EASE YOUR ORGANIZATION INTO AI

A practical guide to building an insights-driven business

CONTENTS

EXECUTIVE SUMMARY

As their colleagues across the business clamor for more and deeper insights, pressure is mounting on IT leaders to offer artificial intelligence (AI) capabilities. As these demands become more complex, AI techniques like machine learning and deep learning can help turn information like data, images and text into insights that inform, automate and enhance business decisions, and ultimately drive business value.

Implementing AI can't be done overnight. It will be a journey of months, maybe years (see figure 1). It starts with building a solid **foundation**, including identifying and prioritizing opportunities, considering impacts beyond return on investment (ROI) (such as ethical and legal regulations), and then organizing people for success.

The often most-underestimated step, is creating your **data pipeline** and aligning it to overall data strategy: determining how you plan to ingest, store, pre-process and ultimately manage it. Once your data is in order, you can **develop** your solution by setting up your technology, then doing development, testing and documentation of results.

Finally, you're ready to **deploy**, which means architecting a solution, implementing it, scaling it as demand grows, and iterating on it with a continuous improvement process.

This guide explores each stage of the AI journey and questions that you as an IT leader can ask yourself, your team and your colleagues to help you along the way.

HELPFUL DEFINITIONS:

Data Analytics: The use of data tools (AI, machine learning, deep learning, statistics, optimizations, simulation, etc) to exploit diverse data to gain business value.

Artificial intelligence (AI): A subset of Data Analytics, AI is the ability of machines to learn from data, without explicit programming, in order to perform functions typically associated with the human mind.

Machine learning (ML): An approach to AI that includes the development and application of algorithms to build and continually improve models over time with more data and input. Machine learning is typically used for solving problems with pre-defined features, for applications such as recommendation engines, quality control/yield, etc. There are three types of ML: supervised learning, unsupervised learning, and reinforcement learning.

Deep learning (DL): A subset of machine learning that uses layered neural networks that learn from vast amounts of data to solve problems. The advantage of DL is the algorithms extract features (vs requiring pre-defined features that are difficult to engineer as in ML) in such applications as computer vision, and speech recognition.

FOUNDATION	**IDENTIFY** Identify business needs/ problems to be solved internally and assess business value of each one	**PRIORITIZE** Prioritize projects based on business value and cost to solve	**CONSIDER** Consider ethical, social, legal, security and other risks and mitigation plans prior to kickoff	**ORGANIZE** Organize internally to get buy-in, support new development philosophy and grow developer talent
DATA	**INGEST** Ingest data using a software tool among the many that run on Intel® Xeon® Scalable processors	**STORE** Store data in object or block storage, databases and/or data lakes	**PROCESS** Process data by performing cleanup and integration using common software tools that run on Intel Xeon Scalable processors	**MANAGE** Manage data via a framework for distributed computation on CPU infrastructure
DEVELOP	**SETUP** Setup compute environment, adding deep learning acceleration if required for training	**MODEL** Model development through training a deep neural network using an Intel® processor-optimized deep learning framework	**TEST** Test the deep learning model using a control data set to determine if accuracy meets requirements	**DOCUMENT** Document the code, process, and key learnings for future reference
DEPLOY	**ARCHITECT** Architect artificial intelligence (AI) deployment with Intel® AI Builders	**IMPLEMENT** Implement AI in production environment	**SCALE** Scale to more sites and users as demand grows	**ITERATE** Iterate on the models with new data over time

Figure 1: The artificial intelligence (AI) Journey

CHAPTER ONE: FOUNDATION

IDENTIFY AND PRIORITIZE BUSINESS CASE

The first step in building an AI foundation is to identify the business problems or questions that need to be solved, and determining whether an AI application is the right way to solve them. As with any new project, it's important to start with a clear idea of how your initiative aligns with business goals and strategy, and what value you can reasonably expect it to generate. AI should be used as the means of answering a compelling question or discovering fresh insight, but it is not a panacea for every yet-to-be-defined business woe.

Typically, an application of AI will help address one of three areas: increasing revenue; driving efficiencies and/or cost savings; and innovating or diversifying your offering. Be as specific as you can about your business goal, and make sure you are asking the right questions to help you define it: where are the greatest opportunities to drive value in the business? Are you being specific enough in defining them? What does success look like, and in what timeframe? What ethical, social, legal, security and other risks need to be mitigated? By addressing these questions up-front, you can more effectively prioritize projects based on business value and cost to solve. This critical thinking guide can help you navigate these discussions, with step-by-step guidance on how to evaluate your business goals for AI initiatives.

> As IT leadership, you can help ensure business goals, priorities and expectations align with data and infrastructure capabilities and roadmap. Coach your line-of-business colleagues to achieve the results they need.

ORGANIZE STAKEHOLDERS

These discussions should be led by the line(s) of business (LoB) concerned, but IT must be involved even at this early stage to help the business understand what's possible and to ensure the expectations being set are realistic and achievable. The type of insights you aim to extract from your data will depend on the needs of the business and the relative analytic maturity of your organization.

It is critical to secure commitment and input from stakeholders around the business – both at this early stage and throughout the project. Bringing together the right combination of people from the outset will help make it a success (see figure 2). This may include (but not be limited to) data scientists and engineers, developers, LoB users and managers, domain experts, insight users, executive leadership, and traditional IT roles.

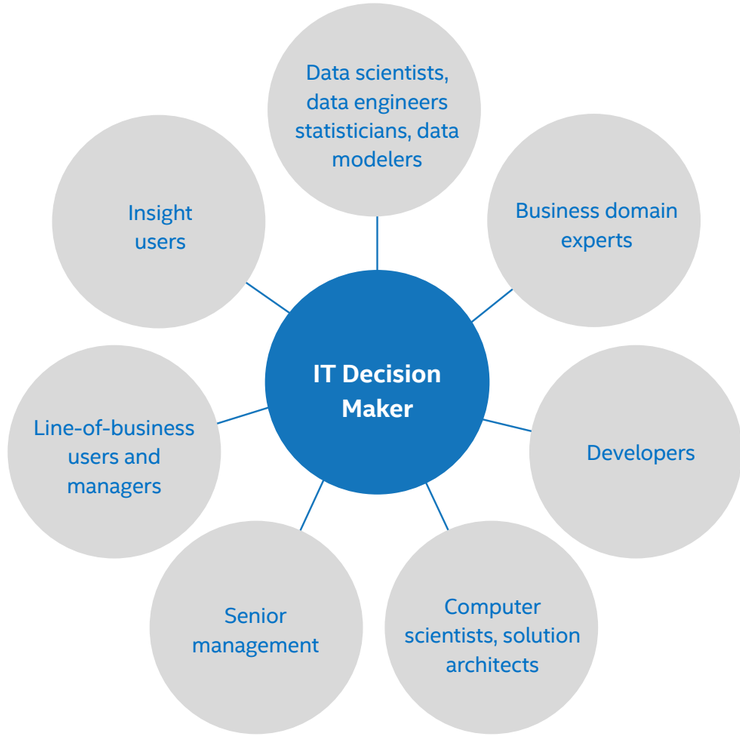


Figure 2: IT can bring together the artificial intelligence (AI) project stakeholders from across the business

Getting everyone bought in up front is essential for long-term success. Work as a group to define roles and responsibilities for everyone. Ensure everyone understands and respects the expertise each person brings to the table to ensure collaboration is as smooth as possible. When stakeholders are active in the development of AI, they can trust the results it gives them and take action. Otherwise, you risk them not using it and the whole project becoming a purely academic exercise with no business value. Help ensure business users are clear on how they will use the insights they're aiming to gain. It's important they are able to consume the output of any analytics or AI quickly and easily (for example through a visualization dashboard). They must also have a clear understanding of where the data comes from and how it is being used in order to effectively explain these data-driven insights to end users.

Achieving this buy-in up front will help foster a data-driven culture that respects and prioritizes the use of data in strategic decision making. As with any cultural change, this can be tricky, so having a team that includes influencers who can help propagate the thinking and behaviors you need across the business is important. Lay the foundations for long-term success from the outset, by working into your plans processes and methodologies that will help you gather feedback regularly – both from consumers of the insights as to the value they deliver to the business, and from data scientists/engineers, solution architects and IT teams around the performance of the algorithms, efficiency of data management, or the need to tweak the underlying infrastructure for example. In this way you can help drive continuous improvement over time.

DECIDE WHICH AI APPROACH YOU NEED

With your business case clearly articulated, you can then start to define the data that will inform and enable the required insights. Once this data is identified then start to work with your data scientists, who will determine the most suitable analytics and AI approaches (see figure 3). Your data scientists can select the methodologies and tools that are most suited to the type of data they are using (see subsequent chapters for more on these). For example, if your data is mostly structured, with pre-defined features, machine learning methods work best for extracting insight and providing transparency. Meanwhile, if your data is unstructured data (e.g. video, images, audio, or text), deep learning is a more suitable approach to extract the necessary features (objects from video and images or words from audio and text). Most organizations tend to begin with machine learning or statistical learning approaches, due to the plethora of structured data and to ensure relative transparency in how the insights are quantified. As the business adopts more model insights, it can introduce more layering of models or enhance the way they feed each other to develop AI solutions. For example, in manufacturing,

Which approach is right?

A large manufacturer uses data to improve their operations, with each challenge using a different approach to deliver maximum business value at a low cost

CHALLENGE	BEST APPROACH	APPROACH	EXAMPLE OUTPUT (ANSWER)
How many widgets should we manufacture?	Analyze historical supply/demand	Analytics/ Business Intelligence	10,000
What will our yield be?	Algorithm that correlates many variables to yield	Statistical/Machine Learning	At current conditions, yield will be at 90 percent with 10% loss expected
Which widgets have visual defects?	Algorithm that learns to identify defects in images	Deep Learning	Widget 1003, Widget 1094…
Can my robotic arm adapt to condition in real-time?	Algorithm that acts and adapts based on feedback	Deep Learning	An intelligent robot that adapts to certain environmental conditions

Figure 3: Choose the right artificial intelligence (AI) approach for your business challenge

if the models used indicate lower yield, it can communicate to the statistical model to update the supply model to update the amount of parts to manufacture. A defect deep learning model can help identify and categorize defects. This model can be paired with the yield model to identify what machines are not performing, sending a technician to calibrate or fix the machines to correct the issue, before it impacts the overall yield metrics.

Remember that AI capabilities are evolving quickly, so stay abreast of other projects or emerging approaches that might help you take your insights to the next level over time. For example, a hospital may see a lot of cases of bronchitis and so apply machine learning to help identify other potential cases within its customer base. It may then be able to share these learnings (without sharing the sensitive data underneath) with other hospitals in its group, so they also have the ability to identify bronchitis cases early without having seen any previously.

THE LONGER READ: BUILDING A STRONG FOUNDATION FOR AI:

- Learn more by reading: How to Build a Great AI Team, Forbes
- Guide your team through your business case evaluations: Analytic Topic Evaluation Guide
- For more on the stages of analytics and AI maturity: Getting started with advanced analytics
- To help you build a data-driven culture: 5 tips to create a data-driven culture at your company, CIO Magazine

CHAPTER TWO: DATA

A good working definition of data strategy is: a central, integrated concept that articulates how data will enable and inspire business strategy[1]. A data strategy should outline:

- The organization's vision for data
- How the organization will execute desired data activities
- How the organization will drive data adoption and use
- The sequence of steps and timings by which the organization will execute its proposed activities
- How the company will monetize its data
- The most efficient and legally compliant way for data to be stored, shared and utilized

Naturally, this is about much more than AI, and AI plans will form a subset of the overarching data strategy. It's important to run AI initiatives within the context of a broader data environment, but also bear in mind current and future AI needs as you set your data strategy moving forward. In doing this, focus on the four key stages of the data pipeline – ingest, store, process and analyze – making sure you have the resources in place to optimize each stage. One of the most critical steps is data integration (see figure 4). Where siloes exist, plan to merge teams and data alike. Where there are data gaps, identify external sources that can help fill them.

> Understand your data characteristics and needs at each stage of its lifecycle. Work with solution architects, hardware and software specialists to map a data pipeline optimized for your artificial intelligence (AI) needs.

Work with your solution architects to understand your data's characteristics, such as source, variety, location, frequency, value and cost to manage. The combination of its needs across these vectors will help you create the most suitable infrastructure for your AI goals, which we'll explore in detail later.

INGEST

The first stage is to **ingest** your raw data, using an approach that is appropriate to its size, source, structure and latency. For example, streaming data is made up of a continuous flow of small, asynchronous messages, which are delivered without expecting a reply. This type of data may be used for real-time analytics use cases, which are becoming increasingly common (for example cybersecurity monitoring). The need for immediate insights and action in these scenarios demands distributed compute and memory (in virtual machines (VMs) and/or at the edge) to enable real-time action. In particular it needs high compute performance at low power. Intel® CPUs such as the mainstream, light Intel® Iris® Graphics, which are optimized for low-power, high-performance edge workloads can help meet these requirements.

Other types of scenario rely on batch data, which tends to consist of large numbers of files that are transferred and stored in bulk, or in relational or NoSQL* databases. For example, a manufacturer wishing to analyze trends and patterns in its supply chain over time may choose to use batch data for its AI workloads. The data in this case can be located on-premises or on other cloud platforms, and ingesting it requires high aggregate bandwidth between the sources and the target. As with ingestion at the edge, high-performance CPUs are essential in ingesting batch data, with the added requirement of massive scalability as data volumes grow. The latest data center technologies from Intel, including Intel® Xeon® Scalable processors, Intel® Optane™ DC persistent memory and Intel® Optane™ DC Solid-State Drives, are architected from the ground up for this type of challenge, with support for a new, bigger tier of memory and extreme performance for fast-growing analytics and AI workloads.

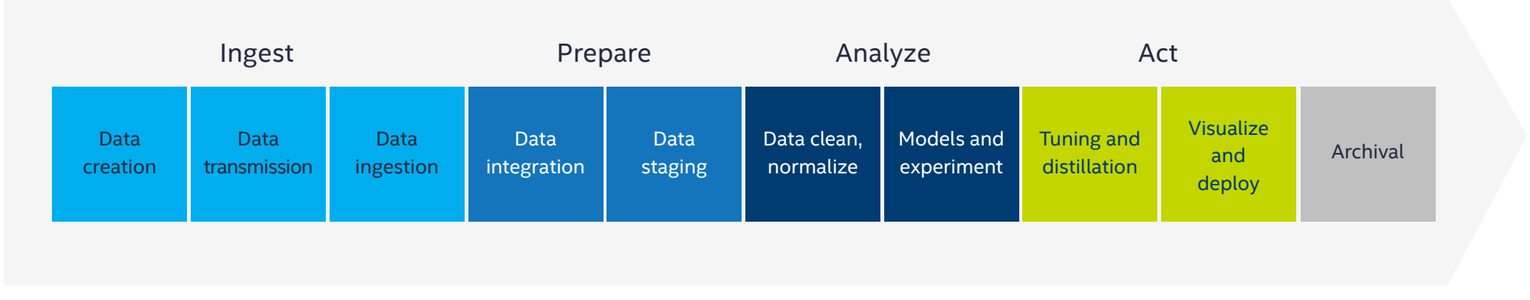


Figure 4: The stages of the data lifecycle, or data pipeline

PREPARE

Once the data has been captured, it must be prepared and stored. Again decisions made here will depend upon the characteristics of the data you are using. Object stores are used for storing structured and unstructured data like logs, database backups, files, images, and videos. For this type of storage requirement, consider QLC NAND devices for cost-effective, high-volume storage, and solid-state drives (SSDs) for caching.

Intel Optane DC SSDs remove performance bottlenecks by delivering high throughput at low queue depths, ideal for caching of temporary data in big data workloads. They deliver improved read latency performance for faster and more consistent time to data analytics insight. When the Intel® Optane™ SSD DC P4800X is used with Intel® Memory Drive Technology, the result is up to 3x faster time to completion in Apache Spark* by cost effectively expanding memory pool capacity compared to an all-DRAM system[2]. Used with Intel Memory Drive Technology, the Intel Optane SSD DC P4800X also enables efficient data set expansion, working seamlessly as system memory, and delivering near-DRAM performance with the cost and capacity benefits of storage.

For databases, a SQL relational database management system (RDBMS) (eg SAP HANA*, Microsoft SQL Server*, Oracle DB*, MySQL*) is generally suited to online transaction processing (OLTP) workloads with structured data that require ACID* compliance, like financial transactions, customer order tracking or checking user credentials. On the other hand, **NoSQL** databases (eg Redis*, Cassandra*, HBase*) tend to support terabyte- to petabyte-scale OLTP workloads that support real-time and/or large-scale analytical workloads, such as real-time app data, Internet of Things (IoT) sensor data or advanced analytical and machine learning tasks. These accelerated workloads demand CPUs for faster processing, and 3D NAND or persistent memory to support real-time operations.

Delivered in conjunction with the 2nd generation Intel® Xeon® Scalable processor, Intel Optane DC persistent memory is a revolution in memory and storage technology, offering a unique combination of affordable large capacity and data persistence. It enables you to keep more hot data closer to the processor, achieving greater cost/performance benefits (and so lower total cost of ownership (TCO)) from your storage. Designed for data-intensive workloads, it delivers breakthrough restart times for in-memory databases and reduced wait times when fetching large data sets from system storage.

Finally, data warehouses or data lakes based on platforms like Cloudera* are common, and most organizations tend to hold a mix of data lakes and the data islands described above. Data lakes can be built on physical servers, or in the cloud or on VMs and containers, but should only be created where they support a specified business need. We are seeing an increasing trend towards the decentralization of compute and storage. This model works well for real-time, edge-based analytics workloads, when supported by a clear storage strategy. Set criteria for tiering your data depending on its urgency and usefulness, and ensure you choose the most cost-efficient memory or storage technology for each tier (see figure 5).

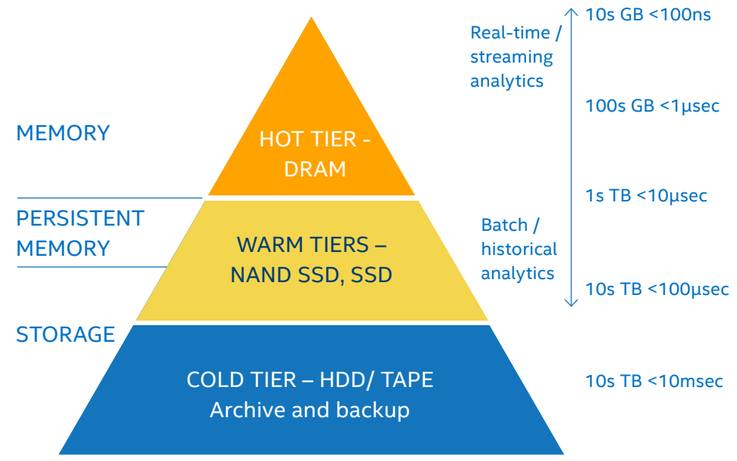


Figure 5: The storage/memory continuum

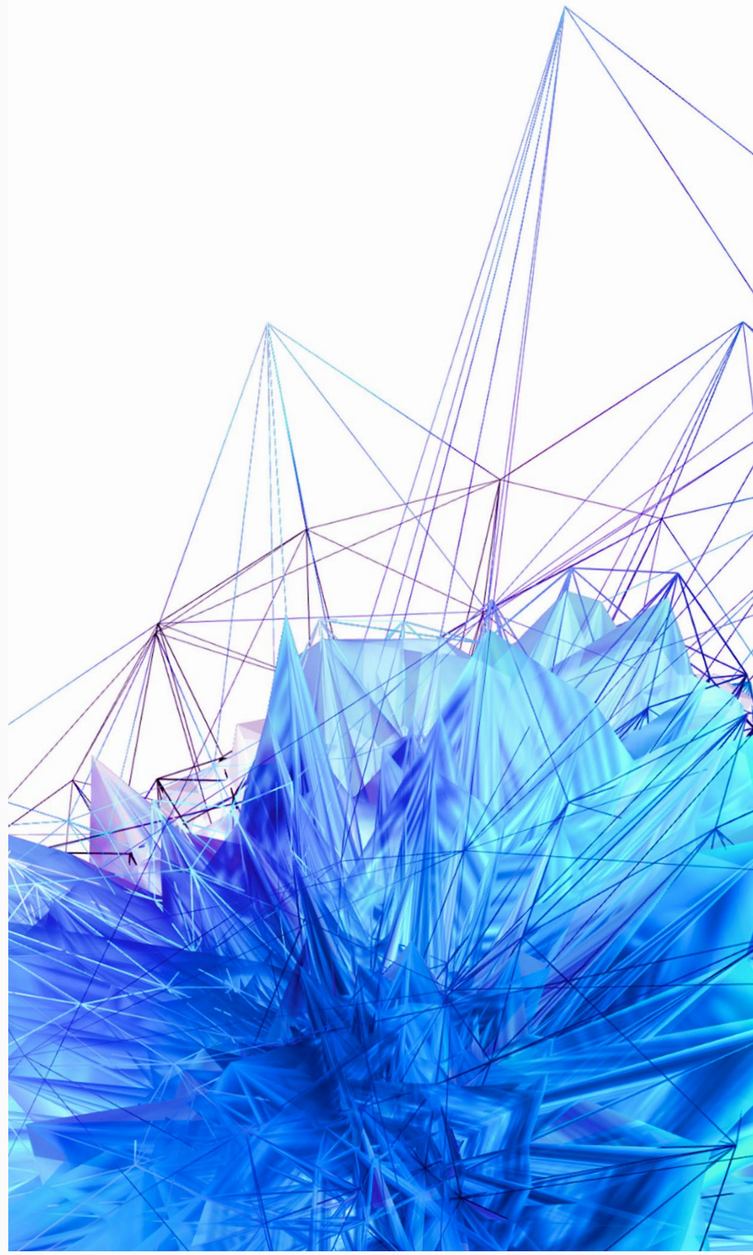
ANALYZE

End-to-end analytics platforms that cover all four stages of the data pipeline, like SAP*, Oracle or Microsoft* generally enable simple integration of business data. Intel works closely with these and other specialist solution providers and key ecosystem players to optimize our customers' analytics and AI environments. For example, the new 2nd generation Intel Xeon Scalable processor and Intel Optane DC persistent memory combine to deliver accelerated insights on in-memory and other analytics workloads, with up to 8x performance improvement on queries[3]. Meanwhile, 25Gb Intel® Ethernet 700 Series connectivity, combined with the latest generation processor technology delivers business insights up to 2.5x faster compared to 1Gb Ethernet products[4].

Further accelerations can be achieved with additional technologies such as Intel® FPGAs, which enable customized acceleration for a range of compute-intensive AI applications; and Intel® Quick Assist Technology (Intel® QAT), which provides a software-enabled foundation for security, authentication, and compression, which significantly enhance the performance and efficiency of standard platform solutions in a software-defined infrastructure.

Software optimizations and frameworks can also help get AI workloads up and running quickly. Common frameworks like TensorFlow*, Caffe*, PaddlePaddle*, and MXNet*, and platforms like Analytics Zoo*, are designed to help you deploy machine or deep learning at speed, so data science and development teams can explore which ones are most suited to their needs. Intel has worked closely with the software ecosystem to develop libraries, toolkits and software optimizations that help your developers boost app performance and cut development time, for little or no cost. For example:

- The Intel® Data Plane Development Kit (Intel® DPDK) is a set of libraries and drivers that accelerate packet processing, available for free, which can be used with any x86 platform.
- Intel® Distribution of OpenVINO™ toolkit enables developers to quickly build applications and solutions that emulate human vision, with a 'write once, deploy anywhere' approach.
- The Intel® Math Kernal Library (Intel® MKL) optimizes code with minimal effort for future generations of Intel® processors.
- The Intel® Data Analytics Acceleration Library (Intel® DAAL) helps software developers reduce the time it takes to develop high-performance applications.



ACT

The final stage of the data pipeline is delivering insights. Data scientists and business analysts will play a key role here, in helping turn these insights into business value. The techniques they use will vary depending on the project. They will perform a range of aggregation and analysis tasks across multiple datasets, which may vary widely in type and structure. Possible techniques include statistical and machine learning methods, like linear/non-linear modeling, clustering techniques, dimensionality-reduction techniques, graph-based methods, and neural networks.

During the insights phase, data visualizations, dashboards and reports can help business users understand the results. The software you use for this should be flexible enough to evolve and develop over time in line with changing business needs and new analytics and AI use cases. To make this easier, popular cloud service providers integrate with a number of reporting and dashboarding tools.

Begin with your existing infrastructure and business intelligence (BI) tools, then evolve as your capabilities and resources expand. For example, an online retailer may begin by using structured data in a relational database to track customer transactions. The next step may be to complement customer information with unstructured data which would enable the organization to build a fuller picture of each customer's experience. For instance a customer looking to place an order online who gets an error message, may then phone customer support and wait 15 minutes for an answer before giving up. They may then post a comment on Twitter. In this instance, the initial issue online is caused by a web server error, while the support call goes into a relational customer relationship management (CRM) system. By bringing these together with the social media post, the organization can gain a more complete, nuanced understanding of the customer and their pain points. This type of user behavior analytics is a great way of extracting more insight from your data over time, and laying the foundation for evolving smoothly towards more complex AI techniques like machine or deep learning.

THE LONGER READ: PREPARE YOUR DATA PIPELINE:

- For more on the four stages of the data lifecycle: From data to insights
- To explore what it means to be data ready: Meet the real AI need

CUSTOMER SUCCESS: ZTO EXPRESS

ZTO Express* is the largest express delivery service in China and one of the largest express delivery companies globally. ZTO provides express delivery service as well as other value-added logistics services.

ZTO packages travel from one of 22,000 service outlets to one of 68 sorting hubs and then on to a destination sorting hub before arriving at its final destination. This entire route is mapped out in just 15 seconds and is measured with real-time analytics. This generates massive amounts of data that must be processed and stored. All that data requires an incredibly agile and responsive technology infrastructure.

ZTO is using 2nd generation Intel® Xeon® Scalable processors and Intel® Optane™ DC persistent memory to increase performance, reduce bottlenecks and keep costs in check. It means fast route processing time, more accuracy and streamlined delivery.

CHAPTER THREE: DEVELOP

SET UP

With a data pipeline in place, you can develop your AI algorithms. Start by setting up a compute environment to run your model training. This can be done using your existing CPU resources, and adding extra performance using FPGAs and/or accelerators if time is constrained.

A range of tools, best practices and use cases are available through the Intel® AI Builders program. This ecosystem, made up of end users, independent software vendors (ISVs), original equipment manufacturers (OEMs) and systems integrators (SIs), can help you accelerate your AI development and deployment.

> Develop and train your AI algorithm on your existing Intel® architecture. Explore the Open Source tools and optimizations available that can help accelerate and streamline the development process.

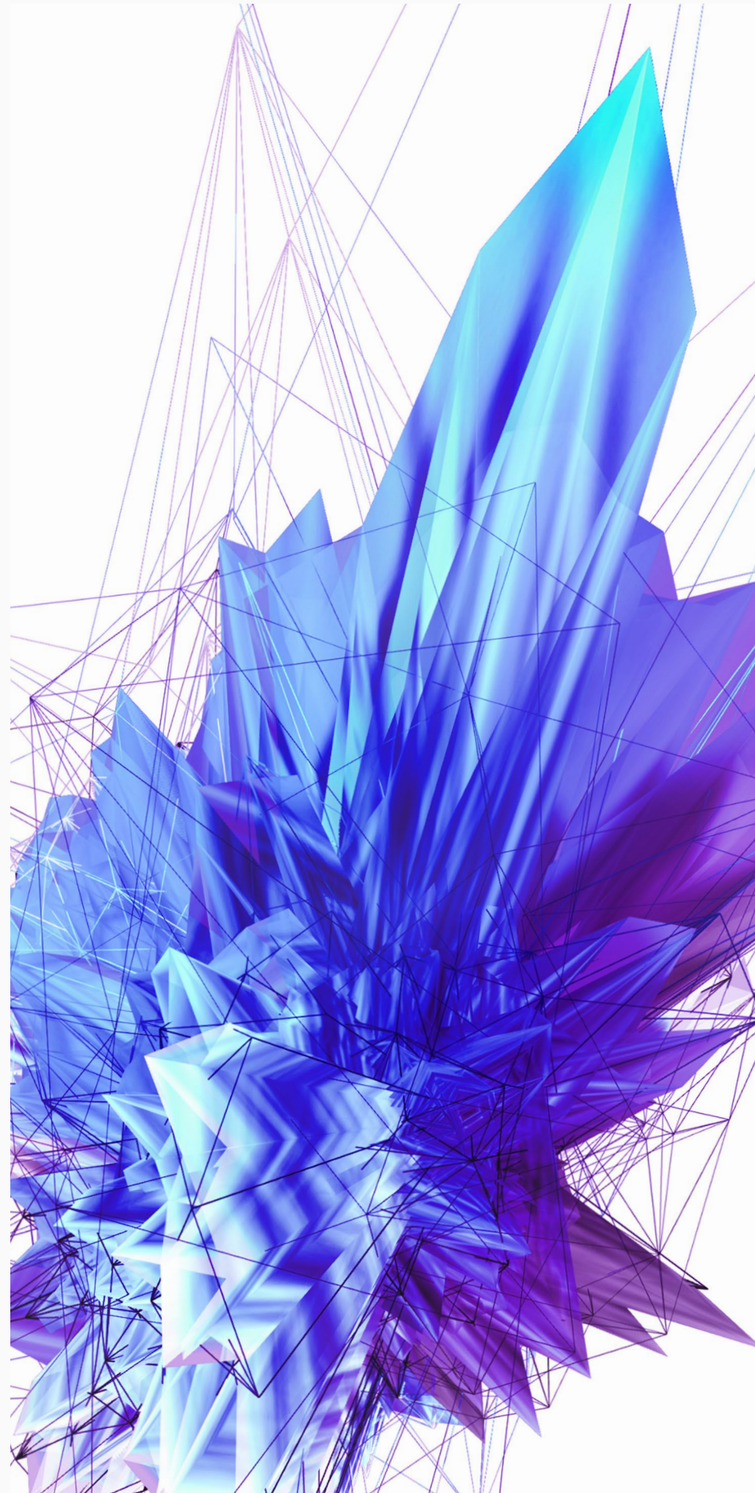
MODEL

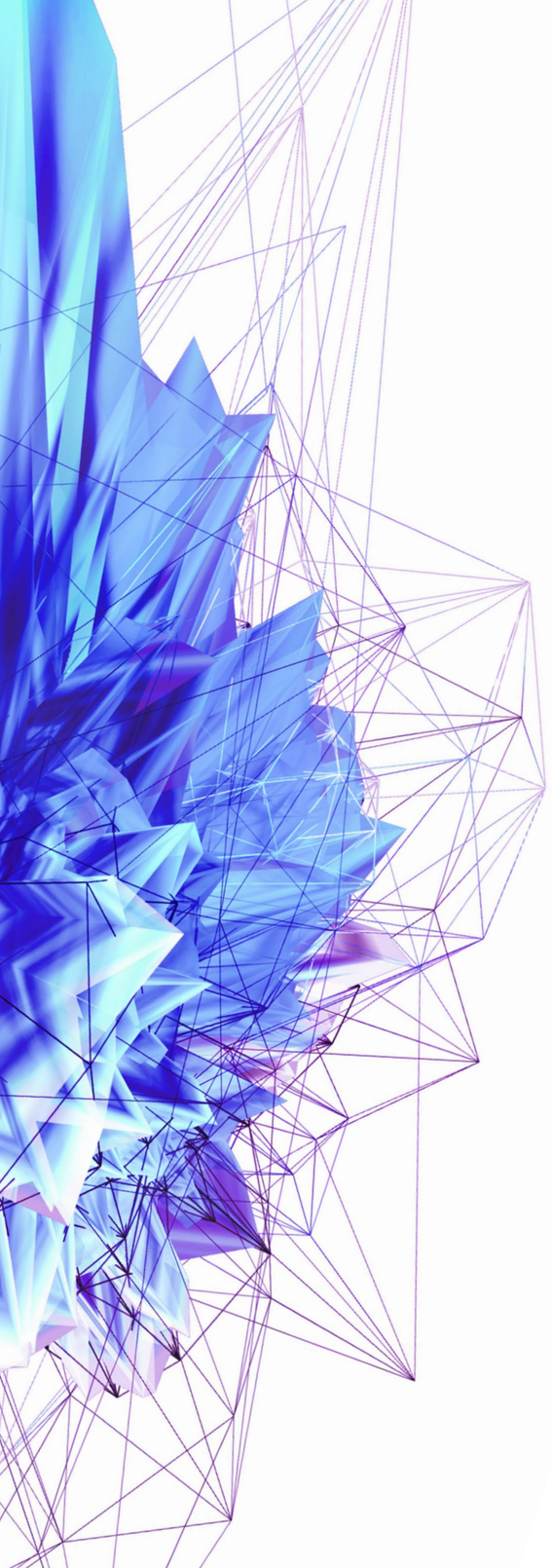
For the vast majority of data scientists seeking to develop a full-stack analytics or deep learning solution on your existing infrastructure, Intel provides a number of optimized frameworks, tools, and services to streamline the development process and meet their performance and accuracy requirements more quickly. Intel's solutions involve layered optimizations that allow you to run open source software – such as TensorFlow* – on an existing Intel® Xeon® processor-based platform in the most performant way possible. An additional strength of such bottom-up optimized solutions is the genuine scalability of the system. The following is a summary of the full-stack optimizations that Intel provides for developing your deep learning model on Intel® architecture:

The **Intel® Distribution of OpenVINO™ Toolkit** enables performance for a variety of different deep learning frameworks and applications running on Intel architecture, such as Intel Xeon processors (CPU), integrated Intel Iris Graphics (iGPU), Intel® Movidius™ technology (VPUs at the edge), and Intel® Stratix® FPGAs. The purpose of this particular distribution is to enable AI application developers to deploy and scale their application diversely without the need for additional framework- and/or architecture-specific changes to the application. Documentation and guides on the Intel Distribution of OpenVINO toolkit can be found here.

TensorFlow* is Google's most popular open source deep learning framework. With its importance in mind, Intel has made a number of optimizations under the hood to ensure training and inference on TensorFlow is as performant as possible on Intel Xeon processors. If you are running TensorFlow on your existing system, be sure to install the pre-built, Intel-optimized wheel. The details on how to do this are included in the white paper in 'the longer read' at the end of this section.

The **Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN)** is an open source performance library that enables compute improvement of deep learning applications and frameworks on Intel® architecture. Intel MKL-DNN has been enabled for a number of applications, including but not limited to Intel® Optimization for Caffe*, Intel® Optimization for TensorFlow*, Microsoft CNTK* and the Intel Distribution of





OpenVINO toolkit. Developer documentation for MKL-DNN can be found here.

Intel® Deep Learning Boost (Intel® DL Boost)'s primary benefit is Vector Neural Network Instruction (VNNI) – an instruction set architecture designed to accelerate inference performance on convolutional neural networks and other artificial intelligence models. Intel DL Boost can optimize inference workloads such as image classification, speech recognition, language translation, and object detection running 2nd generation Intel Xeon Scalable platforms. Explore how Intel DL Boost can enable your AI solution here.

Intel will work closely with your team to help you evaluate your needs and subsequently develop an optimized analytics and AI solution stack on your existing Intel® hardware.

TEST AND DOCUMENT

When training your AI model, you may choose to use simulated data, for example through a generative adversarial network (GAN). Be sure to hold some data back so that you have fresh data with which to test the model after you have trained it. A common approach when training models is to use a hybrid cloud strategy that allows the user to develop their model in the cloud then migrate it to an on-premise solution or to develop it on-premise, using burst capacity in the cloud through an HPC-as-a-Serice provider.

After developing the model, test it using a control data set that the model has not seen before, to determine if its accuracy meets requirements. When you are satisfied with the results, document the code, process and key learnings for future reference.

THE LONGER READ: DEVELOP AND TRAIN YOUR AI ALGORITHMS:

- For in-depth guidance on developing deep learning models: Best practices for scaling deep learning training and inference with TensorFlow* on Intel® Xeon® processor-based HPC infrastructures

CHAPTER FOUR: DEPLOY

Whichever approach you choose, it's important to run a proof of concept (PoC) before full deployment in order to demonstrate deployment sustainability, business value, and secure buy-in from any skeptical stakeholders. You can run a PoC using existing resources and a sample data set. This will make it much easier to operationalize your AI use case in a production environment later. It will also help you to determine whether your data strategy and IT infrastructure are sound, and identify priority areas for investment to build your capabilities.

Work with your solution architects to map network, hardware and software capabilities to your data pipeline and AI requirements.

Develop a strategy to optimize AI workload placement across on-premise, private and public cloud environments, using VMs and containers to help maximize utilization of existing resources.

ARCHITECT AND IMPLEMENT

Each AI use case brings its own infrastructure demands. For example, if you're planning to run AI at the edge to analyze video footage, priorities will be low-latency, high-performance compute and high persistent memory capacity at the edge to support it. If you plan to train algorithms in the cloud and then run iteration on-premise, you'll need compatible environments to enable the necessary programs to transition seamlessly between platforms. Consider each area of your infrastructure to determine how it can be optimized for your AI needs now and in the future. Remember that while processing power is important, it is only part of the picture. For example, bottlenecks in memory or network bandwidth can prevent your compute resources and the software they support from reaching their full potential.

NETWORK

Data rarely resides in one location. With disparate databases and applications across on-premise, private and public cloud, it's important that your network is able to support the secure and seamless transfer of large amounts of data very quickly. Without strong connectivity, your AI applications won't be able to reach their potential.

Consider your connectivity needs in terms of moving data north-south (i.e. into and out of the data center) as well as east-west (within the data center). Harmonizing processor performance and memory speed often means the bottleneck moves to the network, so performance here needs to keep up with the rest of the infrastructure. When connecting the data center to edge devices, plan to explore 5G as it becomes an option.

Built to work better together, the 25Gb Intel Ethernet 700 Series products and 2nd generation Intel® Xeon® Platinum processors accelerate the performance of the data center to deliver business insights up to 2.5x faster compared to 1Gb Ethernet products[5].

CUSTOMER SUCCESS: IFLYTEK CO., LTD.

As a Chinese and global front-runner in the intelligent speech and artificial intelligence (AI) industry, iFLYTEK Co., Ltd., which was founded in 1999, has established a leading international position in such technologies as speech recognition and natural language processing, and has captured a large share of the Chinese speech technology market. It is currently planning to explore new AI applications in cognitive fields on the basis of its iFLYTEK Cloud.

iFLYTEK has already successfully run its AI Cloud on Intel® Xeon® Scalable processors. In addition, it has migrated large numbers of existing GPU-based AI applications to platforms based on Intel Xeon Scalable processors and has witnessed excellent performance optimization results.

"Intel is not only an outstanding hardware product provider, but also a leading full stack solution provider in the field of artificial intelligence. They are a trustworthy, dependable, all-round partner for our artificial intelligence strategy, helping us effectively explore paths to future innovation in artificial intelligence," said Mr. Zhijiang Zhang, Vice President of the iFLYTEK's Cloud Computing Research Institute.

He continued: "The 2nd Generation Intel® Xeon® Scalable processor with integrated Intel® Deep Learning Boost (VNNI) technology, will greatly boost our AI computing. It will also improve the TCO of our AI Cloud by allowing us to remove the previous GPU card. In real workload testing in AI Cloud with 2nd gen Intel Xeon Scalable processors, we can get similar or better performance with VNNI enabled compared to the GPU solution. We built a hot data cache with Intel® Optane® SSDs in the AI Cloud to provide fast access for the models during computing, which improve the average response time of AI Cloud significantly.

SOFTWARE

The applications and tools that your data scientists and business users need will vary depending on the use case you choose. Ensure the suite of apps that you need will work well together and that the key stakeholders understand how to use them.

Most leading solution providers, including SAS*, SAP, Microsoft and Cloudera* offer cross-environment compatibility but it is worth confirming up-front that the independent software vendors (ISVs) you choose to work with offer seamless integration from end to end, across all your on-premise and cloud applications and databases. Lack of integration here can create issues if you wish to move a particular AI workload from the public cloud to your on-premise database (or vice versa), and the software you wish to use is not optimized for one of those environments. Likewise, if the hardware components you wish to use are not certified for particular applications, you may not be able to use them. Using an HPC-as-a-Service provider for a hybrid or multi-cloud solution can be an advantage for seamless migration to and from the cloud. It also allows any independent software vendors (ISVs) you use to reduce the amount of testing they need to do to support it.

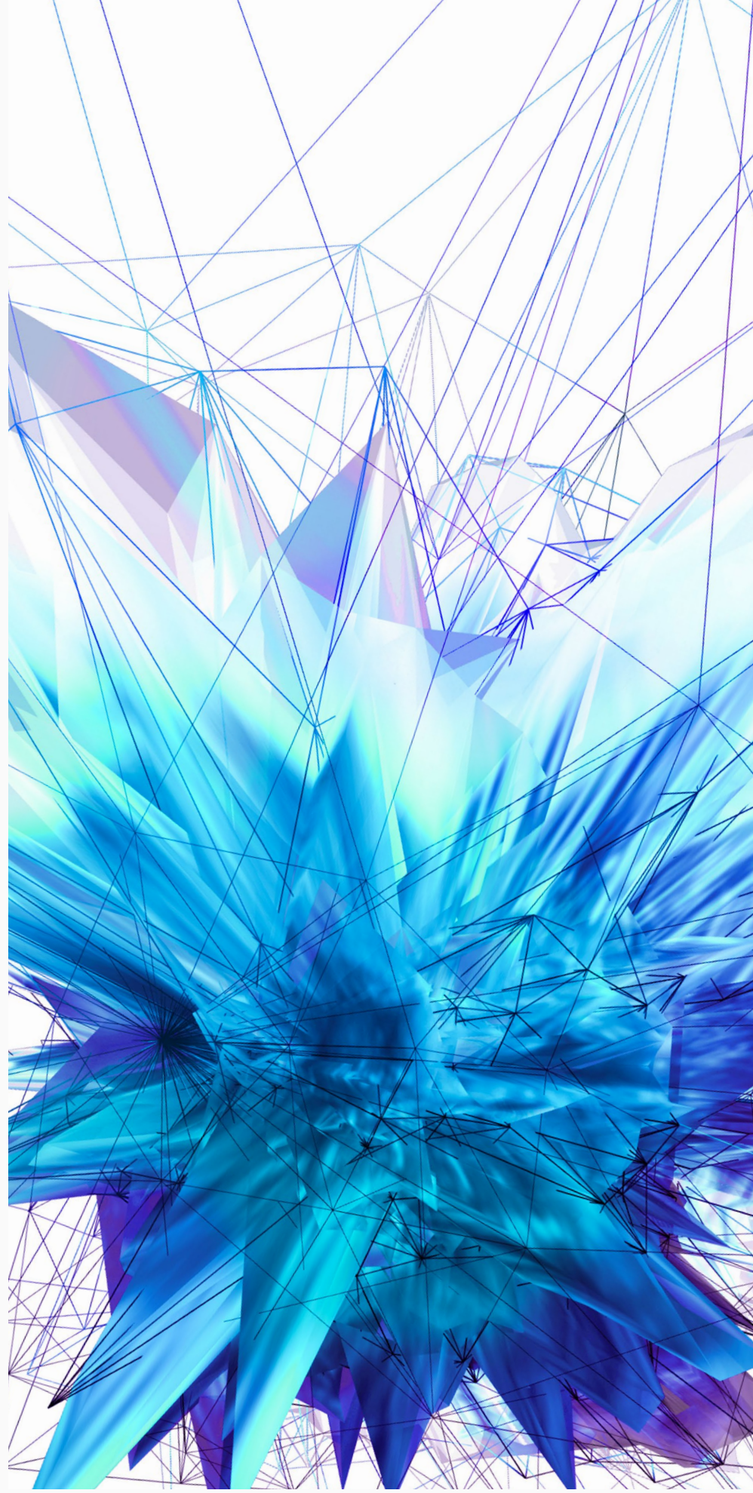
Intel is a longtime believer in democratizing technology and making it accessible to every type of developer so they can work across multiple types of hardware in increasingly complex AI environments. As such, we are committed to open source, and have worked with the software ecosystem and open source community for years to ensure popular libraries and frameworks are optimized to facilitate speedy adoption.

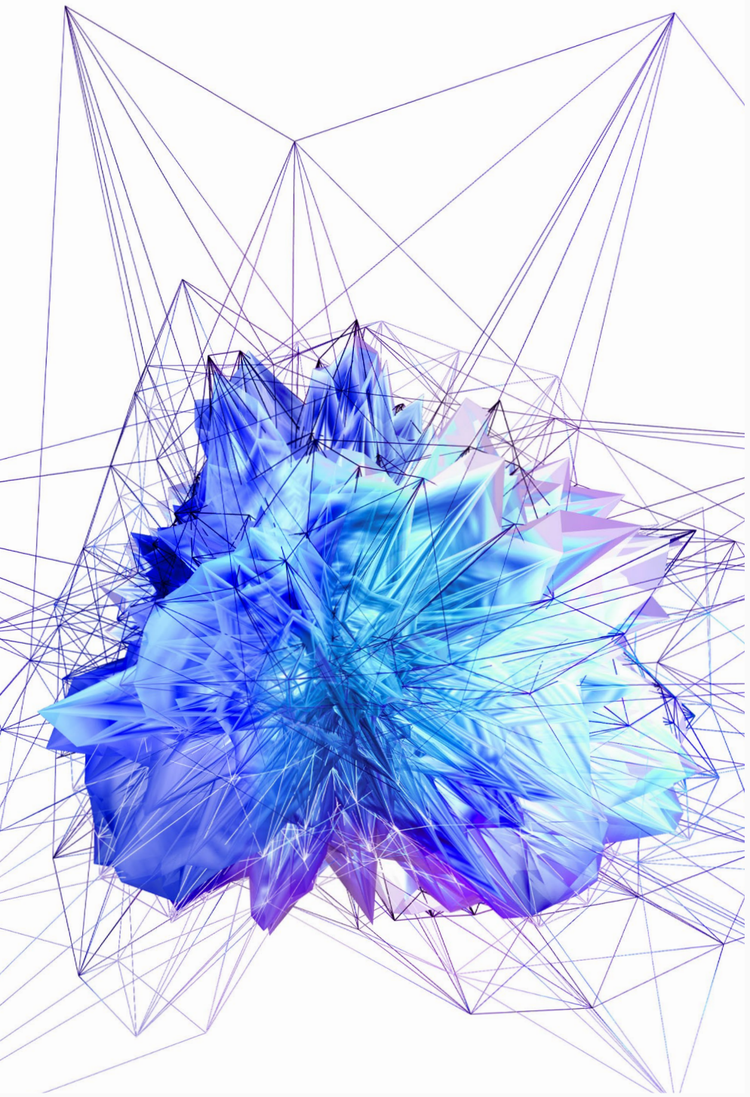
We work with leading framework developers to enhance deep learning performance using software optimizations for Intel Xeon processors in the data center. Intel optimizations available today include TensorFlow, PaddlePaddle, Caffe and MXNet. Also, Intel has worked with Apache Spark*, BigDL*, TensorFlow and Keras* to develop Analytics Zoo, a unified AI and analytics platform. Using these optimized versions allow you to follow a CPU training and inference strategy which has TCO and other benefits.

HARDWARE

Powerful compute performance is of course critical to support complex AI workloads and applications. AI hardware should be matched to the environment, meaning high-throughput, low-latency inference on devices at the edge, or powerful, memory-rich training and inference in the data center. Ensure this performance can be maintained at scale – for example by making sure sufficient data can be held in/near the processor. In many cases, the large datasets needed cannot be contained on an accelerator, meaning time is lots in transferring data on and off the processor. Using a CPU-based environment helps avoid this issue.

Fortunately you can start with the Intel Xeon processor-based infrastructure you know, and Intel is committed to helping its customers build an IT environment for the data era and optimized for AI that starts with their existing resources. Intel CPUs are more performant today than ever, and the latest 2nd generation Intel Xeon Scalable processors are the only processors with AI acceleration built in. With Intel Deep Learning Boost they accelerate AI inference up to 30x[6].





Compute is of course only part of the story. Data-intensive workloads also demand fast, high-capacity storage and memory, although requirements will differ for training and inference applications. For example, training tends to use as much and as varied data as possible. This requires high-capacity persistent memory, which can be hard to achieve with costly DRAM-based systems, or with capacity-constrained GPUs. Make sure you're providing the resources needed here – if not, you risk creating bottlenecks in feeding data to the processor, effectively starving it, and preventing your AI applications from running optimally.

Delivered with 2nd generation Intel Xeon Scalable processors, Intel Optane DC persistent memory delivers the high-capacity memory that AI workloads need at a lower cost than DRAM. This gives you increased VM, container, and application density for improved operational efficiency and compelling TCO. Persistence delivers faster in-memory database restart times, increasing service reliability, decreasing costly downtime, and offering more data security by providing added opportunities for system patching/administration. In addition, Intel Optane DC SSDs enable you to break through NAND SSD bottlenecks to increase system density and accelerate applications. Intel Optane DC SSDs (with Intel Memory Drive Technology) work seamlessly as system memory to expand training set size beyond DRAM-only capacity. The ability to scale to larger and larger data sets helps to deliver more diverse AI training models, boosting accuracy.

Where additional performance is required, you may also be able to deploy devices like FPGAs to accelerate your AI applications and enable you to get started with your existing CPU resources. Using FPGAs also offers the flexibility to adjust your use cases and applications over time. Intel FPGAs are easily modifiable for a wide range of data types and applications, delivering particular value for high-throughput, low-latency inference applications.

As we mentioned above, it is increasingly common to see compute capabilities being disaggregated from the data center and brought to where the data resides, often at the edge. This enables analysis at the point of capture to either automate immediate adjustments (such as monitoring for safety in manufacturing) or to ensure that the network is not being overloaded by sending irrelevant or low-priority data to the cloud or data center (for example, in surveillance, only alerting a human reviewer to footage of suspicious activity rather than requiring them to watch everything all the time). Ensure that if your strategy involves AI at the edge, you choose sensors and devices that have the compute performance, memory capacity and network bandwidth to deliver the results you need.

ACCELERATE AND SCALE IMPLEMENTATION

For organizations wishing to implement AI in their production environments at speed, Intel has developed a range of Intel® Select Solutions, which combine elements of hardware, software, and connectivity that are optimized for particular AI use cases to help you get up and running quickly.

- Intel® Select Solution for BigDL for Apache Spark*: accelerates and simplifies deep learning development and deployment on an optimized, verified infrastructure based on Apache Spark, using Intel® Xeon® Gold processors, Intel® SSDs, and Intel® Ethernet Network Adapters.
- Intel® Select Solution for AI Inferencing: combines the Intel Xeon Scalable processor, Intel SSDs and Intel Ethernet Network Adapters to help optimize price and performance while significantly reducing infrastructure evaluation time.

As you implement AI in your production environment, keep in mind how you will scale to more sites and users as demand grows, and how you will iterate on your models if and when you have new data to draw from in the future. This is where the cloud becomes particularly important. IT leaders today are under increasing pressure to think like cloud service providers, and build a unified, 'one cloud' environment that enables business users to access the data, applications and resources they need on demand, and in a flexible way. This means it is becoming critical to be able to shift your workloads around between on-premise, private and public cloud as if they were all one environment. For example, you may choose to use the public cloud to train a new AI algorithm, but then move back to on-premise or private cloud to run iterations, depending on the location of your data and cost, security or regulatory restrictions around where you can place it.

You should also determine your optimal workload placement strategy for your AI applications, setting decision criteria or guidelines around how and where they should be run, and under which circumstances it makes sense to move them. Criteria may include data gravity and affinity, cost, security, storage or memory capacity, latency tolerance and so on. You should consider how and where VMs and containers can be used to optimize utilization of your existing resources as much as possible.

Intel has worked closely with leading cloud service providers (CSPs) and other ecosystem players to ensure its customers can build seamless multi-cloud environments, running the most compute-intensive workloads, like AI, fast and more cost effectively on Intel technology-based cloud instances.

With your AI model up and running, you move on to a continuous round of iterations. Monitor and measure algorithm performance, efficiency etc. against the goals and key performance indicators (KPIs) that you determined with your key stakeholders at the beginning of the journey.

Ensure you have regular methodologies for adapting and evolving your algorithm as needed to enhance results and/or build in new data sources or address more nuanced business challenges.

THE LONGER READ: DEPLOY AI FOR LONG-TERM SUCCESS

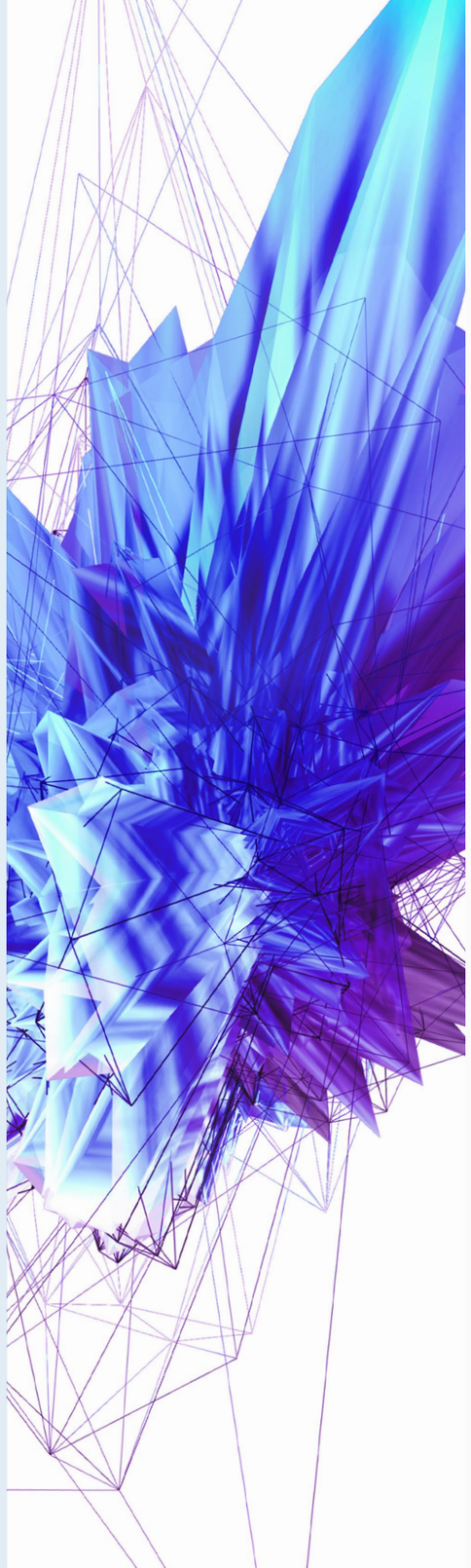
- For more tips on architecting for AI: Intel AI step-by-step guide
- For more on defining a workload placement strategy: Optimal workload placement for public, hybrid, and private clouds
- To explore VM and container migration options: Some like it hot - VM and container migration in hybrid cloud environments
- To explore the transition from PoC to full deployment: From proof of concept to production
- For guidance on collating data center resources to support AI: Select the best infrastructure strategy to support your AI solution
- To discover more about a real-life AI project: Seeking smart cores for AI
- For details on building a multi-cloud strategy: Making multi-cloud work

CONCLUSION

Preparing for and launching AI into your organization is an ongoing conversation between you as IT leadership, your colleagues across the business, and senior management. By integrating business, data and technological considerations from the outset you can make your AI journey a smooth, positive and rewarding one.

LEARN MORE BY READING:

- Solution Brief: Intel Select Solution for AI Inferencing
- White Paper: Five Steps to an AI Proof of Concept



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 http://www.intel.com/performance

[1] https://cisr.mit.edu/reports/create-a-data-strategy/intro.php

[2] Intel, System Configuration for Management Node: S2600WFT Intel white box, 2 sockets, Intel® Xeon® Gold 6140 CPU @ 2.30GHz, 18 cores per socket / 2 threads per core (total 72 vcores), 192GB DDR4, CentOS 7.4* distribution with 4.15.12 kernel, BIOS v13, Horton-Works* Data Platform 2.6.4, Spark 2.2.0*. Configuration for Data Node(s): Same as above plus 2x NVMe* PCIe* Intel® Optane™ SSD DC P4800X 375GB per system, 2x NVMe* PCIe* Intel® SSD DC P4500 3.7 TB per system. Performance results are based on testing as of July 30, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure

[3] 8x improvement in queries result based on testing by Intel on 1 November 2018. Baseline configuration: Platform: S2600WF (Wolf Pass); number of nodes: 1; number of sockets: 2; CPU: Intel® Xeon® Platinum 8280L CPU @ 2.70 GHz; Cores/socket, threads/socket: 28 cores/socket, 2 threads/socket; ucode: (microcode: 0x400000a); HT: Enabled; Turbo: Off; BIOS version: SE5C620.86B.0D.01.0134.100420181737; BKC version: WW06'19; FW version: N/A; System DDR Mem Config slots/cap/run-speed: DDR Mem: 24 / 32GB / 2666 MT/s; System Intel Optane DC persistent memory Config: slots / cap / run-speed: N/A; Total Memory/Node (DDR, Intel DC Optane perst. mem.): 768GB DDR; Storage – boot: SATA SSD 500GB; Storage - application drives: HDD (ST1000NX0313) * 8; NIC: 10-Gigabit SFI/SFP+ Network Connection; Software: Spark Hadoop; OS: Fedora release 29 (Twenty Nine); Kernal: Linux-4.18.8-100.fc27.x86_64-x86_64-with-fedora-27-Twenty_Seven BIOS: SE5C620.86B.0D.01.0299.122420180146; Mitigation log attached: 1,2,3,3a,4, L1TF; Intel Optane DC persistent memory mode: N/A; Run Method: Run 9 I/O intensive queries together in a few iterations; Iterations and result choice: 3 iterations and choose the execution time for second or third iteration; Dataset size: 2TB/3TB; Workload & version: Decision Support I/O intensive queries; Compiler: gcc (GCC) 8.3.1 20190223 (Red Hat 8.3.1-2), JDK 1.8.0_201; Libraries: Memkind; Other software: Spark + Hadoop. New configuration: Platform: S2600WF (Wolf Pass); number of nodes: 1; number of sockets: 2; CPU: Intel® Xeon® Platinum 8280L CPU @ 2.70 GHz; Cores/socket, threads/socket: 28 cores/socket, 2 threads/socket; ucode: (microcode: 0x400000a); HT: Enabled; Turbo: Off; BIOS version: SE5C620.86B.0D.01.0134.100420181737; BKC version: WW06'19; FW version: N/A; System DDR Mem Config slots/cap/run-speed: DDR Mem: 12 / 16GB / 2666 MT/s; System Intel Optane DC persistent memory Config: slots / cap / run-speed: 8 / 128GB / 2666 MT/s; Total Memory/Node (DDR, Intel DC Optane perst. mem.): 192GB DDR + 1TB DCPMM; Storage – boot: SATA SSD 500GB; Storage - application drives: HDD (ST1000NX0313) * 8; NIC: 10-Gigabit SFI/SFP+ Network Connection; Software: Spark Hadoop; OS: Fedora release 29 (Twenty Nine); Kernal: Linux-4.18.8-100.fc27.x86_64-x86_64-with-fedora-27-Twenty_Seven BIOS: SE5C620.86B.0D.01.0299.122420180146; Mitigation log attached: 1,2,3,3a,4, L1TF; Intel Optane DC persistent memory mode: App Direct; Run Method: Run 9 I/O intensive queries together in a few iterations; Iterations and result choice: 3 iterations and choose the execution time for second or third iteration; Dataset size: 2TB/3TB; Workload & version: Decision Support I/O intensive queries; Compiler: gcc (GCC) 8.3.1 20190223 (Red Hat 8.3.1-2), JDK 1.8.0_201; Libraries: Memkind; Other software: Spark + Hadoop.

[4] Testing based on the 2nd Generation Intel® Xeon® Platinum 8260 processor and upgrading from a 1GbE to a 25Gb Intel® Ethernet Network Adapter XXV710 and from Serial ATA (SATA) drives to the NVM Express* (NVMe*)-based PCIe* Intel® SSD DC P4600. Performance results by HeadGear Strategic Communications are based on testing as of February 12, 2019. The comparative analysis in this document was done by HeadGear Strategic Communications and commissioned by Intel. Detailed configuration details: VM Host Server: Intel® Xeon® Platinum 8160 processor, Intel Xeon Platinum 8160F processor (CPUID 50654, microcode revision 0x200004D), and Intel Xeon Platinum 8260 processor (CPUID 50656, microcode revision 04000014); Intel® Server Board S2600WFT (board model number H48104-850, BIOS ID SE5C620.86B.0D.01.0299.122420180146, baseboard management controller [BMC] version 1.88.7a4eac9e; Intel® Management Engine [Intel® ME] version 04.01.03.239; SDR package revision 1.88); 576 GB DDR4 2,133 MHz registered memory, 1 x Intel® Ethernet Network Adapter XXV710-DA2, 1 x Intel® Ethernet Converged Network Adapter X710-DA2; operating system drive configuration: 2 x Intel® SSD DC S3500 in Intel® Rapid Storage Technology enterprise [Intel® RSTe] RAID1 configuration. Windows Server 2016* Datacenter edition 10.0.14393 build 14393, Hyper-V* version 10.0.14393.0, Hyper-V scheduler type 0x3, installed updates KB4457131, KB4091664, KB1322316, KB3211320, and KB3192137. E-mail Virtual-Machine Configuration: Windows Server 2012 Datacenter edition 6.2.9200 build 9200; 4 x vCPU; 12 GB system memory, BIOS version/date: Hyper-V release v1.0, 2012, 11/26), SMBIOS version 2.4; Microsoft Exchange Server 2013*, workload generation via VM clients running Microsoft Exchange Load Generator 2013*, application version 15.00.0805.000). Database Virtual-Machine Configuration: Windows Server 2016

Datacenter edition 10.0.14393 build 14393, 2 x vCPU 7.5 GB system memory; BIOS version/date: Hyper-V release v1.0, 2012, 11/26), SMBIOS version 2.4, Microsoft SQL Server 2016* workload generation DVD Store application* (dell.com/downloads/global/power/ps3q05-20050217-Jaffe-OE.pdf). Storage Server: Intel® Server System R2224WFTZS; Intel Server Board S2600WFT (board model number H48104-850, BIOS ID SE5C620.86B.00.01.0014.070920180847, BMC version 1.60.56383bef; Intel ME version 04.00.04.340; SDR package revision 1.60); 96 GB DDR4 2,666 MHz registered memory, 1 x Intel Ethernet Network Adapter XXV710-DA2, 1 x Intel Ethernet Converged Network Adapter X710-DA2; operating system drive configuration: 2 x Intel SSD DC S3500 in Intel RSTe RAID1 configuration. Storage Configuration: 8 x Intel SSD DC P4600 (2.0 TB) configured as RAID 5 volume using Intel® Virtual RAID on CPU (Intel® VROC), 8 x Intel SSD DC S4500 (480 GB) in RAID5 configuration using Intel® RAID Module RMSP3AD160F, 8 x Intel SSD DC P4510 in RAID 5 configuration using Intel VROC for VM operating system store, Windows Server 2016 Datacenter edition 10.0.14393 build 14393, Hyper-V version 10.0.14393.0, Hyper-V scheduler type 0x3, installed updates KB4457131, KB4091664, KB1322316, KB3211320, and KB3192137. Windows Server 2016 Datacenter and Windows Server 2012 Datacenter Configured with Intel Xeon Platinum 8160 and Intel Xeon Platinum 8160F Processors: Speculation control settings for CVE-2017-5715 (branch target injection)—hardware support for branch target injection mitigation is present: true; Windows* operating system support for branch target injection mitigation is present: true; Windows operating system support for branch target injection mitigation is enabled: true; Windows operating system support for branch target injection mitigation is disabled by system policy: false; Windows operating system support for branch target injection mitigation is disabled by absence of hardware support: false. Speculation control settings for CVE-2017-5754 (rogue data cache load)—hardware requires kernel VA shadowing: true; Windows operating system support for kernel VA shadow is present: true; Windows operating system support for kernel VA shadow is enabled: true. Speculation control settings for CVE-2018-3639 (speculative store bypass)—hardware is vulnerable to speculative store bypass: true; hardware support for speculative store bypass disable is present: true; Windows operating system support for speculative store bypass disable is enabled system-wide: true. Speculation control settings for CVE-2018-3620 (L1 terminal fault)—hardware is vulnerable to L1 terminal fault: true; Windows operating system support for L1 terminal fault mitigation is present: true, Windows operating system support for L1 terminal fault mitigation is enabled: true. Windows Server 2016 Datacenter and Windows Server 2012 Datacenter Configured with Intel Xeon Platinum 8160 and Intel Xeon 8160F Processors: Speculation control settings for CVE-2017-5715 (branch target injection)—hardware support for branch target injection mitigation is present: true; Windows operating system support for branch target injection mitigation is present: true; Windows operating system support for branch target injection mitigation is enabled: true. Speculation control settings for CVE-2017-5754 (rogue data cache load)—hardware requires kernel VA shadowing: false. Speculation control settings for CVE-2018-3639 (speculative store bypass)—hardware is vulnerable to speculative store bypass: true; hardware support for speculative store bypass disable is present: true; Windows operating system support for speculative store bypass disable is present: true; Windows operating system support for speculative store bypass disable is enabled system-wide: true. Speculation control settings for CVE-2018-3620 (L1 terminal fault)—hardware is vulnerable to L1 terminal fault: false. Network Switches: 1/10GbE SuperMicro SSE-X3348S*, hardware version P4-01, firmware version 1.0.7.15; 10/25GbE Arista DCS-7160-48YC6*, EOS 4.18.2-REV2-FX.

[5] Testing based on the 2nd Generation Intel® Xeon® Platinum 8260 processor and upgrading from a 1 GbE to a 25 Gb Intel® Ethernet Network Adapter XXV710 and from Serial ATA (SATA*) drives to the NVM Express* (NVMe*)-based PCIe* Intel® SSD DC P4600.

[6] 30x inference throughput improvement on Intel® Xeon® Platinum 9282 processor with Intel® DL Boost: Tested by Intel as of 2/26/2019. Platform: Dragon rock 2 socket Intel® Xeon® Platinum 9282(56 cores per socket), HT ON, turbo ON, Total Memory 768 GB (24 slots/ 32 GB/ 2933 MHz), BIOS: SE5C620.86B.0D. 01.0241.112020180249, Centos 7 Kernel 3.10.0957.5.1.el7.x86_64, Deep Learning Framework: Intel® Optimization for Caffe version: https://githu b.com/intel/caffe d554cbf1, ICC 2019.2.187, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d9 4195140cf2d8790a75a), model: https://github.com /intel/caffe/blob/master /models /intel_optimized_models /int8 /resnet50_int8_full_conv. , BS=64, No prototxt datalayer syntheticData: 3x224x224, 56 instance/2 socket, Datatype: INT8 vs. Tested by Intel as of July 11 2017: 2S Intel® th Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7. x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb /s, 25nm, MLC). Performance measured with: Environment variables: KMP_AFFINITY='granulari ty=fine, compact', OMP_NUM_THREADS=5 6, CPU Freq set with cpupower frequency-set d 2.5G -u 3.8G -g performance. Caffe: (http:/), /github.com/intel/caffe/ revision f96b759f71b2281835f690 af267158b82b150b5c. Inference measured with "caffe time -forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, synthetic dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github. com/intel/caffe/tree/master /models /intel_optimized_models (ResNet-50). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

Performance results are based on testing as of the date set forth in the configurations and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

Intel does not control or audit third-party data. You should review this content, consult other sources, and confirm whether referenced data are accurate.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com

Intel, Xeon, Optane, Iris, and the Intel logo, are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others.

0719/KR/CAT/PDF

340905-001EN