# Putting Recommendation Engines to the Test

The Dell Technologies HPC & AI Innovation Lab is demonstrating how organizations can build better, faster neural networks to drive recommendation engines.


## ABSTRACT

Tests conducted in the Dell Technologies HPC & AI Innovation Lab demonstrated that organizations can parallelize and greatly accelerate the training of the neural networks used to make recommendation engines. These tests showed that with the approaches used in the lab studies, organizations can train very good recommendation engines very quickly, even with large datasets. This paper summarizes this lab project and its key results. This paper also shares learnings, insights and best practices for organizations seeking to build recommendation engines.



March 2021

## TABLE OF CONTENTS






Published in the USA 3/21.

## BUILDING BETTER RECOMMENDATION ENGINES

Recommendation engines power the suggestive retail space. If you go to Netflix® to look for the next movie you want to watch, a recommendation engine will give you suggestions tailored to your interests and past viewing experiences. When you visit a website, you're likely to see ads based on your browsing history and past purchases. If you shop on Amazon®, you will get all kinds of recommendations based on your purchasing history and the purchasing history of other customers with similar interests.

As Amazon explains, "We examine the items you've purchased, items you've told us you own, items you've rated, and items you've told us you like. Based on those interests, we make recommendations." The company notes that its recommendations change regularly based on a number of factors, including "when you purchase, rate or like a new item, as well as changes in the interests of other customers like you."[1]

And as good as today's recommendation engines are, this is only the beginning. At Dell Technologies, we expect the retail world and its use cases for recommendation engines to be dramatically transformed by advances in artificial intelligence over the next five years. To this end, our HPC & AI Innovation Lab is working to make the process of training AI models faster and more efficient on our servers and solutions, in order to build better recommendations engines.

## THE PATH TO A SOLUTION

In our Lab, we are exploring the use and optimization of neural networks to drive recommendation engines. We're focusing on neural networks because they can outperform more traditional machine learning approaches in terms of accuracy of the recommendation made.

That's the upside. The downside is that neural networks are very hard to train. And they are very sequential. That means you have to pick small samples, little snippets out of the data, and pass them through one at a time and then slowly, over the course of days or weeks, train your neural network to make recommendations for you.

Obviously, it is not very efficient to train a neural network if you have to train it one data point at a time, especially when you could have billions of data points that you need to pass through the network. So we took on the task of making the process of training neural networks more efficient on our infrastructure. In this case, we used Dell EMC PowerEdge™ server nodes with Intel® Xeon® Scalable processors in the Zenith cluster in our HPC & AI Innovation Lab.

In the first step in this process, we set our sights on developing a technique for using parallel processing to train this type of neural network, just as we use parallel processing to train other types of AI networks, such as those we use for image classification and language translation. Basically, we wanted to create a neural network that can process more than one experience at a time, so it could do the task that we wanted it to do quickly and efficiently.

1 Amazon, "About Recommendations," accessed March 31, 2021

At this point, we looked to the world of high performance computing to see what techniques had been used there to solve similar types of problems. This investigation led our lab team to an approach called Markov-Chain Monte Carlo (MCMC), which is basically a way of randomly sampling things at very high rates using parallel processing systems.

We then used that approach as the inspiration for parallelizing the training of Restricted Boltzmann Machines (RBMs), which is the type of neural network used to make recommendation engines that rely on collaborative filtering techniques.

For our experiments, we tested the parallelized training of an RBM model using the open source Horovod distributed deep learning framework and a publicly available MovieLens dataset with a billion movie recommendations. For this process, we wrote a custom optimizer, or a custom set of software tools, that we plugged into the Horovod framework to parallelize the training of our RBM neural network.

## THE RESULTS

In our tests with the MovieLens dataset, we were able to go from two full days of training, or a little over 51 hours, to train our model on a single CPU node to 28 minutes to train our model on 140 nodes. This means that we can now process a billion rows of data in less than a half hour and impart all of that knowledge to a neural network so that it can make recommendations.

In our test case, the recommendations are for movies — for example, if you liked this last movie, you will probably also like this next movie. But this engine isn't specific to movies. You could use these methodologies for any kind of rating system. You could do it for consumer products, books, ads or whatever the particular business case might be.

Even better, in our experiments using our technique to train an RBM neural network with a MovieLens dataset, we showed that both strong and weak scaling could be maintained out to 64 compute nodes while producing quality models in accordance with the scale of the dataset used. In future work, we will focus on training our model at an even greater scale using even larger datasets.

## WHY THIS MATTERS

With technology like this, an online retailer could continuously retrain and refine its product recommendation engine to accommodate a very dynamic product catalog and to consider new insights into the changing tastes and buying habits of customers. A customer who returns to a shopping site on a regular basis might then find new recommendations with each visit — and better recommendations, because the more you train the engine, and the more data you give it, the better it understands the customer.

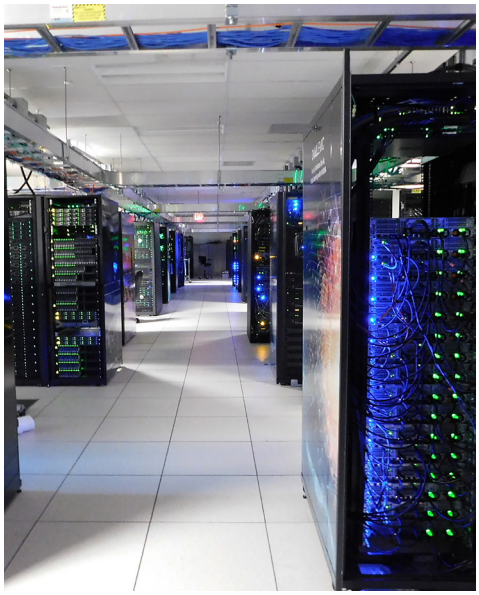
In addition to helping with the follow-up visit, a robust recommendation engine can help retailers gain the coveted additional purchases that come on top of the initial purchases that a customer makes. These unexpected purchases are analogous to the impulse purchases that shoppers make in the checkout line at a grocery store when they add a magazine or a candy bar to the basket. This is what online retailers want to do as well — they want to capture additional purchases, on top of those that prompted the customer's visit.

These unexpected purchases are a bonus for retailers, and robust recommendation engines with continually improving performance and prediction accuracy can help them get there. If we can make those tailored recommendations slightly better, that could translate into significant increases in sales. And if we can make those recommendations much better, that could lead to enormous increases in sales.

## About the lab

The Dell Technologies HPC & AI Innovation Lab encompasses a 13,000 square foot data center in Austin, Texas, devoted to HPC and AI. It houses thousands of servers, three supercomputers, and a wide range of high performance storage and network systems.

The lab is more than world-class infrastructure. Bringing together HPC operational excellence and expertise, it is staffed by a dedicated group of computer scientists, engineers and subject matter experts who actively partner and collaborate with customers and other members of the HPC community. The team gets early access to new technologies, integrates and tunes clusters, benchmarks applications, develops best practices, and publishes white papers.



## TIPS FOR YOUR PROJECT

At the HPC & AI Innovation Lab, we proactively share our learnings, insights and best practices with organizations seeking to capitalize on the technologies for high performance computing and artificial intelligence. With that thought in mind, here are some thoughts on how your organization can get on a path to a successful project.

### START SMALL.

As with all AI projects, the key is to start with the easy stuff. Don't try to start at the end of the journey — where you ultimately want to get to. You might start with a proof of concept for a simple recommendation engine based on a collaborative filtering algorithm that makes predictions based on a user's past ratings for products.

There are a lot of built-in techniques for building recommender systems in packages like scikit-learn, which offers a machine learning library for the Python programming language. Resources like these simplify the work you need to do to build small recommendation engines and other AI applications. And they help accelerate the development of proofs of concept, which are often a critical first step on the pathway to gaining executive buy-in for an AI project.

### KEEP GROWING.

As you gain momentum and get more leeway, collect more data and build more sophisticated models. Keep experimenting and growing until you reach the point where you can get a model into a production environment. And then, as you collect even more data, you can retrain your algorithm to create a more accurate and robust recommendation engine. The idea is to continuously retune, refine and explore more sophisticated techniques to give you even better results.

### LEVERAGE THE AVAILABLE RESOURCES.

Build on the work that others have done. For example, in our research in the Innovation Lab, we work with open source data that your organization can access should you want to try to replicate our results in proofs of concept and other projects.

Similarly, you can draw on the expertise of those who are in the business of helping organizations develop AI applications. At Dell Technologies, for example, we offer workshops that help organizations get on the path to AI, and we have consultants who can help build custom AI models. In addition, the data scientists in our HPC and AI Innovation Lab often work directly with customers to help them select the right platforms, size and configure systems, determine the best software settings and more. The point is, there are teams who can help you jumpstart your project.

### KEEP YOUR EYES ON THE PRIZE.

Building and training a highly scalable, high performance recommendation engine is a complex undertaking. We're trying to create a mathematical model that mimics the human brain. This isn't going to happen overnight. The key is to recognize what's possible, and always work toward the big goal — a continually evolving recommendation engine that will drive your business forward.

## KEY TAKEAWAYS

At Dell Technologies, we expect the retail world and its use cases for recommendation engines to be dramatically transformed by advances in AI in the coming years. To further these advances, our HPC & AI Innovation Lab is working actively to make the process of training AI algorithms faster and more efficient on Dell EMC infrastructure in order to build better recommendations engines.

In our tests in the HPC & AI Innovation Lab, we were able to show that organizations can parallelize the training of the type of neural networks used to make recommendation engines. We further showed that with the approaches we are pioneering, organizations can achieve very good recommendations very quickly, even with large datasets.

This isn't theoretical research. We are demonstrating the foundation for applications that can be developed and deployed today to help retailers, content providers and others build better engines for driving their customers to consumer products, music, video, books and countless other offerings. We believe that a recommendation engine that is easily trained and retrained, that produces high-quality suggestions and scales to huge datasets, can be a significant driver of sales revenue.

## TO LEARN MORE

- For a deeper and more technical dive into the lab's research focused on recommendation engines, see the paper "Parallelized Training of Restricted Boltzmann Machines using Markov-Chain Monte Carlo Methods."
- To learn more about the resources available through the Dell Technologies HPC & AI Innovation Lab, visit hpcatdell.com and delltechnologies.com/innovationlab.
- To explore new HPC solutions for powering AI-driven applications, visit delltechnologies.com/ai.

Leverage Intel®-optimized frameworks, libraries and tools to accelerate processing, increase application performance and reduce development time.

- Intel®-optimized frameworks and tools include TensorFlow, PyTorch, PaddlePaddle, MXNet, Caffe, and OpenVINO,
- Intel® libraries include distributions of Python, Math Kernel Library, MKL-DNN, Machine Learning Scaling Library for Linux®, Data Analytics Acceleration Library and BigDL for Apache Spark™.

Get Intel® optimized frameworks, tools and libraries at software.intel.com/ai

To learn more, visit hpcatdell.com.