WHITE PAPER

Data Center
Artificial Intelligence

The AI Readiness Model

Judging an Organization's Ability to Generate Business Value from Artificial Intelligence

Table of Contents

Understanding Where You Are on the AI Journey 1
Foundational Readiness 2
Operational Readiness. 3
Transformational Readiness 4
Conclusion 4

At Intel, we work with many organizations who are investigating artificial intelligence (AI) solutions. Image recognition, natural language processing (NLP) and predictive maintenance are emerging as particular hotspots. Some businesses are exploring these AI use cases for the first time; others are examining how to advance from a successful starting point.

To aid organizations wherever they are on their AI journeys, Intel has created a Readiness Model to help decision makers understand where to prioritize efforts. We have developed this based on our experience working with customers across a range of scenarios and industry verticals. Examples include manufacturing companies wanting to improve quality control, and financial services organizations looking to use AI in algorithmic trading. This paper provides guidance on how to judge an organization's ability and readiness to use AI to generate business value, and includes a list of questions which you can use to guide your own self-assessment activities.

To start with, we can group organizations into three categories, according to where they are in their AI journey – whether they are using AI for the first time, scaling up, or broadening out use of AI.

Understanding Where You Are on the AI Journey

Some organizations are new to AI

Many organizations are still unfamiliar with deploying AI applications, such as image recognition, natural language processing (NLP), and predictive maintenance. We see three common scenarios:

- **An organization with existing pools of data can benefit from the use of AI.** For example, a solar energy farm that Intel worked with regularly captured images of their hardware in an effort to spot damage. In this scenario, the organization wanted to understand whether image recognition algorithms could be used to identify flaws automatically. Similar situations can be found in healthcare, assessing radiology charts or other medical data.
- **An organization is running a workload in a traditional environment, and wants to apply AI or machine learning to explore opportunities for optimization.** For example, these might be enterprises already familiar with analytics deployments, or research institutions running physics algorithms. Similarly, a manufacturer using fixed patterns for fault detection in machinery, materials, goods or processes may recognize the benefit of machine learning.
- **Some organizations have been researching the potential of AI – for example, the use of NLP in customer service scenarios, or retailers using image recognition to optimize analytics of in-store behaviors.** In some cases, the problems that AI can solve the best may not be obvious – it can, for example, involve a behind-the-scenes use-case, such as predictive inventory management.

For these organizations, it can be a challenge to map out the benefits of AI in advance and data is not always available or presented in such a way that is suitable for AI.

Other organizations are ready to scale up AI

These organizations have researched the possibilities for AI, have an idea of where they want to use it and have successfully implemented test models. They may fall into a number of scenarios:

- **An organization may have developed a proof of concept AI solution running on a workstation or a single device.** The challenge is to migrate the solution to a data center environment so that it can be moved to production. One example would be enabling operations to work with predictive maintenance insights.
- **An organization has developed a 'home-grown' solution, and is now looking to use industry standard infrastructure and/or software.** Whether hardware, software or both, the challenge can come from migrating to an architecture which (in the shorter term) yields less optimal results.

While these organizations are more advanced in their AI journey, they can still lack the skills they need to 'scale up' their use of AI beyond a smaller number of engineers with a relatively simple hardware and software configuration. In addition, moving to a multi-node solution will need to address the fact that AI does not scale linearly. For example, whereas a single-node configuration can process hundreds of images per second, a move to 50 processors will not deliver 50 times the performance. Data sources may also bottleneck if required to be used 'at scale' versus one-off sampling.

Others are broadly implementing AI

A third category of organizations are using machine learning or AI to some extent, and are now looking to broaden their adoption across a wider range of use cases. We see the following scenarios:

- **An organization may be using AI successfully in a line of business, and is now looking to expand.** For example, a company might be using image recognition in manufacturing for quality control, and now wants to deploy NLP in call centers.
- **An organization is successfully using AI to learn from and interpret data, and now wants to extend into inference-based maintenance and updates to models.** They may also want to use outcomes of AI to drive automation, for example, using inventory data to drive spare-part inventory management, planning and acquisition. From an infrastructure point of view, the same organization might also be looking to improve power efficiency or performance and reduce total cost of ownership (TCO).

In these scenarios, challenges come from ensuring that the platform can fit more than one use case, and managing resources as requirements and resource demands fluctuate between each case.

The Three Types of AI Readiness

While organizations can be at various stages of their AI journeys, their progress to the next stage or to ongoing success is dependent on having the right elements in place across skills and resources, infrastructure and technology, processes and models.

We can also consider readiness in terms of the following areas:

- Foundational – a prerequisite for AI is appropriate infrastructure and interfaces
- Operational – suitable management and governance mechanisms are key to the sustainability of AI solutions
- Transformational – the ability of an organization to maximize the value it gets from AI

Foundational readiness is the first step, but the success of AI hinges on operational readiness, and then how receptive the business is to AI – transformational readiness. This feeds new and updated requirements, which further drive the foundation for AI deployments.

Foundational Readiness

A prerequisite for AI is appropriate infrastructure and interfaces. While this stands to reason, it is not always evident what the needs will be in advance of testing and evaluating potential scenarios. At the same time, skills and expertise may be in short supply. With these factors in mind, the following should be taken into account.

Infrastructure platform

Many organizations want to understand whether existing data center facilities will be suitable for AI workloads. While the answer may be yes for a proof of concept, some facilities may not suit the massively scalable processing required for machine learning and AI. Depending on the scenario, the data flow required by AI can severely tax your network bandwidth. Some AI solutions can be scheduled at off hours to maximize throughput. Meanwhile other, more time-critical scenarios may impose a greater burden on the network, for example if the data was linked into a predictive maintenance capability.

For further information about platform requirements for AI, across processing, storage and networking, see our paper on creating a proof of concept.

Cloud resources

Consider cloud-based services as a basis for AI, particularly in areas such as image and natural language processing. The cloud offers the advantage of low-entry-point, pay-per-use services, making it an excellent choice for training and testing. As organizations look to scale up use of AI, they will need to check if their cloud-based resources are still suitable. Cloud

service providers may have graded pricing and service level agreements, perhaps defined by the underlying compute infrastructure running the workload. Decision makers will need to strike the optimum balance between performance and cost – possibly on a workload by workload basis.

Data sources

Critical data sources need to be available and accessible. Decision makers and engineers will require an assurance that data is available in the right quantity and quality for deep learning algorithms, whether it is internally or externally sourced: in the latter case, this may need contractual arrangements to be agreed with third parties.

Software packages

A wide variety of inference, machine learning and AI software packages are available today, each with benefits and constraints. Open source packages such as, TensorFlow* and BigDL sit alongside commercial solutions such as the Intel® Saffron™ software and cloud-based services. Each of these needs to integrate with tools for data management, visualization and so on: each use case will require a software architecture that selects the best tools for the job in hand, potentially taking into account downstream systems, customization, optimization and other modifications.

Operational Readiness

Suitable management and governance mechanisms are key to the sustainability of AI solutions. IT decision makers can examine several areas to ensure their AI deployments will be operationally ready. We recommend a review of the following.

Agile delivery

Best practice models such as DevOps and other agile methodologies can significantly help organizations at any stage of their AI journey – in earlier stages, continuous development and delivery responds to quickly changing requirements, unclear outcomes and the need for repeated review, evaluation and testing. And for more advanced use, continuous improvement can drive changes in both data sources and models.

Operational management

The 'Ops' side of DevOps also needs to incorporate criteria around effective service delivery (across the platform selected as a foundation for AI), as well as effective management of all data sources, both internal and external. Additionally, criteria should exist around measuring the business effectiveness of AI – for example, are the generated insights or automation delivering the expected business value?

Skills and expertise

Lack of skills is a frequent challenge for organizations in the earlier stages of the AI journey. At the outset, you may need to bring in skills from outside, in particular solution architects who can tune proof-of-concept solutions; however, it is beneficial to build up in-house skills across both IT and lines of business. Proceeding beyond proof of concept without in-house skills adds unnecessary risk to an AI project, particularly in the review and evaluation stages.

Cybersecurity

Given how the relationship between AI and automation can minimize human intervention and therefore oversight, cybersecurity of data, infrastructure and algorithms should be seen as a high priority. Potential security risks come from the corruption of input data into the AI, tampering with the models, or unauthorized access to the resulting insights.

Governance, compliance and risk

The relationship between AI and governance is complex and many-faceted. In earlier stages of an organization's journey, governance questions will not be any different to other data-centric IT projects – can the project deliver, is customer privacy protected and so on. As its use is extended, AI brings additional ramifications: for example, in predictive planning and maintenance, how much human involvement is required in spares purchasing (if any)?

AI software optimizations for Intel® Xeon® processors

To allow data scientists and developers to work with their preferred framework, Intel has optimized a number of deep learning libraries for many of the most popular AI frameworks including, TensorFlow*, Theano*, and MXNet*.

Operating underneath these frameworks, the Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN) is a new accelerator with specific math for deep learning, optimized on top of x86 with Intel® Advanced Vector Extensions 2 (Intel® AVX-2) and Intel® Advanced Vector Extensions 512 (Intel® AVX-512) instructions. As an open source project, it will continue to track new and emerging trends in all major frameworks.

Alternatively, BigDL is an open-source distributed deep learning library for Spark* that can run directly on top of existing Spark or Apache Hadoop* clusters. Developed by Intel, it allows for loading of pre-trained Torch* models into the Spark framework, and can efficiently scale out to perform data analytics at big data scale.

Learn more about both of these solutions in our infographic The Anatomy of an AI Proof of Concept.

From data to insight to innovation to revenue: Intel® Xeon® Scalable processors

Intel® Xeon® processors are the heart of the data center, running the majority of the most critical and most innovative workloads. While existing Intel® architecture in your data center will allow you to get started with machine learning and deep learning, the new Intel® Xeon® Scalable processors offer the most agile Intel platform for AI, capable of taking your AI to the next level. In terms of the compute capability required for deep learning training and inference, the Intel Xeon Scalable processor delivers up to 2.2x performance compared to the previous generation for deep learning training and inference performance. And, benefitting from additional software optimizations – for example, TensorFlow*, Theano* and Torch* – can achieve up to 113x performance compared to non-optimized three-year-old servers for deep learning, providing a solid foundational architecture for AI workloads[1]. Six memory channels compared to previous generations' four makes for significantly increased memory bandwidth and capacity for memory-intensive workloads.

Transformational Readiness

The third (and possibly most important) area of readiness is whether an organization can maximize the value it gets from AI. Whether AI is being used for pattern recognition, insight generation or process automation, it will generally have one of the following impacts:

- Support better decision making, at senior or line management level
- Automate a part of a business process, or drive an automated response

In either case, AI can have a major impact on the day-to-day running of the business – which means that the business needs to be able to embrace resulting changes.

Strategic leadership

It is not essential that the organization's board sees digital technologies in general, and AI in particular, as a driver for business growth. However, it certainly helps if the right mindsets exist, to drive delivery from the very top of the organization. If AI is seen as a source of strategic advantage, priorities and budgets will be set accordingly.

Business opportunity

In the same vein, AI's chances of success can be improved if it unlocks new opportunities for business growth, new ways to engage with customers or new types of operational process. If this is the case, the organization should be looking to structure itself in order to leverage the AI-driven business opportunities that may exist. It is important to be clear on how existing and desired operating models can integrate the results and benefits of AI, through automation or augmentation.

Clarity of business case

A precursor to any business change is to have a clear picture of the benefits that change will bring. In Intel's experience, organizations earlier on in their AI journey have more focus on TCO: whether AI delivers the expected results (and saves money through automation) at an acceptable cost. More developed projects are looking to increase AI performance, and the most advanced are looking to see ROI in business terms, for example the amount of time that is released to perform other tasks. The business case should present clear, costed criteria for what constitutes success.

Business acceptance

The solution should be adapted to business needs right through to the daily activities of front-line staff, and the people impacted. Achieving acceptance may not always be straightforward, particularly if job roles and responsibilities change as a result of implementing AI.

Conclusion: Start Planning Now for AI Success

Whether your organization is at the beginning of its AI journey, or has already made some progress, you can ask yourself a number of questions to assess where you are and increase your readiness as a result. Some suggested questions that may assist your planning are below:

Using AI for the first time

- Is the scenario, use case or problem to be solved with AI clearly defined?
- Are priorities set around where AI will deliver the most business value?
- Is the planned infrastructure architecture clear and appropriate?
- Are all necessary data sources clearly understood and accessible?
- Can your chosen software packages deliver the AI solution end-to-end?
- Are sufficient skills and resources available (either in-house or externally)?
- Have expectations been set around training and learning times?

Making the most of your existing Intel® Xeon® processor-based infrastructure

Leveraging your existing data center infrastructure is an ideal opportunity to prove the value of AI to your business from a flexible, general purpose foundation.

Intel® Xeon® processors use a consistent infrastructure and programming model for existing analytics pipelines and support the large memory requirements of AI models. The previously mentioned software optimizations mean Intel Xeon processor-based infrastructure can continue to support an organization's AI journey from experimentation to proof-of-concept to production, depending on the workload.

Learn more about both of these solutions in our infographic The Anatomy of an AI Proof of Concept.

- Is the TCO of the end-to-end solution clear and signed off?

Scaling up use of AI

- Can the planned solution scale beyond initial testing and evaluation?
- Is a clearly defined business case confirmed with a business unit?
- Is sufficient direct resourcing available, with time allocated and reserved?
- Is network bandwidth sufficient to ensure timely data delivery at scale?
- Are operational management processes in place which cover AI delivery?
- Does the architecture align with industry standards and best practices?
- Has a cybersecurity risk assessment been undertaken and acted upon?
- Have realistic deployment plans been set and communicated?

Broadening out use of AI

- Is a team in place to oversee AI-based continuous improvement?
- Are the broader AI opportunities for the organization researched and clear?
- Are AI solutions developed and deployed following agile best practice?
- Are measures in place to monitor business effectiveness of AI solutions?
- Is the architecture for AI provided as a platform, rather than as one-off solutions?
- Are lines of business fully engaged in how AI will affect their processes?
- Are the governance needs of the AI solution clearly understood?
- Is AI seen as a central pillar of an IT-enabled business strategy?

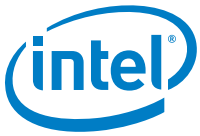
By addressing these questions, you will increase the probability of success – and will increase the acceptability of AI-based solutions in your organization. Getting everything right from the start is not a realistic expectation: rather, you should be looking to build skills and expertise as you discover the benefits of AI.

Most important is to have a clear grasp of the problem you are looking to solve. Establish a problem statement, then work to make the solution a reality. Whether it is automating processes or delivering insight, the ultimate goal of AI is to give yourself, and your organization, freedom to innovate and grow.

Learn More

To read more about the Intel AI portfolio and how it can support your journey to AI, visit: ai.intel.com

Intel's performance-optimized machine and deep learning libraries and frameworks are available here: https://software.intel.com/en-us/ai-academy



[1] INFERENCE using FP32 Batch Size Caffe GoogleNet v1 256 AlexNet 256.

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system. 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 Source: Intel measured as of June 2017 Optimization Notice: 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.

Configurations for Inference throughput
Processor :2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.46GB (12slots / 32 GB / 2666 MHz).CentOS Linux-7.3.1611-Core , SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework caffe version: f6d01efbe93f70726ea3796a4b89c612365a6341 Topology :googlenet_v1 BI OS:SE5C620.86B.00.01.0004.071220170215 MKLDNN: version: ae00102be506ed0fe2099c6557df2aa88ad57ec1 NoDataLayer. Measured: 1190 imgs/sec vs Platform: 2S Intel® Xeon® CPU E5-2699 v3 @ 2.30GHz (18 cores), HT enabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 256GB DDR4-2133 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.el7.x86_64. OS drive: Seagate* Enterprise ST2000NX0253 2 TB 2.5" Internal Hard Drive. Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact,1,0', OMP_NUM_THREADS=36, CPU Freq set with cpupower frequency-set -d 2.3G -u 2.3G -g performance. Deep Learning Frameworks: Intel Caffe: (http://github.com/intel/caffe/), revision b0ef3236528a2c7d2988f249d347d5fdae831236. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy 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 (GoogLeNet, AlexNet, and ResNet-50), https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19), and https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet benchmarks; files were updated to use newer Caffe prototxt format but are functionally equivalent). GCC 4.8.5, MKLML version 2017.0.2.20170110. BVLC-Caffe: https://github.com/BVLC/caffe, Inference & Training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was st ored on local storage and cached in memory before training BVLC Caffe (http://github.com/BVLC/caffe), revision 91b09280f5233cafc62954c98ce8bc4c204e7475 (commit date 5/14/2017). BLAS: atlas ver. 3.10.1.

Configuration for training throughput:
Processor :2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.28GB (12slots / 32 GB / 2666 MHz).CentOS Linux-7.3.1611-Core , SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework caffe version: f6d01efbe93f70726ea3796a4b89c612365a6341 Topology :alexnet BIO S:SE5C620.86B.00.01.0009.101920170742 MKLDNN: version: ae00102be506ed0fe2099c6557df2aa88ad57ec1 NoDataLayer. Measured: 1023 imgs/sec vs Platform: 2S Intel® Xeon® CPU E5-2699 v3 @ 2.30GHz (18 cores), HT enabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 256GB DDR4-2133 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.el7.x86_64. OS drive: Seagate* Enterprise ST2000NX0253 2 TB 2.5" Internal Hard Drive. Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact,1,0', OMP_NUM_THREADS=36, CPU Freq set with cpupower frequency-set -d 2.3G -u 2.3G -g performance. Deep Learning Frameworks: Intel Caffe: (http://github.com/intel/caffe/), revision b0ef3236528a2c7d2988f249d347d5fdae831236. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy 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 (GoogLeNet, AlexNet, and ResNet-50), https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19), and https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet benchmarks; files were updated to use newer Caffe prototxt format but are functionally equivalent). GCC 4.8.5, MKLML version 2017.0.2.20170110. BVLC-Caffe: https://github.com/BVLC/caffe, Inference & Training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was st ored on local storage and cached in memory before training BVLC Caffe (http://github.com/BVLC/caffe), revision 91b09280f5233cafc62954c98ce8bc4c204e7475 (commit date 5/14/2017). BLAS: atlas ver. 3.10.1.

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

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 www.intel.com/benchmarks

Benchmark results were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown". Implementation of these updates may make these results inapplicable to your device or system.

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