managing our data centers with over 10,000 nodes and growing," Pisetzky emphasizes. "We're heavily invested in iDRAC. Without its advanced remote capabilities, I couldn't efficiently run my data centers."

Taboola uses iDRAC with PowerEdge servers to handle routine administrative tasks agent-free, including server deployment, firmware upgrades and BIOS updates. iDRAC also streams critical metrics for monitoring server performance and analytics—such as CPU errors, memory or power usage and even server operating temperatures—providing Pisetzky and his team with alerts so they can proactively respond to potential problems.

"It would be an impossible task to manage that number of servers with only 12 SREs," notes Pisetzky. "iDRAC running on PowerEdge servers gives us the ability to monitor, deploy and update servers while optimizing our resources."

An ideal high-performance computing platform

In the past, Taboola viewed its infrastructure as just a collection of servers. Today, the company takes a more holistic view of its data centers as high-performance computing (HPC) clusters able to process an enormous number of requests per second.

"We now emphasize rack awareness in our logistics—how much density and bandwidth we have in each rack in our data centers," Pisetzky relates. "Rack awareness allows us to understand where the various compute units are and what different nodes within a data center cluster are running—for better resiliency. Rather than just add servers or racks, we look at everything as a single HPC machine, and reshuffle servers to achieve significant performance improvements and greater cost efficiencies."

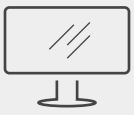
Pisetzky believes there's much more to be gained by further upgrading the utilization of the platform—leading to continuing processing and software improvements in the near future.

"We can always improve," comments Pisetzky. "This includes not only engineering, but also how we deal with the occasional setback—these are opportunities for us to learn and improve."

He concludes, "We've evolved from a startup that buys sporadically from different IT vendors to a company that is truly Dell powered today. As part of our partnership with Dell, we have a real sharing of information that goes both ways and benefits both companies."

"PowerEdge R740xd servers provide the performance to access our massive databases to train our models and push them back for front-end inferencing."

Ariel Pisetzky,
VP of IT and Cybersecurity
Taboola



Learn more about Dell EMC solutions

Contact a Dell EMC Expert

Connect on social

DELL Technologies

intel