Solution Brief
Artificial Intelligence

Use Automated Machine Learning To Speed Time-to-Value for AI

With DataRobot's AutoML platform and the latest Intel® technologies, enterprises can quickly train large datasets and build production-ready machine-learning models

Solution Benefits

- **Fills the data science skills gap.** Empowers a wide range of business users to develop machine-learning models
- **Delivers price/performance for machine-learning training.** Cost-effectively trains multiple models with large data sets simultaneously
- **Builds AI success.** Quickly produces robust, transparent machine-learning models, smoothing the path to AI adoption

DataRobot

Executive Summary

The growing desire to gain business value from artificial intelligence (AI) has created a gap between the demand for data science expertise and the supply of data scientists. DataRobot's automated machine learning (AutoML) platform, running on Intel® architecture, addresses this challenge by automating many tasks needed to develop AI and machine-learning applications.

DataRobot users can build accurate, transparent predictive models within minutes. Data science experts can work more efficiently. Business users can create robust machine-learning models by applying their understanding of enterprise data and business processes. Organizations can apply AI to important business challenges and position themselves for success in the emerging algorithm economy.

Optimized for the latest Intel technologies, the DataRobot AutoML solution delivers unrivaled performance, memory capacity, and scalability for creating, training, and deploying machine-learning models on familiar, cost-effective infrastructure. Using 2nd Generation Intel® Xeon® Scalable processors and Intel® Optane™ persistent memory, organizations can train models on datasets of up to 100 GB. In benchmark tests, a system with Intel Optane persistent memory trained at practically the same speed as a DRAM-only system, depending on the dataset size and training method. The system with Intel Optane persistent memory was projected to train up to a 1.33x larger dataset at the same memory cost compared to a DRAM-only system.[1]

Unlock the Power of Machine Learning

Industry Challenges

- Tedious, time-consuming development process
- Rising demand for machine learning
- Shortage of data science expertise

DataRobot Benefits

Figure 1. Running on Intel® technologies, DataRobot builds AI success by automating the development of robust machine-learning models.

Authors

Lokendra Uppuluri
AI Solutions Architect
Intel Data Platforms

Felix Huthmacher
Engineer
DataRobot

Snehal Adsule
AI Solutions Engineer
Intel Data Platforms

Suleyman Sair
Senior Architect
Intel Architecture, Graphics, and Software

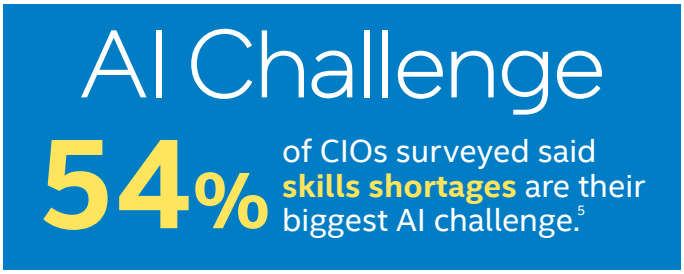
About DataRobot

DataRobot, Inc., was founded in 2012 and received its first patent for data analytics in 2015. Organizations worldwide use DataRobot to empower the teams they already have in place to rapidly build and deploy machine-learning models and create advanced AI applications. The DataRobot platform encapsulates best practices and safeguards to speed and scale data science capabilities while maximizing transparency, accuracy, and collaboration. In April 2019, the company announced that its customers had built one billion models on its Amazon Web Services cloud platform.[2] DataRobot's investors include Intel Capital.

Business Challenge: Fill the Data Science Talent Gap

Organizations of all sizes are eager to apply AI to their toughest challenges and most exciting opportunities. Many recognize machine learning and other forms of AI as powerful ways to gain competitive advantage by deriving fresh insights from their growing data stores. The worldwide AI market, valued at USD 20.67 billion in 2018, is projected to grow to USD 202.57 billion by 2026, a cumulative annual growth rate from 2019 of 33.1 percent.[3]

The rising demand for AI solutions has led to significant shortfalls in AI talent. According to a January 2020 report from TalentSeer, demand for people with AI skills grew 74 percent in each of the four preceding years.[4] In a Gartner survey of 3,000 enterprise CIOs from 89 nations, 54 percent identified skills shortages as their biggest AI challenge.[5]



Along with the shortage of data science experts, machine-learning development is hampered by tasks that are often complex, tedious, and time consuming. As a result, data scientists spend valuable time performing these tasks instead of taking full advantage of their expertise. In addition, many people with knowledge of business data lack the specific skills to create machine-learning models. These issues slow AI development and prevent enterprises from deploying AI as rapidly and extensively as business needs require.

Solution Overview: Automated Machine Learning with DataRobot and Intel

DataRobot uses automated machine learning (AutoML) to help fill the AI skills gap. The DataRobot solution automates and replaces much of the tedious manual work required by traditional data science processes. It empowers data-savvy users at all skill levels to rapidly develop, test, model, and deploy machine-learning algorithms, using best practices and safeguards to help avoid human error.

With DataRobot, users across an enterprise can build on their knowledge of business data to generate advanced machine-learning models, without needing to create code or understand the intricacies of specific algorithms. Data scientists can productively apply their unique expertise to selecting and fine-tuning models. Organizations can quickly create accurate machine-learning models and capture greater value from enterprise data. Figure 2 shows the DataRobot solution's graphical user interface (GUI).

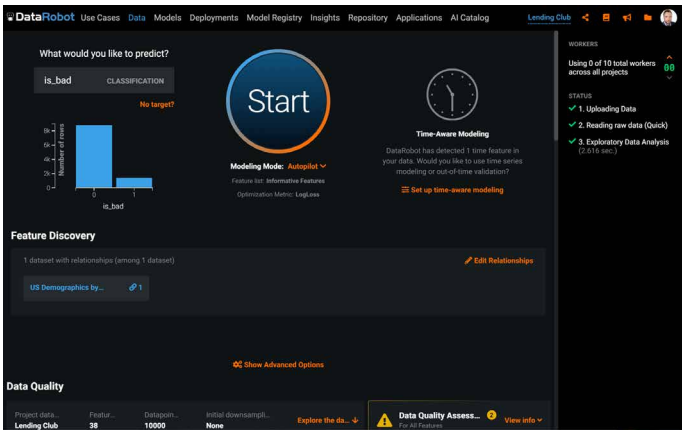


Figure 2. DataRobot's intuitive GUI helps users with business data skills develop machine-learning models without mastering the details of algorithm development, feature training, and other aspects.

DataRobot uses machine learning and Intel technologies to analyze huge data volumes and capture relationships, trends, and patterns that may be too subtle for previous business intelligence and analytics to detect. Users input the relevant data and select the variable they want to predict. DataRobot chooses the most appropriate algorithms and optimizes data preprocessing, feature engineering, and parameter tuning for each algorithm. It builds and trains hundreds of predictive models, ranks and scores the models, and recommends the best model to deploy for the data and prediction target. Instead of spending weeks or months developing and testing a few hand-coded models, users can build and explore hundreds of models and deploy the best-performing model—all within hours.

DataRobot is designed for transparency, so users can understand and explain how models were built and why models made the predictions they did. Built-in visualizations show which types of data have the greatest impact on a model, delivering insights into how individual variables affect the business. The solution uses the performance, scalability, and memory capacity of Intel technologies to build, train, and evaluate machine-learning models, as well as to handle growing datasets and use cases.

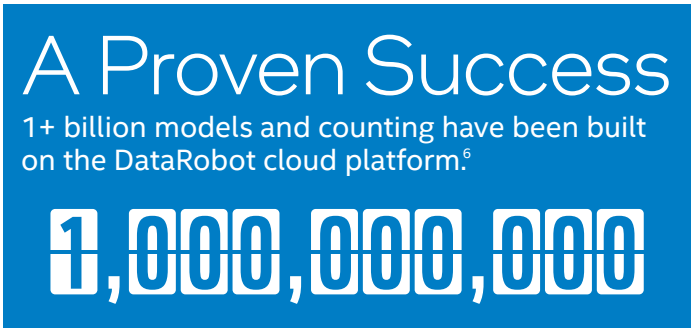
Generating Insights and Value with AutoML

Diverse industries are using the DataRobot AutoML solution to create predictive models that augment human expertise, enhance data-driven decision making, improve efficiencies, and more. Here are some examples:

- **Insurance companies** are targeting areas ranging from underwriting to marketing. They're using machine-learning-enabled insights to optimize pricing algorithms, sharpen risk assessment, and reduce fraudulent claims.
- **Financial technology companies** are predicting fraudulent credit card transactions and creating new investment products. They're strengthening Blockchain security by detecting anomalous behavior within the chain and boosting marketing response rates through improved targeting.
- **Retailers** are gaining new insights into customer spending patterns and shopping behavior across all channels. They apply these insights to better align product mix, promotions, messaging, and media choices to select the right product at the right place and the right time.
- **Manufacturers** are taking the next steps in factory automation and supply chain optimization, driving further productivity gains, cost savings, and quality improvements. Using predictive maintenance and real-time data streams from connected assets, they're optimizing costs and uptime by servicing assets before they have a chance to break down. They're incorporating machine-learning models into the design of next-generation smart products.
- **Public sector agencies** are using machine-learning models with real-time data feeds to predict potential terrorist activities, fraudulent activities, and threats to cyber security. Scalable machine-learning solutions are a key enabler for smart-city functionality that can help improve public safety, traffic efficiency, and more.
- **Healthcare organizations** are augmenting the judgment of clinical care teams with machine-learning models that flag patients at high risk of developing life-threatening infections or requiring costly readmissions. Pharma companies are optimizing the logistics of drug shipments, improving delivery costs and customer service.

Solution Value: Streamlined Path to an AI-Driven Enterprise

The AutoML solution from DataRobot and Intel changes the speed and economics of predictive analytics and provides a rapid path to AI success. This industrial-grade platform addressees the skills shortage by making data scientists more productive. It empowers data professional who have data skills and business savvy to rapidly develop and deploy accurate predictive models. It also addresses the need many DataRobot users have to train models on very large datasets. Organizations can scale their machine-learning efforts to complete more projects, iterate and explore new use cases, and apply AI more broadly throughout their businesses. They can democratize AI and create AI-driven enterprises.



DataRobot is a comprehensive solution that adds value throughout the critical phases of developing and deploying machine-learning models.

- **Ingest data.** DataRobot transforms structured and unstructured data into the specific format each algorithm needs for optimal performance. It follows best practices for data partitioning.
- **Engineer features.** DataRobot develops new features from existing numeric, categorical, and text features. It knows which algorithms benefit from extra feature engineering and which don't, and only generates features that make sense given the data characteristics.
- **Explore and select algorithms.** DataRobot provides access to hundreds of algorithms along with the appropriate pre-processing for users to test against their data. It helps users select the algorithms that make sense for their data and their AI challenge.
- **Train and tune machine-learning models.** DataRobot trains models on the user's data, using smart tuning to optimize the most important hyper-parameters for each algorithm.
- **Find optimal algorithm combinations.** Ensemble or blender models typically outperform individual algorithms. DataRobot finds the optimal algorithms to blend together and tunes the weighting of the algorithms within each ensemble model.

- **Compare models head to head.** DataRobot builds and trains dozens of models, compares the results, and ranks the models by accuracy, speed, and the most efficient combination. Users can explore the models with DataRobot's intuitive GUI and choose which ones to move forward with.
- **Build trust.** To help ensure transparency, DataRobot explains its model decisions, showing which features have the greatest impact on the model's accuracy and the patterns fitted for each feature. It provides explanations to illustrate the rationale behind a specific prediction.
- **Deploy production-ready models**. DataRobot produces production-ready models that users can integrate with enterprise applications with just a few lines of code. Models can be deployed for real-time predictions, batch deployments, scoring on Apache Hadoop, or other methods. Users can develop their own models using R, Python, Apache Spark, MLlib, H2O, and other tools and call the DataRobot library to activate them.
- **Monitor and manage.** Post-deployment, DataRobot makes it easy to compare predictions to actual results and train a new model on the latest data. DataRobot proactively highlights if a model's performance is deteriorating over time.

Solution Architecture for Automated Machine Learning

Powerful Intel technologies help DataRobot optimize performance to simultaneously automate, train, and evaluate multiple machine-learning models and deliver AI applications at scale (see Figure 3).

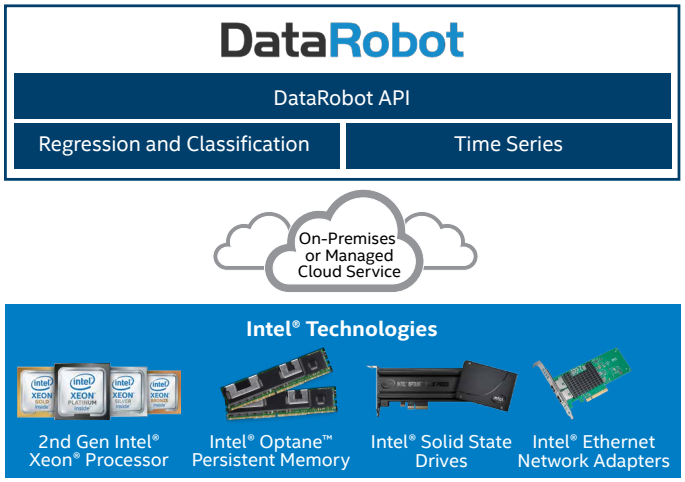


Figure 3. DataRobot takes advantage of the latest Intel® technologies to deliver outstanding performance for automated machine-learning development.

For each new model, DataRobot searches its growing library of thousands of open source machine-learning models. It evaluates possible combinations of algorithms, pre-processing steps, and other attributes to select or construct the most appropriate elements for a given dataset and prediction target. It trains the top models on the user's data and presents the highest performers for users to evaluate. Deployed models can analyze billions of data combinations to deliver new insights and discover signals that may have previously hidden by "noisy" data. The solution can be deployed in an on-premises private cloud or an Amazon Web Services (AWS) cloud managed by DataRobot.

DataRobot integrates easily within the ecosystem of technologies that already exist in the enterprise. These include security and data privacy technologies, data integration and visualization tools, and infrastructure platforms such as Apache Hadoop and SQL databases. Structured and unstructured data can be ingested from data lakes, tables, and other enterprise sources, and users can interact with the system through graphical or programmatic interfaces.

The DataRobot platform includes two independent but interrelated products:

- **Regression and Classification** incorporates a variety of regression techniques—from simple linear regression to statistical classic regression models to more complex techniques such as gradient boosting and neural networks. The platform solves simple binary classification problems as well as complex, multiclass problems with up to 100 categories.
- **Time Series** automates the development of sophisticated models that predict the future values of a data series based on its history and trends. The platform integrates time-series feature engineering to discover predictive signals. It uses both basic and advanced time series models to optimize forecasting accuracy and can visualize insights over time and deploy models to production.

Intel® Technologies for High-Performance, Cost-Effective AutoML Training

Intel's latest generation of data center technologies is built from the ground up for AI workloads. They deliver outstanding performance, scalability, and memory capacity for DataRobot workloads, which are both CPU- and memory-intensive. Organizations can advance their use of AI while maintaining a consistent, cost-effective environment for AI development and model deployment.

- **Intel® Xeon® Scalable processors** provide powerful platforms for data-centric workloads. 2nd Generation Intel Xeon Scalable processors incorporate a built-in hardware accelerator and Intel® Deep Learning Boost with Vector Neural Network Instruction (VNNI) to increase inferencing performance. They also add hardware-enhanced security features to help build a trusted computing foundation. New 3rd Generation Intel Xeon Scalable processors add further performance features, including the industry's first x86 support for Brain Floating Point 16-bit (bfloat 16) for increased training performance.
- **Intel® Optane™ persistent memory** is a new class of non-volatile memory that fills the gap between fast but expensive DRAM and lower-cost, lower-performing NAND SSDs. This innovative memory approaches DRAM performance levels, but at a lower cost per gigabyte. It resides on the memory bus and allows more than 3 TB of memory per CPU socket. In Memory mode, Intel Optane persistent memory can be used transparently as a volatile extension of DRAM.

- **Intel® Solid State Drives** (Intel® SSDs) combine high throughput, low latency, and high endurance to enhance performance for data-bound applications. The Intel® SSD D3-S4510 is a SATA-based SSD optimized for read-intensive workloads. Designed for increased data storage per rack unit, these large-capacity SSDs are available in sizes from 240 GB to 3.8 TB. The Intel SSD DC P4610 is designed with 64-layer, tri-level cell Intel® 3D NAND technology to help data center managers optimize storage efficiency and proficiently manage at scale.
- **Intel® Ethernet Network Adapters XXV710** offer flexible, scalable performance with the ability to auto-negotiate for 1/10/25 GbE connections. These adapters provide intelligent offloads and accelerators to unlock network performance on Intel Xeon Scalable processor-based servers.

Together, these technologies enable enterprises deploying DataRobot to train massive datasets and multiple models simultaneously with high performance.

Benchmarking for AutoML Training

Machine-learning training is a data-intensive task that can require significant amounts of memory. The demands can be especially steep for an AutoML solution such as DataRobot, which trains multiple models simultaneously using the customer's data before ranking them. While models can be trained with varying amounts of data, a larger dataset can help increase model accuracy.

To explore DataRobot's memory requirements, a team from Intel's AI Solutions Group used DataRobot in Autopilot Mode to randomly select and train models from the DataRobot model catalog. We found that training multiple, randomly chosen models required a memory footprint of 6 to 25 times the size of the dataset. The range depended on the model types as well as the percentage of data used for training.

Because of the high ratio of dataset-to-memory footprint, organizations training large datasets may need a large data pool to avoid the performance drain of memory-capacity-bound workloads. Yet, configuring a large, all-DRAM data pool can be prohibitively expensive.

The benchmarking team wanted to see how well Intel Optane persistent memory could address this situation. Could Intel's memory innovation provide DataRobot users with a cost-effective solution for high-performance AutoML training on large datasets?

The team started by training DataRobot in Autopilot mode with a 50-GB test dataset. Then, they selected several models from the leaderboard at random and re-trained them on two systems that differed only in the type of memory in their worker nodes. One used all DRAM, and the other used Intel Optane persistent memory. Figure 4 and Table 1 summarize the benchmarking systems.

We configured both systems with the same memory capacity and compared the performance (training time) of the two configurations. We expected Intel Optane persistent memory to provide somewhat lower performance than the DRAM-only system. However, when we analyzed the training time for the selected models, we found that performance on the system with Intel Optane persistent memory was similar to the all-DRAM system, depending on the model being trained.

Then, using an Intel pricing model, we reconfigured the two systems for the same memory cost instead of the same capacity. Our analysis projected that Intel Optane persistent memory would provide as much as a 1.33x larger dataset capacity for training than the all-DRAM configuration, again depending on the model being trained.

Figure 4 shows these results for training on the Gradient Boosted Trees Classifier for the Intel Optane persistent memory system relative to DRAM-only system. The left half of the chart illustrates performance and performance per dollar for the same memory capacity. The right side shows the projected training dataset capacity and training dataset capacity per dollar for the equivalent memory cost.

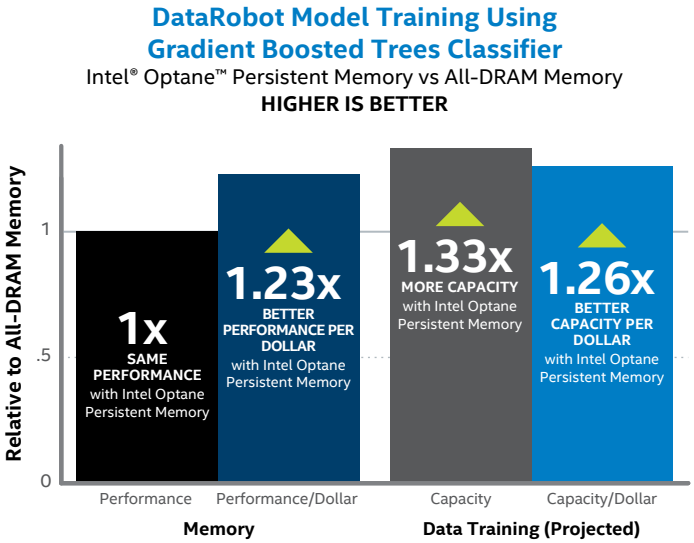


Figure 4. Intel® Optane™ persistent memory provided 1.23x better performance per dollar at the same capacity (left side). It is projected to provide 1.33x more training data capacity and 1.26x better training data capacity per dollar than an all-DRAM configuration.[7]

In summary, our tests demonstrated the following:[7]

- Organizations can train at practically the same speed on a system with Intel Optane persistent memory as on a DRAM-only system, achieving as much as a 1.23x improvement in performance per dollar.
- Organizations are projected to be able to train up to a 1.33x larger dataset at the same cost on a system with Intel Optane persistent memory compared to a DRAM-only system. This is projected to produce an indexed capacity per dollar of up to 1.26x.

Table 1. Typical Configuration for Deploying DataRobot for Training Large Datasets

System Element	Worker Nodes	Management Node	DataRobot Application Node
CPU	• Intel® Xeon® Platinum or Gold processor • 2x 2nd Generation Intel Xeon Platinum 8260 processor, 35.75M cache, 2.40 GHz or 2x 2nd Generation Intel Xeon Gold 6230 processor, 27.5M cache, 2.10 GHz	• 2x 2nd Generation Intel Xeon Gold 6230 processor, 27.5M cache, 2.10 GHz	• 2x 2nd Generation Intel Xeon Gold 6230 processor, 27.5M cache, 2.10 GHz
DRAM Memory	384 GB 2666 MHz, DDR4 ECC RDIMM	192 GB, 2666 MHz, DDR4 ECC RDIMM	192 GB, 2666 MHz, DDR4 ECC RDIMM
Persistent Memory	• 12x 128 GB Intel® Optane™ persistent memory (1.54 TB)	NA	NA
Boot Drive	2x 240 GB Intel® SSD DC D3-S4510, M.2	2x 240 GB Intel SSD DC D3-S4510, M.2	2x 240 GB Intel SSD DC D3-S4510, M.2
Capacity Storage	3x 1.6 TB Intel SSD DC P4610	1.0 TB Intel SSD DC P4510	1.0 TB Intel SSD DC P4510
Network	Intel® Ethernet Converged Network Adapter XXV710-DA2 (10/25GbE)	Intel Ethernet Converged Network Adapter XXV710-DA2 (10/25GbE)	Intel Ethernet Converged Network Adapter XXV710-DA2 (10/25GbE)
Software	• DataRobot v5.2.0 Parcel • Cloudera CDH 5.16.1: – HDFS DataNode – Yarn Node Manager – Spark Gateway – Hive Gateway	• DataRobot v5.2.0 Parcel: – DataRobot Master Service – DataRobot ETL Controller – DataRobot ETL Default – DataRobot ETL Quick Worker Services • Cloudera CDH 5.16.1: – Cloudera Manager – Yarn Resource Manager – HDFS Name Node – Zookeeper – Spark History Server	• DataRobot v5.2.0: – Docker services

Typical Configuration for Deploying DataRobot

Table 1 summarizes a typical system configuration to run DataRobot on premises with Hadoop deployment for training datasets of up to 100 GB. Depending on the size of your enterprise and the number and size of datasets, you may need multiple worker nodes. Please contact your DataRobot representative to learn more about optimal sizing for your training requirements.

Conclusion: AI at Scale

AI has become a core element of business operations and a critical source of competitive differentiation. With DataRobot's AI and AutoML platform and industry-leading Intel technologies, enterprises can address the shortage of data scientists and remove a major roadblock to AI success. They can create production-ready machine-learning models quickly, increasing the productivity of data scientists, scaling their AI development efforts, and applying machine-learning to their biggest business challenges and opportunities.

By taking advantage of AI-optimized Intel technologies, organizations can take advantage of the full power of AutoML. They can deploy powerful training platforms with up to 3 TB of Intel Optane persistent memory per CPU socket. They can also train large datasets at a lower cost than all-DRAM memory configurations. Whether they choose on-premises or cloud-based infrastructure, they can run on versatile, industry-standard architecture with outstanding performance, scalability, and reliability. With DataRobot's AutoML solution and Intel technologies, organizations can focus on AI innovation and creating an AI-driven enterprise.

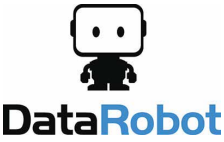
Find the solution that is right for your organization. Contact your Intel representative or visit **intel.com/ai**.

Learn More

You may also find the following resources useful:

- DataRobot
- Intel AI
- Intel Xeon Scalable processors
- Intel Optane Persistent Memory
- Intel SSD DC P4610
- Intel Ethernet Products
- Intel Deep Learning Boost

Solution Provided By:



[1] **Configuration Details for Benchmark Tests:**
DRAM System: Test by Intel as of May 1, 2020. 1-node, 2x Intel® Xeon® Platinum 8260L processors, 24 cores HT On Turbo ON Total Memory 1.54 TB (24 slots/64 GB/2933 MHz), BIOS: SE5 C620.86B.0X.02.0094.102720191711 (ucode:0x500002c), CentOS 7.6.1810, kernel 4.19.94, DataRobot Gradient Boosted Trees Classifier training, score=1.0 (Normalized training time).
Intel® Optane™ Persistent Memory System: Test by Intel as of May 1, 2020. 1-node, 2x Intel Xeon Platinum 8260L processors, 24 cores HT On Turbo ON Total Memory DRAM 384 GB (12 slots/32 GB/2666 MHz) Intel Optane persistent memory 1.54 TB (12 slots/128 GB/2666 MHz), BIOS: SE5C620.86B.0X.02.0094.102720191711 (ucode:0x500002c), CentOS 7.6.1810, kernel 4.19.94, DataRobot Gradient Boosted Trees Classifier training, score=1.03 (Training time relative to DRAM system).

[2] DataRobot Press Release, "DataRobot Celebrates One Billion Models Built on its Cloud Platform," by Libby Botsford, April 16, 2019. datarobot.com/news/press/datarobot-celebrates-one-billion-models-built-on-its-cloud-platform

[3] Technology & Markets, "Artificial Intelligence (AI) Market size, Share and Industry Analysis, 2019-2026." fortunebusinessinsights.com/industry-reports/artificial-intelligence-market-100114

[4] TalentSeer, "2020 AI Talent Report Highlights: Current Talent Landscape & 2020 Market Trends," January 22, 2020. talentseer.com/post/2020-ai-talent-report-highlights-current-talent-landscape-2020-market-trends

[5] Gartner Press Release, "Gartner Survey Shows 37 Percent of Organizations Have Implemented AI in Some Form," January 21, 2019. gartner.com/en/newsroom/press-releases/2019-01-21-gartner-survey-shows-37-percent-of-organizations-have

[6] See endnote 2.

[7] See endnote 1.

Performance results are based on testing as of May 1, 2020, and may not reflect all publicly available security updates. No product or component can be absolutely secure.

Software and workloads used in performance tests may have been optimized for performance only on Intel® microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit intel.com/benchmarks.

Intel technologies may require enabled hardware, software or service activation. Your costs and results may vary.