Business Brief

AI Data Science Considerations for the Data Center Leader

How thinking like a data scientist can help you align artificial intelligence (AI) goals with IT infrastructure priorities.

The infrastructure and data science teams must work together to meet the compute, latency and throughput demands of AI applications, while maintaining data center efficiency.

Artificial intelligence (AI)—usually meaning machine and deep learning training and inference—has been an area of rapid technological advancement over the last decade. The steady innovation within the field of AI has made it possible for companies to build applications capable of analyzing vast or complex data types that were previously only interpretable by humans—such as images, text, speech, and audio. It also makes this possible at a scale and speed that humans could never approach. However, this new class of applications have massively increased compute demands, often requiring strict latency and massive throughput to maintain accuracy.

The infrastructure team must therefore work closely with their data science and AI application development colleagues to ensure these needs are met. At the same time, they must stay focused on larger data center priorities, such as maintaining cost-effective operations and flexibility. To do this, infrastructure architects must promote and foster a more holistic, data-led approach among their app-focused colleagues.

AI workloads run on CPU-based infrastructure

You have already invested in IT infrastructure. So when adding new workloads—AI or otherwise—it's important to ensure that new capabilities optimize your existing resources before you consider investing in new compute. Your current Intel® architecture provides:

- **Flexibility:** The multi-purpose nature of the Intel® Xeon® platform means it supports a wide range of workloads, including machine and deep learning (see figure 1). With extensive software optimizations across machine and deep learning, and with built-in Intel® Deep Learning Boost inference acceleration, AI on the CPU is faster than ever before.
- **Efficiency:** Map your AI workload needs to current system utilization to identify where and how to best provision for them. You can often provision additional usage for AI workloads on your spare capacity with existing CPUs. Most machine learning and deep learning inference run on CPUs today, and in many cases CPUs are ideal for deep learning training. Use optimizations of common AI software frameworks (such as TensorFlow and PyTorch), libraries and tools to help ensure performance per watt and performance per dollar remains strong and PUE ratio is as close to 1 as possible.
- **Scalability:** Scale your AI training workloads easily across multiple Intel® data center or edge nodes as demand requires. Maintain efficiency as you scale with a system design that takes into account network and memory, using technologies like Intel® Ethernet 700 series, and Intel® Optane™ technology respectively. This enables you to make the most of your existing hardware investment while scaling your deep learning workloads for higher throughput across even the largest data sets. Facebook has demonstrated this approach well.

	RECOMMENDATION ENGINES	CLASSICAL MACHINE LEARNING	RECURRENT NEURAL NETWORKS	MODELS USING LARGE DATA SAMPLES	OTHER REAL-TIME INFERENCE	TRAINING ON SPARE CYCLES
PURPOSE	Recommend ads, search, apps etc	Gain insights from data	Speech recognition	Medical images, seismic exploration, 3D environments	Image recognition, speech recognition, natural language processing	Any purpose
CLASS	Multilayer Perceptron (MLP)	Regression, classification, clustering, etc	Recurrent Neural Network (RNN)	Convolutional Neural Network (CNN)	Multiple	Any class
CPU BENEFIT	Training and inference. Larger memory for embedding layers	Faster cores for large datasets and difficult-to-parallelize algorithms	Real-time inference. Faster cores for sequential, difficult-to-parallelize data	Training and inference. Larger memory required	Faster cores for small batches which are difficult to parallelize	Data center capacity

Figure 1. Your existing Intel® technology-based infrastructure can support a wide range of AI use cases and workloads

Intel® technology powers AI performance

2nd generation Intel® Xeon® Scalable processors deliver scalable performance for a wide variety of artificial intelligence (AI) applications, along with an average **42 percent performance per dollar improvement** over the prior generation[1].

Intel and Google engineers have been working together to optimize TensorFlow, a flexible open-source AI framework, for the Intel® Xeon® platform, providing **up to 3.75x inferencing speed** up when using Intel® Deep Learning Boost[2].

Implementing AI: The data scientist's perspective

Your data science colleagues may come at things from a different point of view. They will likely associate the quickest model development time for AI with a GPU-based hardware platform, which can deliver very high throughput for some deep learning training workloads.

However, it very much depends on the AI workload, data type, and requirements in question. Work with your data scientists and help them to keep an open mind when it comes to what might be the best platform. Start by characterizing the workload itself, asking your data scientists questions like:

- What types of models do you need to run?
- What size are each of these models (number of parameters)?
- How large is the data used for modelling?
- What's the typical batch size of each model?
- What's the maximum number of live activations?
- What's the arithmetic intensity of each model?
- What are your latency constraints?

Understanding these workload characteristics will enable you to determine the underlying compute requirements for your data scientists' AI workloads. If they are running dedicated deep learning training, using dedicated hardware acceleration may make sense. In most other cases, running the AI workloads on your existing compute infrastructure is likely to be your best option for achieving a happy compromise with your data scientists. It will enable you to deliver the acceleration and performance they demand, while helping you meet your efficiency, scalability and flexibility objectives.

Think also about how any projects will need to scale over time. Data scientists often experiment with new algorithms and workloads on a relatively small scale, using a small number of GPUs. However, when these projects need to be deployed into production on enterprise-scale, few IT teams have the budget to provide the same platform. Making sure your data scientists are running their experiments on a platform comparable to that of a large-scale deployment will help them avoid unnecessary challenges.

A holistic view helps lay shared foundations for AI

As a next step, focus on making sure your data science team is fully on board with your proposed approach, and be ready to address any doubts or objections they may raise. These may often arise as a result of focusing on application acceleration in isolation. As the person with a full view of the IT infrastructure, you can offer valuable insight on how taking a broader view of the application within the IT environment can present opportunities for optimization that the data scientist may not initially have thought of.

Talk to them about the full analytics/AI pipeline, from data ingestion to consumption (see figure 2). Before any AI algorithm can be run on a piece of data, it must be captured, ingested, and cleaned up. Most of that process today runs on the Intel® Xeon® platform, meaning that running that last AI step on the same platform reduces complexity and saves development time across the whole pipeline.

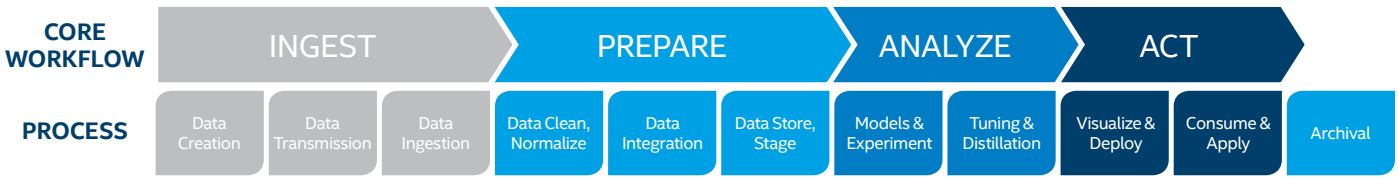

Figure 2. The data pipeline underpins every AI application. Run yours using existing, CPU-based infrastructure

This integration is enhanced further for organizations that have clusters built on Spark and Hadoop, as Intel's optimizations for these clusters increase their efficiency when running machine and deep learning workloads. Tools like Analytics Zoo, an end-to-end open source analytics and AI platform that also integrates with Intel's Spark and Hadoop optimizations, can help you scale these workloads seamlessly with full control and visibility as demand increases.

Build your partnership for long-term success

By partnering with your data science colleagues you can build a solid foundation for your AI strategy. If and/or when you're ready to build on that foundation, Intel has a flexible edge-to-cloud AI acceleration portfolio that meets your evolving power, performance and memory needs. Intel also has a broad ecosystem of solution providers in the Intel® AI Builders community who can help you get started.

Alongside the data scientists in your organization, you play a critical role in the success of the crucial AI initiative, so it's important for you keep the lines of communication open. Creating an understanding of each other's priorities and concerns is the best first step.

Learn More

- **Business Brief**
  Accelerating AI Adoption
- **Solution Brief**
  AI-Driven Solutions Improve Healthcare Access and Quality (JLK)
- **Case Study**
  Kongsberg Maritime's Marine Solutions
- **Solution Brief**
  Using AI to Analyze Fashion and Luxury Market Performance (IFDAQ)
- **Solution Brief**
  AI-Powered Next-Generation Contact Centers

[1] **36% More Estimated Performance and 42% More Estimated Performance/Dollar: Geomean of SPECrate®2017_int_base(est), SPECrate®2017_fp_base(est), STREAM Triad, and Intel® Distribution for LINPACK* Benchmark Across Ten New 2-Socket 2nd Gen Intel® Xeon® Gold Processors Vs. First Generation.** 2nd Gen Intel® Xeon® Gold R processors: 1-node, 2x 2nd Gen Intel® Xeon® Gold processor (62xxR/$$) on Intel Reference platform with 384GB (12 slots / 32 GB / 62xx@2933,52xx@2666) total memory, ucode 0x500002c, HT on for all except off for STREAM (GB/s), LINPACK (GFLOPS/s), Turbo on, with Ubuntu19.10, 5.3.0-24-generic, 6258R/$3950: SPECrate®2017_int_base(est)=323, SPECrate®2017_fp_base(est)=262, STREAM=224, LINPACK =3305; 6248R/$2700: SPECrate®2017_int_base(est)=299, SPECrate®2017_fp_base(est)=248, STREAM=224, LINPACK =3010; 6246R/$3286: SPECrate®2017_int_base(est)=238, SPECrate®2017_fp_base(est)=217, STREAM=225, LINPACK =2394; 6242R/$2529: SPECrate®2017_int_base(est)=265, SPECrate®2017_fp_base(est)=231, STREAM=227, LINPACK =2698; 6240R/$2200: SPECrate®2017_int_base(est)=268, SPECrate®2017_fp_base(est)=228, STREAM=223, LINPACK=2438; 6238R/$2612: SPECrate®2017_int_base(est)=287, SPECrate®2017_fp_base(est)=240, STREAM=222, LINPACK =2545; 6230R/$1894: SPECrate®2017_int_base(est)=266, SPECrate®2017_fp_base(est)=227, STREAM=222, LINPACK =2219; 6226R/$1300: SPECrate®2017_int_base(est)=208, SPECrate®2017_fp_base(est)=192, STREAM=200, LINPACK =2073; 5220R/$1555: SPECrate®2017_int_base(est)=257, SPECrate®2017_fp_base(est)=220, STREAM=210, LINPACK =1610; 5218R/$1273: SPECrate®2017_int_base(est)=210, SPECrate®2017_fp_base(est)=188, STREAM=199, LINPACK =1290, test by Intel on 12/25/2019. First Gen Intel® Xeon® Gold processors: 1-node, 2x Intel® Xeon® Gold processor (61xx/$$) on Intel Reference platform with 384GB (12 slots / 32 GB / 61xx@2666,51xx@2400) total memory, ucode 0x500002c, HT on for all except off for STREAM (GB/s), LINPACK (GFLOPS/s), Turbo on, with Ubuntu19.10, 5.3.0-24-generic, 6152/$3655: SPECrate®2017_int_base(est)=224, SPECrate®2017_fp_base(est)=198, STREAM=200, LINPACK =1988; 6148/$3072: SPECrate®2017_int_base(est)=225, SPECrate®2017_fp_base(est)=198, STREAM=197, LINPACK =2162; 6146/$3286: SPECrate®2017_int_base(est)=161, SPECrate®2017_fp_base(est)=175, STREAM=185, LINPACK =1896; 6142/$2946: SPECrate®2017_int_base(est)=193, SPECrate®2017_fp_base(est)=176, STREAM=185, LINPACK =1895; 6140/$2445: SPECrate®2017_int_base(est)=202, SPECrate®2017_fp_base(est)=183, STREAM=188, LINPACK =1877; 6138/$2612: SPECrate®2017_int_base(est)=189, SPECrate®2017_fp_base(est)=195, STREAM=189, LINPACK =1976; 6130/$1894: SPECrate®2017_int_base(est)=172, SPECrate®2017_fp_base(est)=165, STREAM=185, LINPACK =1645; 6126(proj)/$1776: SPECrate®2017_int_base(est)=141, SPECrate®2017_fp_base(est)=157, STREAM=170, LINPACK =1605; 5120(proj)/$1555: SPECrate®2017_int_base(est)=148, SPECrate®2017_fp_base(est)=148, STREAM=159, LINPACK =924, 5118/$1273: SPECrate®2017_int_base(est)=134, SPECrate®2017_fp_base(est)=132, STREAM=149, LINPACK =818, test by Intel on 2/18/2020.

[2] Up to 3.75x improvement with AI Inferencing Intel Select Solution. The solution was tested with KPI Targets: OpenVINO/ ResNet50 on INT8 on 02-26-2019 with the following hardware and software configuration:

Base configuration: 1 Node, 2x Intel® Xeon® Gold 6248; 1x Intel® Server Board S2600WFT; Total Memory 192 GB, 12 slots/16 GB/2666 MT/s DDR4 RDIMM; HyperThreading: Enable; Turbo: Enable; Storage(boot): Intel® SSD DC P4101; Storage(capacity): At least 2 TB Intel® SSD DC P4610 PCIe NVMe; OS/Software: CentOS Linux release 7.6.1810 (Core) with Kernel 3.10.0-957.el7.x86_64; Framework version: OpenVINO 2018 R5 445; Dataset:sample image from benchmark tool; Model topology: ResNet 50 v1; Batch Size: 4; nireq: 20. The solution was tested with KPI Targets: TensorFlow/ ResNet50 on INT8 on 03-07-2019 with the following hardware and software configuration:

Base configuration: 1 Node, 2x Intel® Xeon® Gold 6248; 1x Intel® Server Board S2600WFT; Total Memory 192 GB, 12 slots/16 GB/2666 MT/s DDR4 RDIMM; HyperThreading: Enable; Turbo: Enable; Storage(boot): Intel® SSD DC P4101; Storage(capacity): At least 2 TB Intel® SSD DC P4610 PCIe NVMe; OS/Software: CentOS Linux release 7.6.1810 (Core) with Kernel 3.10.0-957.el7.x86_64; Framework version: intelaipg/intel-optimizedtensorflow:PR25765-devel-mkl; Dataset: Synthetic from benchmark tool; Model topology: ResNet 50 v1; Batch Size: 80

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 teWsts 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 www.intel.com/benchmarks.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

0620/JL/CAT/PDF ♻ Please Recycle 343569-001EN