THE JOURNEY TO AI-ENABLED SAAS

AN AI PRIMER FOR IT DIRECTORS AND OPERATIONS LEADERS

INTRODUCTION

Machine Learning (ML) and Artificial Intelligence (AI) will impact practically every application in every industry in the coming few years. Modern applications are increasingly delivered as cloud-based services and as a result, Software-as-a-Service (SaaS) and cloud-based application vendors can now deliver the competitive benefits of AI to their customers.

However, ML and Deep Learning (DL) in particular, are relatively new concepts for many developers. As such, many SaaS providers could use guidance in prioritizing the most important projects, curating the required data sets, building the needed core competencies, and creating the AI models and infrastructure to run them.

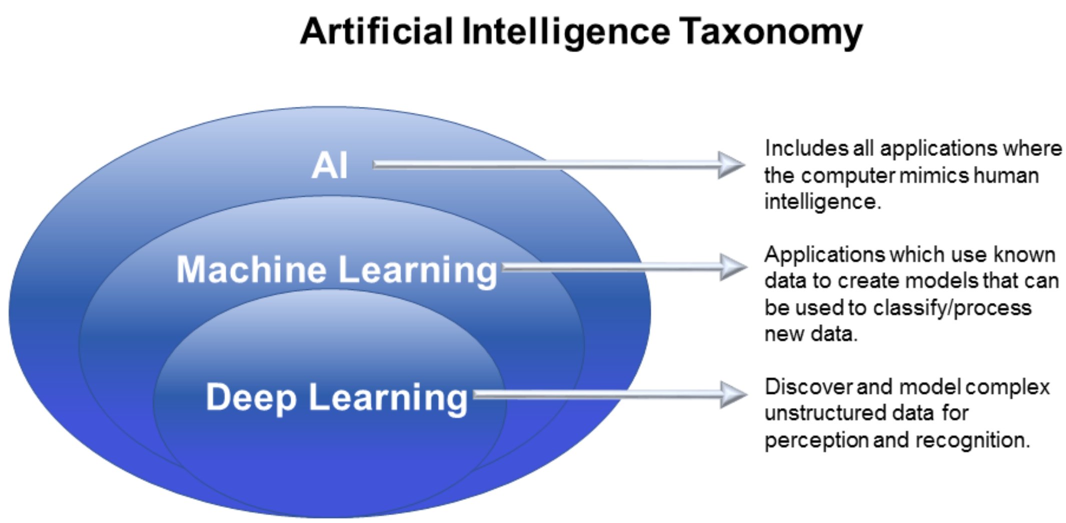
This paper introduces the key concepts of ML and AI, explores common use cases in SaaS, and outlines the required technologies and skills for SaaS application professionals. The paper is not an exhaustive reference, but rather is intended to help business and technical decision makers understand the art of the practical today. This includes best practices and ideas for developing and/or acquiring the prerequisite talents and technologies to realize the benefits of AI for SaaS business.

Organizations are now focusing on AI as a disruptive technology with the potential to transform the computing industry, change the way businesses program, reimagine the hardware used, and redefine what one can expect a computer to accomplish.

TAXONOMY: ARTIFICIAL INTELLIGENCE (AI), MACHINE LEARNING (ML), AND DEEP LEARNING (DL)

AI is a branch of computer science dealing with the capability of a machine to imitate human intelligence. This broad term can be used to characterize something as “simple” as recognizing an object in an image, to the ability to apply vision and decision-making in real time to guide vehicles and drones. While there has certainly been a tremendous amount of hype around this technology, its impact on science and business will indeed be dramatic.

FIGURE 1: THE RELATIONSHIP BETWEEN AI, ML, AND DL.



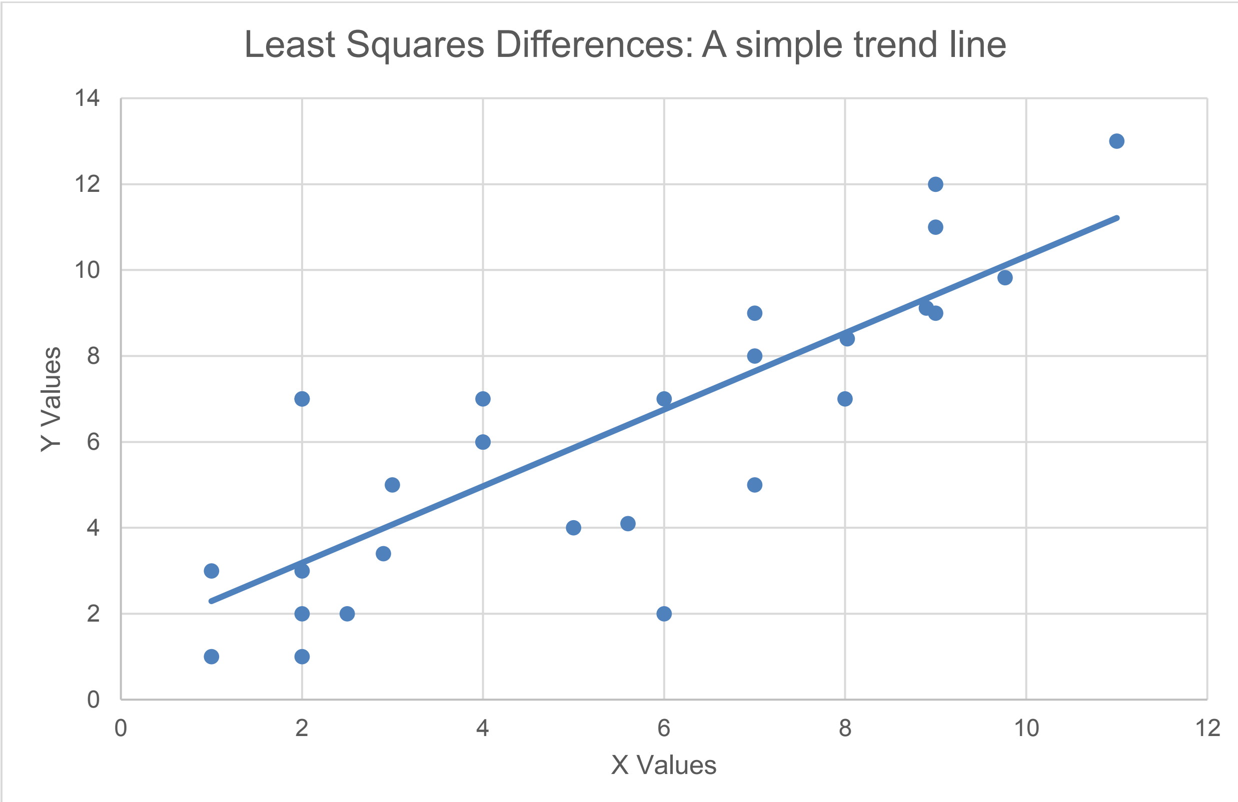
Source: Moor Insights & Strategy

MACHINE LEARNING

ML is a specific field of AI that evolved from the study of pattern recognition in data whereas DL is a branch of ML that uses deep neural networks (DNN) to “learn” to classify input data from large data sets. “Deep” refers to the number of layers in the neural network, an advance made possible by significant leaps in computational capacity. The most widely used ML approaches use statistical analysis techniques such as linear and logistic regression, decision trees, random forests, support vector machines, gradient boosting, and similar iterative classification techniques. Typical use cases include resource allocation, predictive analytics, predictive maintenance, text classification, trend identification, forecasting, face detection, and pricing optimization.

These statistical approaches are generally well-suited to identify trends and categories in numerical data and do not typically require massive training datasets nor hardware accelerators (although the recently introduced RAPIDS software from NVIDIA can now accelerate many of these codes). ML offers a far easier path to building “smart” applications than DL for many numerical problems and is fairly resilient when dealing with missing or noisy data.

FIGURE 2: LINEAR REGRESSION FITS A LINE OR A CURVE TO DATA POINTS USING SUM OF LEAST SQUARES DIFFERENCES.



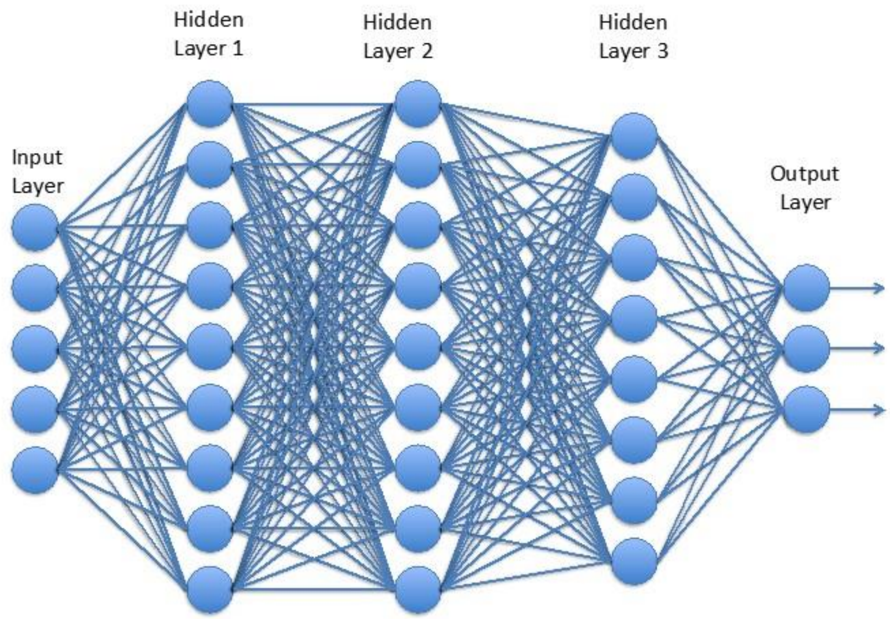
Source: Moor Insights & Strategy

DEEP LEARNING

DL is responsible for much of the recent attention, investment, and publicity for AI. DL is now widely used in hyperscale datacenters and mobile devices for computer vision, natural language processing, social network processing, autonomous driving, object detection and classification in image and video, financial asset modeling, and many other application areas in science and business. Its relevance exploded as two factors converged: the availability of adequate data sets to build these networks and the immense computational power of graphics processing units (GPUs) needed to train them.

DL is used primarily for unstructured data and is primarily based on variants of DNNs, which "learn" from massive datasets using "tagged" data. In general, DL is used to solve multi-variant problems and queries that are beyond the scope of the traditional ML techniques. The network nodes contain "weights" which are iteratively tuned using forward (guessing) and backward (correcting) passes through a large training dataset.

FIGURE 3: A TYPICAL DEEP NEURAL NETWORK



Source: Moor Insights & Strategy

This approach and many others detailed in this paper are described as "supervised learning". However, recent advances in unsupervised learning (using techniques such as reinforcement networks) are starting to show significant progress where DNNs come up short. Reinforcement learning (RL) is a neural network that is trained by trial and error. This is ideal for creating automated actions such as those found in gaming and used by Google to defeat the world champion GO player in 2016.

PROJECT BRAINSTORMING: FINDING POTS OF GOLD (AND AVOIDING RAT HOLES)

Most SaaS providers begin this journey with a lengthy list of possible projects from service chatbots to product enhancements that eliminate work for their application users. Below are a few ideas to guide management teams on their initial journey into ML and DL:

1. First, tap into leading customers for direct “voice of the customer” input on what they would find to be of value.
2. Start small, perhaps by extending existing functionality with ML techniques. Moonshots are rarely successful unless one has significant expertise and a clear line of sight to a compelling solution and ROI.
3. Deliver customer value by automating tasks customers can avoid. Look at the data and processes customers use in a SaaS or application and consider how these could be enhanced through analytics with ML or DL.
4. For a first DL project, consider starting with a well-defined area backed by existing data, such as adding speech recognition and natural language processing as input modes for existing applications.
5. If possible, acquire adequate data to enable the training of DL networks and have the right hardware infrastructure to support project(s).
6. Learn from analysis of competitors’ SaaS enhancements: what do customers see that would be valuable to add?

AI in Action at Pure Storage

The Pure Storage “Sev1” support team wanted to know how an array’s performance changes in response to the workload being run on it in order to help customers predict and prevent failure caused by reaching the performance limits in the Pure FlashArray. Using the principles above, Pure Storage tackled the problem creating Pure1 Meta to help customers improve availability and performance of their flash storage arrays and gain a better understanding of their workloads. Lean more at this blog: https://blog.purestorage.com/closer-look-machine-learning-pure1-meta/.

Areas Where AI Can be Used by SaaS Providers

Below are a few areas where AI/ML are being deployed, specifically for SaaS organizations. The common foundation throughout these examples are the datasets from which the models can be created. Data typically needs to be prepared for processing, cleansed of questionable, low-confidence data points, and tagged for DL. DL places the most demands here, requiring massive datasets with which supervised learning can take place. Additionally, it is critical in all ML projects for the data to be inclusive in order to avoid the pitfalls of data biases.

Personalization

AI can be used to personalize the customer experience to each user’s use patterns and recommend specific actions they can take. These areas represent low-hanging fruit –

using pre-trained neural networks for voice input and applying understanding of a user's workflow and data. Consider analyzing historical data to create a recommendation engine relevant to a SaaS domain that can suggest next steps and ideas.

AUTOMATION (E.G. OF CUSTOMER SUPPORT)

Chatbots can be used to pre-screen customer support inquiries, either to resolve the customer's query or to better inform the customer support representative. Chatbots can be natural language (text) input or voice input, but must be carefully built and thoroughly tested to avoid creating a poor customer experience. Of course, these technologies can also add language translation to an organization's customer services, enabling more efficient global support services. Chatbots-as-a-service are a relatively easy approach to enhance customer services interactions and can represent additional sources of revenue for some SaaS providers. Another area of automation worth exploration is providing suggestions for field input and actions in forms.

PREDICTIVE ANALYTICS

Predictive analytics is the use of data, ML (statistical algorithms), and more recently DL to estimate probable future outcomes based on historical data. Predictive analytics has been around for decades and its usage continues to expand across many industries and application domains. Common uses include fraud detection, device failure prediction, marketing campaign optimization, improving operational efficiencies, and risk reduction. Predictive analytics might be an area ripe for development focus and enhancement on an organization's platform – as these areas touch nearly all aspects of a business and great deal of analytic software exists.

Three of the most widely used predictive modeling techniques are decision trees, regression, and neural networks. Decision trees are used to partition data into categories that are most different from each other and are popular because they are easy to understand and interpret. They also handle missing values well and are useful for preliminary variable selection.

While decision trees are used to predict whether a data sample is or is not a member of a category, regression analysis (of which there are now at least a dozen algorithms) can predict a numerical value (e.g., the y axis) of a continuous function being applied to an input variable. Regression techniques, long a staple in statisticians' toolboxes, have expanded into new areas to find relationship patterns in large and small data sets. These tools are often used to determine how much specific independent factors, such as temperature, price, or age, influence the movement of a dependent variable of interest, such as purchase volume or probability. The use of neural networks in predictive analytics is still relatively new, but advocates claim it can deliver excellent accuracy with relatively small datasets.

Pricing Optimization

If an organization's SaaS offerings include enterprise resource planning (ERP), e-Commerce, and marketing services, customers can optimize their pricing to reach company objectives with pricing modules built on ML and DL. Current tools allow retailers to set initial, discount, and promotional prices by accounting for factors such as competition, weather, season, costs, and local demand patterns.

There has been much discussion in the industry around the impact AI may have on traditional per-seat SaaS pricing. As AI improves the individual's efficiency, the argument goes, companies will buy fewer seats, all while the SaaS provider spends more money to develop the AI-enhanced service. Clearly, one solution here is to move to a usage-based pricing model, where the customer's fees are commensurate with the value received. Venture Beat has published an interesting article on this topic here: https://venturebeat.com/2017/03/18/ai-is-going-to-kill-seat-based-saas-models/

As with predictive analysis, most of these tools can be built using general linear models such as regression. The use of DNNs in this area is still in its infancy.

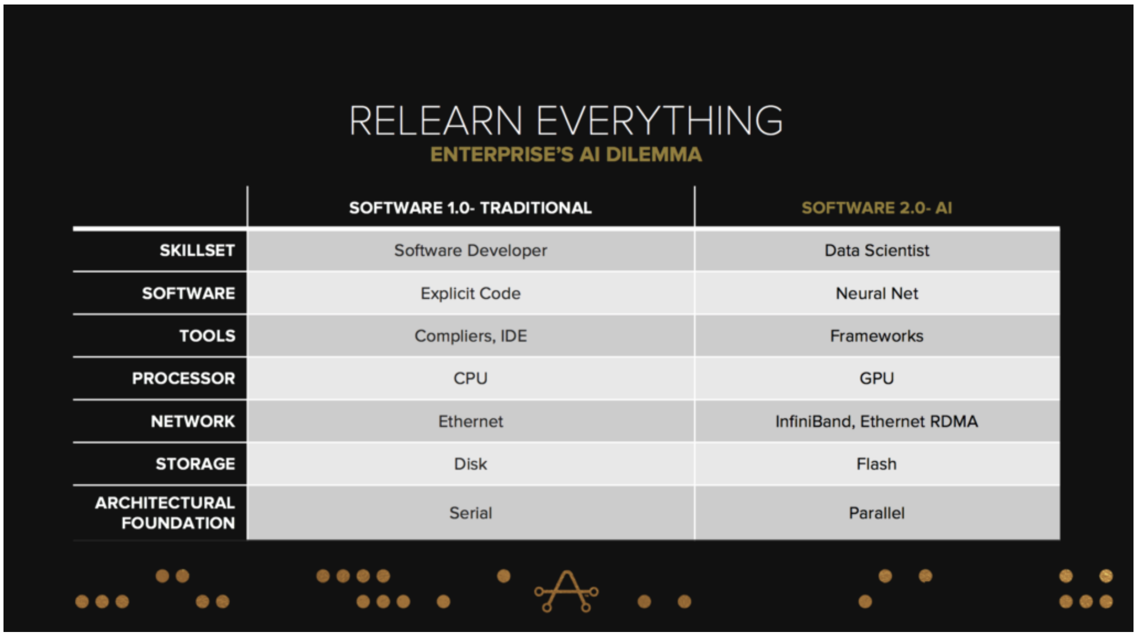
Enhanced Development

The world of software development is another area where AI is having a dramatic impact for businesses and their customers. AI improves the traditional development process for C++, Java, and Python teams – as it completely transforms some software by replacing the traditional logic-based programming of the last five decades with human-like intelligent data processing.

In traditional procedural or functional programming, say in C++ or Python, ML can speed prototyping, provide programmer assistance with access to documentation and code suggestions, analyze errors, create test scenarios, and optimize code for portability and maintenance.

The transformation of software development using DNNs has just begun, but could completely revamp programming in the fields of computer vision, gaming, speech processing, robotics, natural language translation, and even databases. Andrej Karpathy, head of AI at Tesla, opines that "Neural Networks are not just another classifier, they represent the beginning of a fundamental shift in how we write software. They are Software 2.0".

FIGURE 4: AI IS USHERING IN A NEW ERA, WHICH MANY CALL SOFTWARE 2.0.



RELEARN EVERYTHING
ENTERPRISE'S AI DILEMMA

	SOFTWARE 1.0- TRADITIONAL	SOFTWARE 2.0- AI
SKILLSET	Software Developer	Data Scientist
SOFTWARE	Explicit Code	Neural Net
TOOLS	Compliers, IDE	Frameworks
PROCESSOR	CPU	GPU
NETWORK	Ethernet	InfiniBand, Ethernet RDMA
STORAGE	Disk	Flash
ARCHITECTURAL FOUNDATION	Serial	Parallel

Source: Pure Storage

ENHANCED SECURITY

AI is being used to protect physical and IT infrastructures and applications from malicious attacks. Cameras and sensors are equipped with facial recognition chips for securing public and private venues, while ML and DL is being applied to detect digital threats. Application security remains challenging to realize with ML as the training datasets are typically unavailable or quite expensive to create. On the other hand, using ML to detect anomalies in infrastructure is fairly advanced as numerous scholarly and commercial publications exist that can help one understand how ML can help protect businesses. Security-as-a-service providers like FireEye provide solutions in this area.

ORGANIZATION AND TALENT CONSIDERATIONS

SPONSORSHIP

AI projects can be somewhat controversial since they tend to grab attention and reprioritize resources. With this in mind, it is helpful to identify a passionate high-level executive sponsor to launch and steer the effort across the organizations that will contribute people, data, and/or budget. Having a formal project kickoff, called and facilitated by the sponsor, will help get everyone on the same page and make the project's business goals and schedule clear to everyone involved.

TALENT SOURCING

Securing the necessary talent for AI can be one of the most challenging hurdles in embracing Software 2.0, but there are resources and several staffing options which can help clear this hurdle.

Teach Yourself

Building in-house ML and DL expertise is perhaps an inevitable requirement to remain competitive, even if one chooses to start with outsourcing. There are several good sources for learning materials, including the suppliers of ML software and online courses.

ML open-source repositories and independent software vendor (ISV)solutions are popular and include extensive training and documentation resources. Open-source DL framework repositories also have rich training, example code, and documentation and enjoy a large pool of contributors who can help along the way.

There are also excellent online education offerings by Coursera, Udacity, and Udemy. Many of these are taught by some of the leading researchers in DL, such as Geoffrey Hinton, Yoshua Bengio, and Andrew Ng. Moor Insights & Strategy (MI&S) highly recommends these classes as they can ensure an organization's team truly understands the underlying mathematics, use cases, and limitations of the various methodologies available.

Hire the Talent

Hire the talent needed, possibly augmenting hard to find data scientists with data engineers. While the search can be challenging, there is really no substitute for having strong technical leaders on board that are committed to success. Consider offering leading engineers the opportunity to enhance their skills and grow with the organization, working alongside new hires.

Rent the Talent

Smaller companies often retain the services of consultants to help bring their team up to speed and get the project started. This can be a good way to start, but long-term success will ultimately depend on having a strong in-house team that understands the business, has close customer relationships, and thoroughly understands the application space. Solicit input from prior clients to ensure said consultants understand AI and have a solid track record of producing quality results.

While outsourcing early projects can speed time-to-market, it will be critical to transfer knowledge used to build early AI features to a permanent in-house team. To accomplish this while moving forward quickly, a hybrid strategy could be a great option, outsourcing with partners while building up an in-house team. A blog post on Dataiku builds a case for such a strategy: https://bit.ly/2F7HFhC.

A second article published in the *MIT Sloan Management Review* also speaks to outsourcing analytics where scholars share their perspective that "In contrast to analytically challenged companies, which are usually happy to outsource their analytics requirements, analytically superior companies wanted to expand their internal analytic capabilities". Link to article here: https://bit.ly/2LPBP4Y.

Below are four consulting firms that focus on building ML applications:

- https://appliedai.de/about/
- https://www.elementai.com/en/what-we-do
- https://algorithmia.com/
- https://www.cognitivescale.com/get-started/

Final Project Selections and Priorities

Return to project brainstorming work and prioritize projects once established leaders are equipped with the knowledge of the business and application services and of AI, ML, and DL.

It is critical to select projects for which an organization or their clients have access to adequate data to build the AI models. Per usual in this phase of planning, evaluate each project's expected costs, timeframe, return, and risk.

The process of final project selection can be the catalyst for an executive sponsor to gain buy-in from the organization's thought leaders and stake holders, such as finance, marketing, development, and sales. Once all stakeholders are on board, get budgets and critical success factors lined up, define ROI metrics, gain final approval, and formally kick off the AI Journey.

Technology Selection and Acquisition

Cloud Vs. On-Prem Hosting

For ML and DL projects, consider the benefits and costs of both cloud-based services and on-premises solutions. In the cloud, Google, Amazon AWS, and Microsoft all provide AI tools and hardware. Microsoft also offers application programming interfaces (APIs) to access pre-trained neural networks, some of which can be customized with specific data such as vocabulary and images. Google and Microsoft both offer "automatic" model builders, which promise a quick route to model development. As these "Auto ML" tools are in their infancy, MI&S suggests they be used judiciously, perhaps to give an organization a first approximation of a model.

On-premises model development and deployment is often the better approach for production AI work, especially if it is going to become a major company emphasis with significant investments in compute cycles and data acquisition. Frequently, companies wisely choose to initiate early model development using cloud-based ML hardware and software services, avoiding the time and expense required to configure and manage GPU clusters and software stacks. As model development progresses, many of these companies then choose to move production ML to on-premises infrastructure to minimize data movement, lower hardware costs, and improve productivity as use cycles increase.

ML and DL Hardware

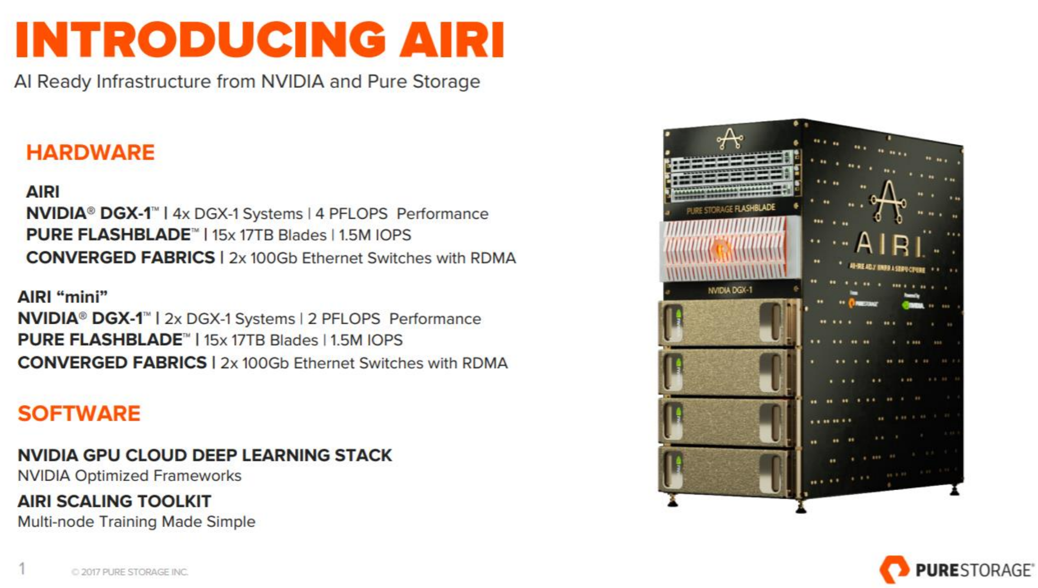
ML code typically runs on standard or high performance 2-socket servers with ample memory and storage capacity. ML tools do not require a great deal of training data or hardware acceleration.

However, DL training demands high performance GPUs to accelerate the massive computation required to build DNN models. DL inference queries (the use of trained neural networks) usually runs fine on standard servers. However, complex data types such as computer vision and natural language processing require accelerators (GPUs). DL also needs lots of data, which is often sprawled across several silos such as SQL, Hadoop, Elastic clusters, etc. As a result, data engineering and data science teams are always moving data sets around rather than focusing on model development. As Andrew Ng, co-founder of Google Brain and former vice president and chief scientist at Baidu, stated, "First thing that any organization looking to power their businesses must do is to consolidate their data lake/warehouse."

Pure Storage has a strong portfolio of compute and storage solutions that can meet performance needs and budgets and the company has considerable experience tailoring these systems for AI projects. Pure Storage FlashBlade provides a consolidation solution (data hub). As a solid-state array, it can handle many different workload types, large and small files, and random and sequential data access, reducing the need for silos that are specialized for workloads.

The company also collaborated with NVIDIA, the leader in GPUs for AI acceleration, to engineer AIRI™, the industry's first AI-ready infrastructure. AIRI is designed for integrated all-flash storage, central processing units (CPUs), GPUs, high performance networking, and the NVIDIA DL software stack, which meets the demanding performance requirements of DL.

FIGURE 5: AIRI FROM PURE STORAGE



Source: Pure Storage

For on-premises hardware, organizations will want to select vendors with significant experience in AI projects as they are better able to help size and configure servers, accelerators, storage, and networking to give them the best performance for their money. Major hardware vendors also offer consulting and training services to help with AI project planning and implementation.

ML AND DL SOFTWARE

There are three sources of ML software to build and manage a ML project: open source code, ML cloud services (MLaaS), and ISV solutions. Software selection should be based on the type of data (structured or unstructured, numeric or images, language, etc.), existing contracts/services, and the problem at hand.

For projects working with classical ML, open source tools such as SPARK MLib, R-Server, Hive, SciKit-Learn, and NVIDIA's Data Science Stack provide state-of-the-art open source solutions and come complete with examples, documentation, tutorials, and large user support communities. Notebooks such as Jupyter can help manage the development lifecycle and are widely used. Common algorithms include decision trees, regression, Bayesian analysis, and support vector machines, among others. For organizations with ISV licenses and support agreements, robust ML solutions from software companies like SAP, SAS, and Oracle provide tools and training as well as extensive support and implementation services.

Nearly all production projects building DNNs now use open-source AI "frameworks". These tools are typically championed by universities and large search, e-Commerce, and social networking companies. Each framework has its adherents and supporters and often favors a data type or use model (DNN, convolutional neural network, recursive neural network, reinforcement learning, etc.) for which the framework was originally developed for internal use cases. Google's TensorFlow, Amazon's MXNet, Facebook's CAFFE2 and PyTorch (which are merging), and the high-level Keras framework are a few of the most popular frameworks used to develop and train a variety of DNNs. The Pure Storage AIRI infrastructure comes with a full stack of optimized software from the NVIDIA GPU Cloud container repository and the AIRI toolkit for multi-node training at scale.

Once a model is developed, the DNN's execution graph and "weights" should be optimized for subsequent inference processing on specific chips. Two notable tools here are NVIDIA's TensorRT and the open-source ONNX, the interchange format recently introduced by Microsoft, Amazon, and Facebook to provide a framework-independent DNN format that can be transferred and compiled to run on different devices. Intel's OpenVINO tools can optimize inference for execution on specific Xeon processor models, including Xeons with the newly announced DLBoost feature, and eventually the Intel Nervana neural network processors (expected in late 2019 or 2020).

Similarly, predictive model markup language (PMML) provides an open standard extension to XML on the machine learning front and is commonly used to transfer machine learning models between frameworks such as Apache Spark's Mlib and the execution environment.

DATA PREPARATION

Preparing data set(s) for machine learning model development is a critical and time-consuming process. There are a number of good sources here that can be helpful, but below are basic steps that should help organizations understand the task that lies ahead for their data scientist.

- Articulate the problem statement carefully. Knowing precisely what one is trying to predict is essential as it will inform exploration for available relevant data in the organization, determine if one is trying to classify (yes/no), cluster into categories, perform regression, rank results, or tackle deep learning. As a rule, keep it simple.
- Examine relevant data sets in the organization and determine if they are adequate. Some organizations determine at this point that they need to change the data they routinely collect in order to feed the machine learning process with adequate features.
- Format the data to make it consistent at the record level, especially if the data comes from different databases.
- Reduce the data through attribute sampling (selecting the relevant attributes) and record sampling (eliminating samples with missing, erroneous, or less representative values). Consider aggregating similar records into broader record types where the detail is too fine grain. "Outliers" can at times significantly improve prediction accuracy, so prune data sets judiciously and retain the ability to backtrack as the datasets evolve.
- Clean data and take care to fill in missing data values (perhaps with mean values) as unintentional zeroes can skew the model results.
- Normalize (rescale) data, if necessary, to ensure large values do not overshadow smaller values.
- Consider discretizing data, turning numerical values into ranges or categories.

MODEL DEVELOPMENT AND DEPLOYMENT

Once the problem statement is clearly defined, goals and metrics have been established, and a core team built, it is time to start building AI. Model development is an iterative process and can take considerable time and data to get the accuracy and consistency needed. Below is a broad overview of the process for developing deep learning networks.

1. First, and perhaps most importantly, gather and prepare the data as discussed.
2. Next, or in parallel, select the algorithm(s) that can best meet the organization's needs, possibly two or three and work iteratively to determine which algorithm works best with the data and problem at hand.
3. Continue to iterate and tune the model and hyper-parameters until the desired results are achieved. Deep learning dashboards can help track progress and backtrack when needed.
4. For deployment, continue to collect new data and constantly review model accuracy to ensure the model continues to learn and improve.

Once satisfied with the ML or DL models, it is time to deploy the feature in the SaaS offerings. Many new features will require a new software release, such as voice or data input features. However, some AI features can be rolled out as standalone modules, generating additional revenue. For example, ML-driving pricing or forecasting optimization. For integration into existing applications, ML and DL calls to framework APIs are added to the code, looking like familiar SQL query calls.

CONCLUSIONS AND RECOMMENDATIONS

AI is already changing many businesses and the software they depend on to run their operations. SaaS providers have a choice. Organizations can lead, choose to keep up, or fall behind in smart competitive features their customers will demand for their own competitiveness and efficiency. If built with the required skills and data sets and started small with a focus on creating tangible business value, organizations can create valuable offerings with AI that can thrive in the new era of smart applications and data-driven Software 2.0.

Pure Storage recognizes the scope of the challenge and the potential business impact of adding AI features into SaaS businesses. The company has demonstrated its ability to help customers realize their potential with innovative solutions like the new AIRI infrastructure offering for deep learning.

MI&S recommends harnessing the power of AI in an organization's SaaS business and partner with companies like Pure Storage to ensure the best infrastructure is in place to support these efforts.

Important Information About This Paper

Contributor

Karl Freund, Senior Analyst at Moor Insights & Strategy

Publisher

Patrick Moorhead, Founder, President, & Principal Analyst at Moor Insights & Strategy

Inquiries

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

Citations

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

Licensing

Disclosures

This paper was commissioned by Pure Storage Inc. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.