WHITE PAPER

Data Center
Artificial Intelligence

Select the Best Infrastructure Strategy to Support Your AI Solution

Buy, build, repurpose or outsource data center resources to implement image recognition, natural language processing or predictive maintenance workloads

Table of Contents

Introduction

For many organizations, the question is not about whether, but when, and most importantly how, to deploy Artificial intelligence (AI). As the focus of IT strategy moves from data management to intelligent action, enterprises increasingly recognize the role of AI to support humans in problem-solving, decision making and creative endeavors. AI systems learn from vast amounts of complex, structured and unstructured data, and turn it into actionable insights.

Enterprises recognize that implementing and using AI is critical for their continued growth in a competitive environment. Many potential opportunities exist, including:

- One-off, custom machine learning solutions to solve specific issues
- More generalized solutions to drive better decision making across the enterprise, from predictive to pre-emptive
- New opportunities using AI to drive innovation, make connections, identify and monetize new developments

To take advantage of the new and exciting AI opportunities, one of the first considerations is suitable infrastructure. AI solutions frequently demand new hardware and software, for example around collation and annotation of data sources, in scalable processing or creating and fine tuning models as new data becomes available. For any given AI solution, the options include:

- Repurpose existing hardware to deliver the AI solution at minimal cost
- Buy a one-off AI solution to address the needs of the use case only
- Build a broader platform that can support the needs of multiple AI solutions
- Outsource AI solution delivery to a third-party resource, including cloud

AI and the opportunities it represents are evolving rapidly, so budget holders may be reluctant to invest too heavily. A lack of in-house expertise, for example, can undermine solution delivery, creating the potential for reputational risk if mistakes or delays occur. And a lack of trust in the effectiveness of AI can create a substantial barrier to realizing its full value.

With these challenges in mind, the purpose of this guide is to help decision makers to select an AI infrastructure approach that accelerates adoption and enables experience to be gained without embedding cost or creating longer-term issues. In the following sections, we consider the merits of each option.

Repurpose existing hardware

Enterprises nearer the start of their AI journeys often look to use 'spare cycles' in their data centers to run AI workloads, or they develop solutions based on a single, 'spare' server or workstation node or a small cluster. Doing so enables you to:

- Test ideas and see what is viable for your organization
- Research hardware and software options
- Build skills and experience in a real scenario
- Engage with lines of business about the benefits of AI

The configuration for repurposed hardware will vary depending on the scenario: in the table we present one potential configuration, used as a basis for deep learning training and testing using Intel® Optimization for Caffe*.

Item	Model/Version
Hardware	
Intel® Server System	R1208WT
Intel® Server Board	S2600WT
(2x) Intel® Xeon® Scalable processor	Intel® Xeon® Gold 6148 processor
(6x) Crucial* 32GB LRDIMM DDR4	CT32G4LFD4266
(1x) Intel® SSD 1.2TB	S3520
Software	
CentOS Linux* Installation DVD	7.3.1611
Intel® Parallel Studio XE Cluster Edition	2017.4
Intel® Distribution of Caffe*	MKL2017
Intel® Machine Learning Scaling Library for Linux* OS	2017.1.016

The **benefits** of repurposing existing hardware to meet your AI needs are as follows:

- As it makes use of existing hardware resources, a repurposed solution has a low procurement cost of entry and can, potentially, be set up in minimal timeframes.
- The tight scope of a single-node or small-cluster solution enables research to focus on a tightly bound environment, focusing efforts on the AI itself, rather than (for example) network bandwidth or operational management.
- The repurposing approach offers an opportunity to use spare processing cycles across the infrastructure, which reinforces the benefit of AI as 'augmenting' existing capabilities.

The **disadvantages** of repurposing existing hardware to meet your AI needs are as follows:

- A single-use, minimal solution will not always integrate easily with broader solutions or user-facing tools, limiting its applicability, scope and longevity. This can also result in multiple technology 'silos'.
- Unless the available hardware is aligned to the need, you can incur overheads converting or redirecting less appropriate resources.
- Without controls in place, test configurations can become live configurations and may create more overheads as a result (and risks, for example if you 'borrow' resources to test a need, that need to be returned).

Is this approach right for you?

Based on the above, repurposing existing hardware is recognized as a useful shorter-term option. Some organizations have used servers destined as part of a data center refresh (as these are often procured en-masse and installed over time). You can look at this option as a way of convincing budget holders, but equally, you need to keep the longer-term in mind even if you are aiming to develop a short-term solution.

Buy a one-off solution

Many organizations we speak with are considering the option of procuring a custom solution to meet a well-defined use case. The common scenario is that, while decision makers can articulate one clear need, they may be less able to think beyond this one-off response (and towards more strategic use of AI across the business). This can be for several reasons, not least that developing a business case for broad AI adoption is significantly more involved than doing so for a specific requirement.

Figure 1 shows solution architecture which delivers on a well-bounded need, in this case predictive maintenance, and as such can be procured as a custom AI solution.

The example is built on servers using Intel® Xeon® processors, which are exceptionally well suited for inference-based AI models. Input is provided by a range of data sources and sensors, coming from existing systems and end-points – for example, manufacturing machinery, vehicle or building data. Enabling frameworks and software deliver training and inference capabilities: Intel aims to ensure that all major deep learning frameworks and topologies run well on Intel® architecture. This feeds inventory management software and visualization tools alike, via a web-based API.

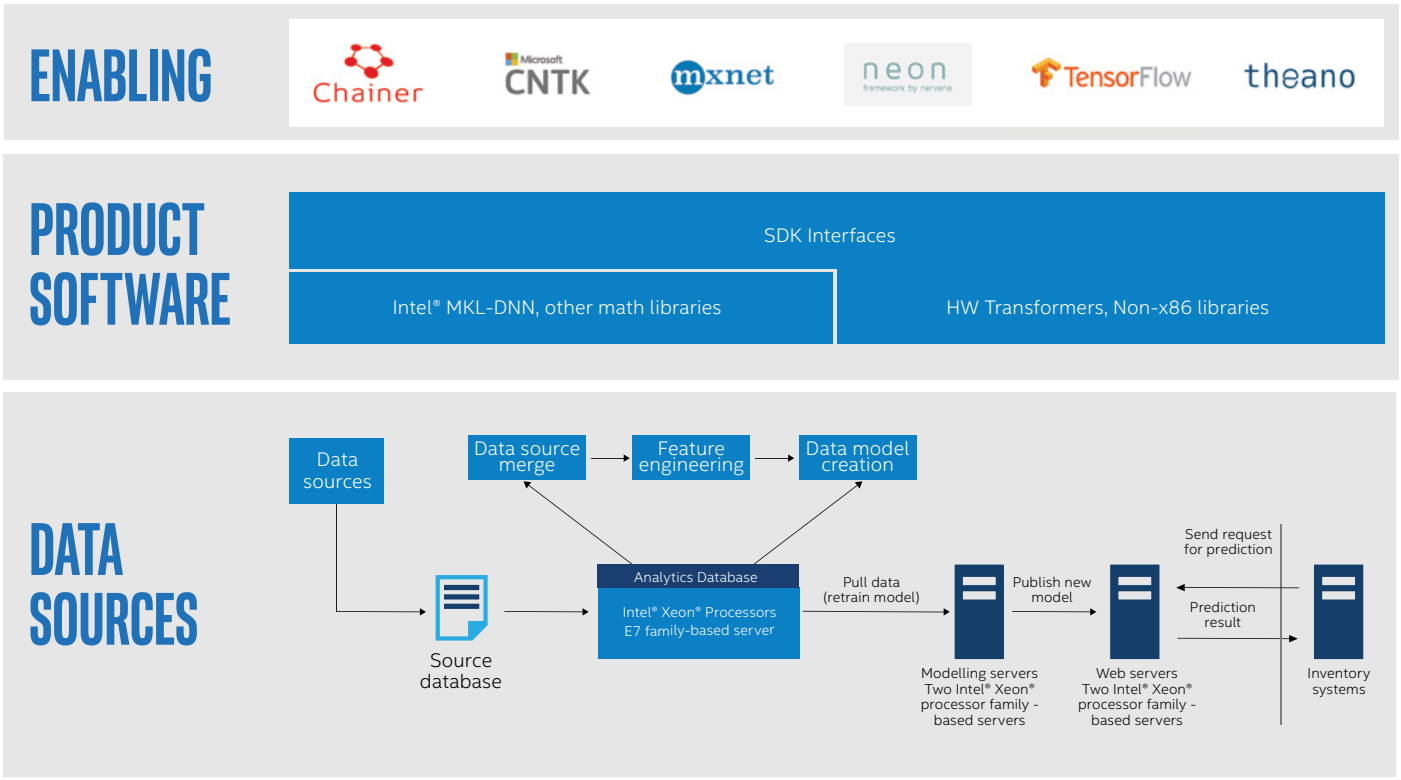


Figure 1. Hardware and software architecture for high performance AI use cases such as predictive maintenance

The **benefits** of buying an AI solution specific to a single use case are as follows:

- An off-the-shelf approach enables potentially faster deployment and adoption than a more generalized solution, because it only needs to support the needs of the lines of business and stakeholder groups concerned.
- The tighter scope means that it is easier to learn skills and expertise around AI, as the (potentially small) team can focus on optimizing a single solution, before concerning themselves with broader optimization and scale questions.
- A one-off solution may be the best option for specific use cases as it benefits from potentially increased efficiency and performance compared to solutions designed for multiple, shared-resource scenarios.
- The solution may be cheaper than an architecture designed to support a broader range of use cases, in terms of hardware costs.

The **disadvantages** of buying an AI solution specific to a single use case are as follows:

- The chosen solution may become outdated, if the solution-specific elements do not evolve in parallel with changing requirements of the scenario concerned.
- A one-off approach can result in multiple 'silos' of AI, where individual solutions exist in isolation of each other, and therefore need to be developed and managed in parallel.
- Single solutions may be cheaper individually but can prove more expensive overall, for example if the organization develops (potentially repeating) technology architectures.

Is this approach right for you?

Overall, the benefits of speed and specificity of buying a custom AI solution need to be weighed against potentially higher costs and the risk of ending up with multiple 'silos' of scenario-specific AI. This can be the case if the organization decides to adopt AI to a greater extent, following initial, successful forays.

To determine if this approach is for you, consider how likely it is for your organization to move towards broader adoption of AI, following this scenario. If you have already deployed one-off solutions, or if you are likely to in the future, you should perhaps be considering the benefits of building a broader platform. However, if you are still testing ideas, then repurposing existing hardware, or outsourcing deployment, might be better options.

Build a broader platform

Organizations with more experience of AI, or those who are looking to respond to needs across multiple areas of their business, may look to adopt a broader infrastructure solution that supports more general AI workloads. This approach is similar to the emerging 'platform' architecture we now see prevelant across IT – that is, an approach that

provides a highly scalable infrastructure layer that can be managed as a single pool, using virtualization and software-defined orchestration across server processing, storage and networking.

In the case of AI, this platform can be used with using a wide variety of open source and commercial software packages, configured to meet the needs of individual workloads. Figure 2 shows how this architecture can support a particular scenario, in this case facial recognition, based on the following layers:

- Hardware – comprising compute, input/output (I/O) and ancillary processing nodes, together with scalable storage and network connectivity. Communications between devices and systems is based around an ultra-high speed backbone such as the Intel® Omni-Path Fabric (Intel® OP Fabric).
- Software – made up of an operating system and virtualization layer, on top of which a library of AI-specific modules can run, enabling algorithmic processing and analytics, data management and I/O, as well as ingestion and delivery of data sources and visualization of analysis results.
- Process – this involves the 'business logic' of the AI application, using library modules to deliver such capabilities as facial recognition. The process layer takes into account both training of learning algorithms, and evaluation/inference of results.

The **benefits** of building a broader platform to meet your AI needs are as follows:

- A platform-based approach offers a single point of configuration and unique deployment target. It should therefore have a lower ongoing cost for general-purpose use of AI, through reduced management overheads.
- From a skills and experience perspective, while the platform may be more complex than a single solution, it provides a focus for building expertise.

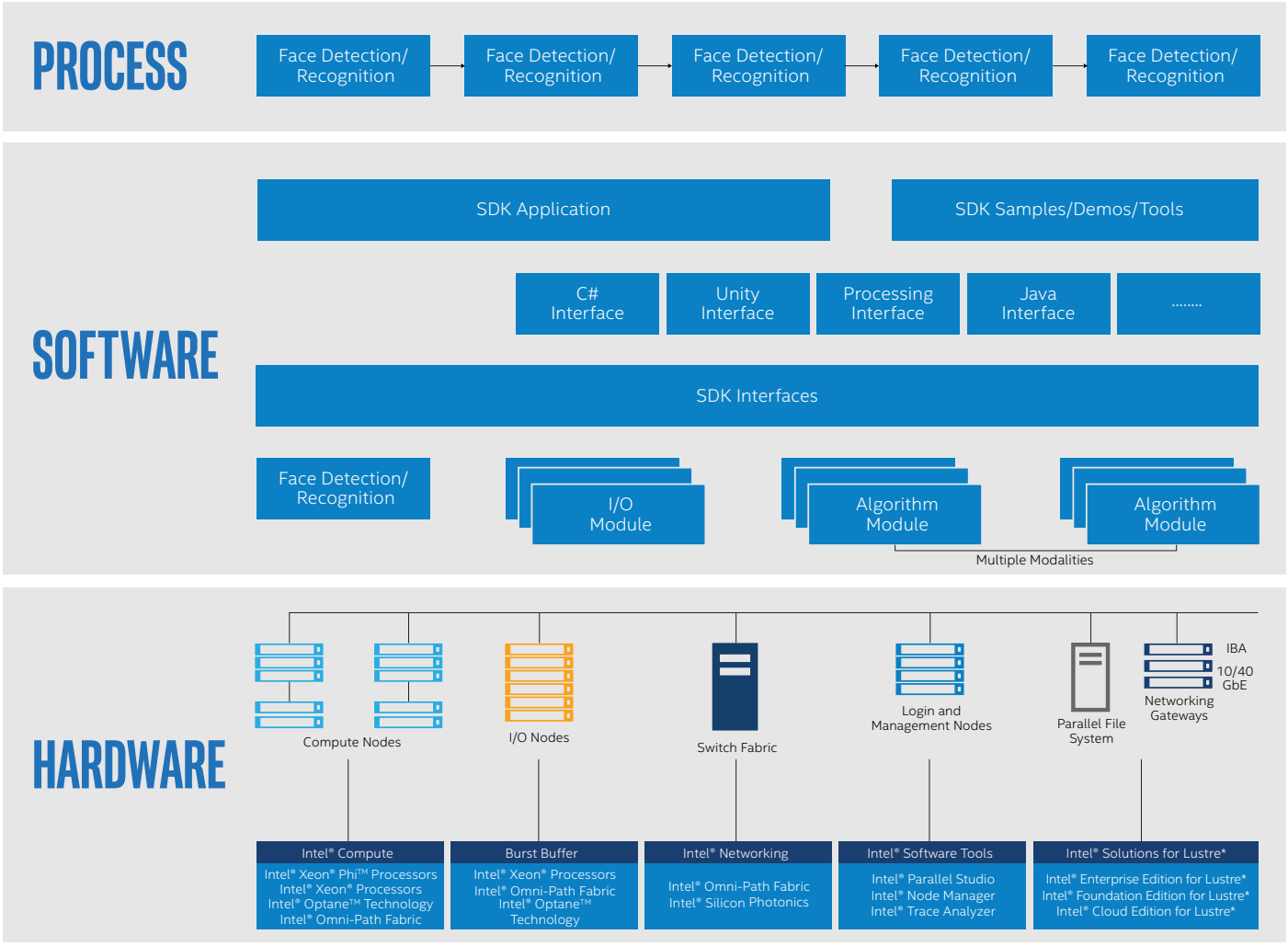


Figure 2. Hardware and software architecture for a broader range of AI use cases

- Organizationally, it can be managed by a single team rather than multiple teams, enabling improved communications both internally and with lines of business.

The **disadvantages** of building a broader platform to meet your AI needs are as follows:

- The initial build of a platform for AI may be seen as more complex, and costlier, than a single-use solution (note that it doesn't have to be – while it can be architected for broad use, it can be deployed as a smaller version initially).
- A broader platform would benefit from having in-house skills available, particularly as it is first configured. This may be a challenge for organizations and could increase deployment risk.
- Creation of a broader platform also means a greater potential risk should the architecture prove to be inappropriate – for example, it may be sized too small or too big for the actual need.

Is this approach right for you?

For organizations looking to expand their use of AI, building a broader platform makes sense. While its benefits remain true for those at an earlier stage of their journey, a lack of skills and/or inclination to embrace AI as part of business-as-usual may make a broader platform seem like too big a step.

To determine if this approach is suitable for your enterprise, you can test the need for AI across lines of business, for example running proof-of-concept studies (based on existing or outsourced infrastructure) to gain both experience and buy-in. If it still appears that your needs for AI are limited, you can look at a one-off solution. For further information about platform requirements for AI, across processing, storage and networking, see our paper on creating a proof of concept.

Outsource solution delivery

Organizations at various stages of their AI journey may look to use third-party resources (including cloud-based options) and skills, either to deliver a full-stack solution or to work with existing resources. Specific elements of an outsourced AI solution may include:

- **Infrastructure hardware** – pay-per-use infrastructure as a service, including GPU and SSD, can be allocated on demand
- **AI-specific software** – some providers now offer libraries of AI functions, including voice and image recognition
- **Data management** – service-based platforms provide a highly scalable basis for data collation and delivery

In-house engineers can work with a third-party provider to deliver a solution architecture that is entirely, or partially outsourced.

The **benefits** of outsourcing solution delivery to meet your AI needs are as follows:

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 Theano* and TensorFlow*.

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, Intel's BigDL is a distributed deep learning library for Spark* that can run directly on top of existing Spark or Apache Hadoop* clusters. 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.

- Capabilities may be available 'off the shelf', minimizing deployment and configuration issues
- Pre-developed options impose less overhead for new starters in terms of skills and resourcing
- Costs for pay-per-use services are bounded, fitting with situations where resources are difficult to pin down
- External services can be used to augment and test new solutions before bringing them in-house
- Organizations can benefit from third-party skills and knowledge

The **disadvantages** of outsourcing solution delivery to meet your AI needs are as follows:

- Managing the relationship with an outsourcer adds cost and reduces efficiency, particularly with lines of business, which can make it harder to drive innovation.
- The resulting infrastructure architecture may create data bottlenecks, depending on where data is sourced – for example, the organization may need to upload data from its internal systems into the cloud.
- It can be harder to build in-house experience and skills, particularly around solution architecture and data science, if solution delivery is outsourced to a third party. Opportunities to gain valuable knowledge may be lost.
- The cost of using outsourced resources may be greater over time, than running systems in-house.

Is this approach right for you?

Outsourced AI solutions have the advantage of lower cost of entry, and cloud-based options are a good fit for experimentation and shorter-term research. However, these advantages need to be balanced against longer-term costs, the importance of developing in-house skills and experience and constraints of working at scale. For example, if the scenario is particularly data-heavy (taking information from manufacturing systems for example, or within a retail environment) it may make more sense to use internal resources.

Deciding the Best Option

As these examples show, there is no "one size fits all" for AI solutions – each needs to be considered in terms of:

- Type, size and models of business
- Need and scope of use cases
- Availability of in-house infrastructure
- Skills, expertise and experience in both IT and lines of business
- Strategic view and level of commitment to AI
- Availability of data from internal or external sources
- Short- versus longer-term capacity planning

Many, if not all of these factors depend on where an organization is on its AI journey. Those looking to at least begin understanding the benefits of AI may see repurposing existing hardware, or making use of cloud services, as the simplest way to deliver value quickly. Organizations further down the line may look at buying a one-off AI solution aimed at a specific goal, and those looking at the longer term will see a broader platform for AI as most appropriate, particularly if dealing with large quantities of internal data.

As we show in Figure 3, this progression also aligns with increasing skills and experience, user trust and overall ROI of AI across the organization: these topics are also covered further in our paper The AI Readiness Model. While a broader platform may offer the most benefit in the longer term, it may also be too much to digest for an organization whose skills or trust levels are still being established.

Above all and whichever option is selected, the most important starting point is to understand what scenarios AI is looking to address. We would recommend that an organization consults with peers or speaks to experts before engaging in any procurement or deployment activity. And it is never too soon to start building skills and knowledge, for example by way of the Intel® AI Academy.

Considering running an AI proof of concept? Check out Intel's 'Anatomy of a successful PoC'.

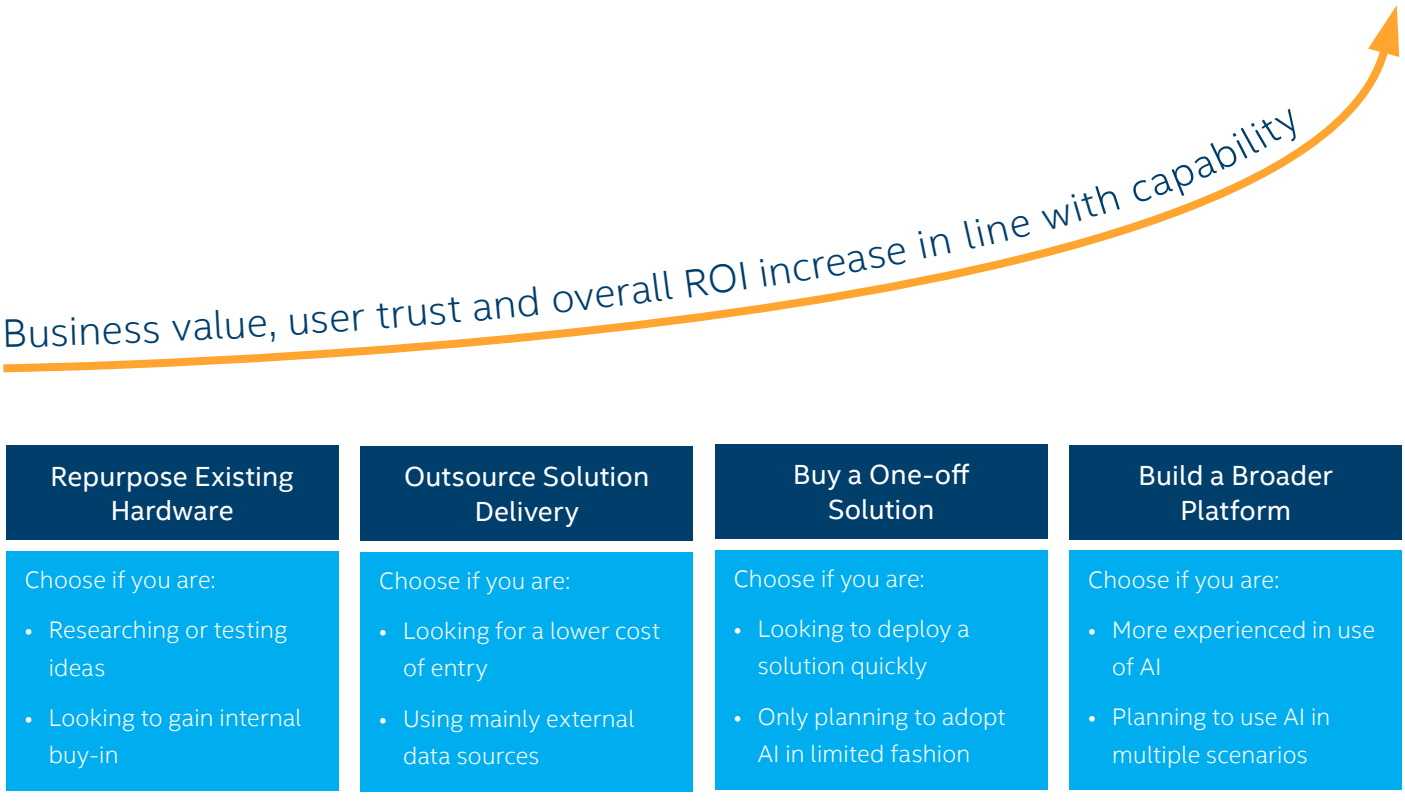


Figure 3. Business value, user trust and overall ROI increase in line with maturity

Learn More

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

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

References and Resources

Intel® AI Academy, https://software.intel.com/en-us/ai-academy

The Challenges and Opportunities of Explainable AI https://ai.intel.com/the-challenges-and-opportunities-of-explainable-ai/

The Future of Retail is All About Artificial Intelligence https://ai.intel.com/future-retail-artificial-intelligence

Intel® AI academy – learn the basics https://software.intel.com/en-us/ai-academy/basics

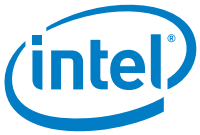
Loihi – Intel's New Self-Learning Chip Promises to Accelerate Artificial Intelligence https://newsroom.intel.com/editorials/intels-new-self-learning-chip-promises-accelerate-artificial-intelligence/

Partnership on AI, Yinyin Liu, Head of Data Science, Intel Artificial Intelligence Products Group, https://ai.intel.com/partnership-on-ai/

Intel® RealSense™ SDK 2016 R2 Documentation, SDK Architecture, https://software.intel.com/sites/landingpage/realsense/camera-sdk/v1.1/documentation/html/index.html?doc_essential_programming_guide.html

IT@Intel: AI Optimizes Intel's Business Processes: An Audit Case Study, White Paper, November 2017, https://www.intel.co.uk/content/www/uk/en/it-management/intel-it-best-practices/ai-optimizes-intels-business-processes-an-audit-case-study-paper.html

Deep Learning Training and Testing on a Single Node Intel® Xeon® Scalable Processor System Using Intel® Optimized Caffe, Intel AI Academy, https://software.intel.com/en-us/articles/deep-learning-training-and-testing-on-a-single-node-intel-xeon-scalable-processor-system



0318/RD/CAT/PDF 337358-001EN