DELL Technologies NVIDIA

Training AI models for content recommendation on a massive scale

Taboola relies on powerful, scalable Dell EMC PowerEdge servers with NVIDIA GPUs to continually train AI models that make 30 billion content recommendations daily

Taboola | Content Marketing | Worldwide

Business needs

Taboola needs leading-edge high performance computing systems to train and run sophisticated artificial intelligence models that provide billions of relevant content recommendations every day.

Solutions at a glance

- Dell EMC PowerEdge R740 servers with NVIDIA GPUs
- Dell EMC Integrated Dell Remote Access Controller (iDRAC)
- Docker Kubernetes open-source containers

Business results

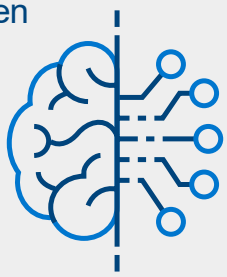
- Providing 30 billion personalized content recommendations delivered each day
- Delivering real-time recommendations in as little as 50 milliseconds
- Enabling continual training of cutting-edge machine learning models
- Achieving 6x improvement in AI-based inferencing over time

Each day, Taboola provides

30 billion
content recommendations

Taboola processes AI-driven requests at a rate of

150,000
per second



Four billion web pages? Taboola has it covered

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 4 billion web pages, processing up to 150,000 requests per second.

The engine driving all of 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. To get there, Taboola turned to Dell Technologies for their artficial intelligence (AI) solutions.

150,000 requests a second — an edge in AI inferencing

Given the scale of its business, Taboola knew that it couldn't simply add more servers as its global demands grew. It found a better solution in the Dell Technologies AI portfolio, according to Ariel Pisetzky, vice president of information technology and cybersecurity for Taboola.

The company now relies on the Dell EMC PowerEdge FC640 servers to run its sophisticated homegrown inferencing algorithms based on an open-source TensorFlow machine intelligence framework. This architecture enables Taboola's recommendation engine to seamlessly respond to as many as 150,000 requests every second. 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.

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

> *"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*

"PowerEdge FX servers let us serve more clients faster with better content recommendations," Pisetzky says. "To do that with the same investment in hardware is a great win."

With up to 64 servers per rack, this modular architecture ensures Taboola highly scalable business performance to meet the computing needs. Taboola also enjoys the versatility and simplicity necessary to support a "Lego block" approach, allowing Pisetzky's team to meet changing demands cost-effectively — using the same servers interchangeably as AI inferencing nodes, database servers or storage nodes with simple configuration changes.

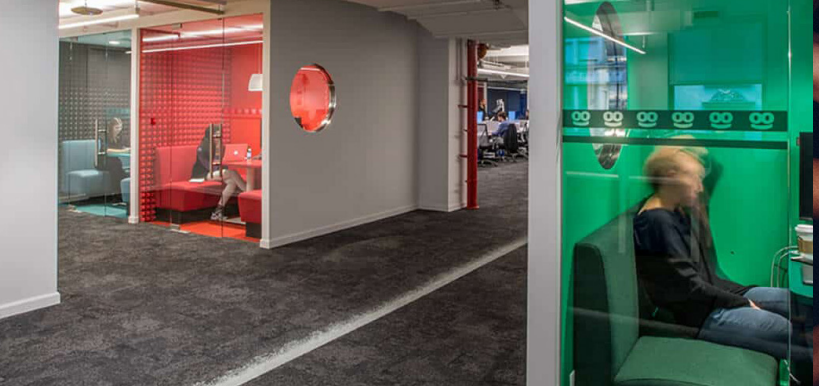
"With PowerEdge servers, we now get up to six times the performance on our AI-based inferencing compared to when we started," states Pisetzky. "This helps reduce our costs, and we believe there's a lot more to be gained over time."

Machine learning models with deep neural networks

To support Taboola's powerful AI engine, the company's back-end data centers require GPU-accelerated servers and neural networks that accurately and reliably train deep learning models to infer user preferences. For this work, Taboola opted for Dell EMC PowerEdge R740xd servers with lightning-fast NVIDIA GPUs.

He notes that training done on the back end is much different from the real-time inferencing done on the front end.

"The demands aren't in terms of response times, but rather the time it takes to process large volumes of data," Pisetzky explains. "PowerEdge R740xd servers with NVIDIA GPUs provide the performance to access our massive data to train our models and push them back to our front-end data center for inferencing."



For this work, Pisetzky and his team use a rich mix of software and database tools, including Vertica, Cassandra and MySQL databases across a variety of nodes.

"To further boost our response times, we have created high performance computing clusters across our data center to take advantage of additional computing power," he says. "Rather than just adding servers or racks, we look at everything as a single HPC cluster. This delivers significant performance improvements and greater cost efficiencies."

Ongoing system optimization

From its CPU-based inference systems on the front end, to the GPU-accelerated training clusters on the back end, to the billions of data points hit in between, Taboola strives to achieve a maximum balance of performance, energy efficiency, and accurate, high-quality results. This work is driven by intensive, ongoing optimization techniques.

"We didn't want specialized hardware just for training when we started out, although to date, we only use this hardware for training," Pisetzky says. "We didn't know how much local storage we would need, but now we do — there has been a lot of learning along the way. We didn't have a use case we could model ourselves after, so much of what we've learned has been the result of careful thought and optimization, from compute and acceleration down to the impacts on the larger storage and network infrastructure."

"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

Pisetzky notes that it takes partnerships to build production AI-ready systems that can meet the demands of billions of hits against Taboola's many AI and analytics systems. For instance, while it's one thing to work with vendors to put together the most balanced training servers, it's another to think about the impact on file systems, storage and networks. With Taboola's close relationship with Dell and NVIDIA, Pisetzky says that his team could see both the big picture from a systems-level view, even across many data centers, while drilling down to the node level for a deep understanding of what might provide the right cost/efficiency/performance balance.

Moving the technology forward

When Pisetzky and his team began the deep learning training journey, they were at the beginning of the AI era and still did not have a keen sense of how their AI systems would interface with the rest of the analytics backbone that comprises the company's multifaceted recommendation engine. He says they started out running training operations on Dell EMC PowerEdge R730 servers with two NVIDIA P100 GPUs per server. Soon after, they saw the value of moving to Dell EMC PowerEdge R740xd servers with three NVIDIA V100 GPUs.

Pisetzky notes that going forward, NVIDIA GPUs will continue to be central to all of Taboola's training operations.

"Making the jump from the NVIDIA P100 GPU to the V100 GPU brought immediate improvements," he says. While his teams at Taboola did evaluate some of the AI chip startups, much of the focus there was on inference, which is a completely different side of the company's deep learning story from a hardware perspective.

"There was never any doubt that we could handle the training and retraining demands we've had over the years with NVIDIA GPUs," he says. "It has been a matter of growing our understanding, as well as adopting newer generations of NVIDIA GPUs over time, which has given us definite performance advantages."

An ideal HPC 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 HPC clusters are 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,” he says. “This includes not only engineering, but also how we deal with the occasional setback — these are opportunities for us to learn and improve.”

In this ongoing progress, Pisetzky expects to work closely with Dell Technologies and NVIDIA.

“We’ve evolved from a startup that buys sporadically from different IT vendors to a company that is truly Dell-powered today,” he says. “As part of our partnership with Dell, we have a real sharing of information that goes both ways and benefits both companies.”