

White Paper

Accelerating the AI Journey with Cisco

Powering the AI Data Lifecycle from End to End

By Mike Leone, ESG Senior Analyst; and Leah Matuson, Research Analyst
August 2019

This ESG White Paper was commissioned by Cisco
and is distributed under license from ESG.

Contents

Introduction	3
AI Is a Priority	3
The Skills Gap	3
Data Pipeline and Technology Stack Complexity	4
Cisco’s Unified Architecture.....	6
Simplicity and Manageability Across the Data Pipeline.....	6
Enabling AI through Full-stack Curation.....	7
Cisco Solutions and Services	8
Deep Learning with Cisco.....	8
Systems Management with Intersight Unified Management Platform.....	9
The Bigger Truth.....	9

Introduction

Many organizations recognize the value of AI and how the next-generation technology can enable digital transformation unlike anything in the past. AI can help increase efficiency by automating business processes, decrease operational expenses, provide timely and valuable business insights from data analysis, and enhance the customer/user experience—all of which converge into positively impacting the bottom line.

But while businesses across the board are viewing AI as a lifechanging technology, there are many barriers to adoption. Whether due to skills gaps throughout organizations, copious challenges across the data pipeline, or ineffective data management and orchestration tools across hybrid cloud environments, organizations fear their AI initiatives will not be achievable. Organizations need assistance in determining how AI can assist different teams and business units. They must be able to identify AI business use cases, understand how to grow diverse data sets, ensure efficient data pipelines with intelligent automation, and enable personnel within an organization to remain productive in their specific roles.

Because AI is brand new to many organizations, companies are turning to vendors that can enable them to succeed across the entire data pipeline by offering a proven infrastructure foundation that unites operational silos, reduces complexity across the AI infrastructure stack, and improves data lifecycle efficiency.

AI Is a Priority

With nearly one in three organizations (32%) surveyed by ESG citing the need to improve data analytics for real-time business intelligence and customer insight as one of the business initiatives expected to drive the most technology spending at their organization over the next 12 months,¹ it's no surprise these organizations are turning to AI to enhance operational efficiency, improve customer satisfaction, and gain further predictive insights into future business scenarios or outcomes.

Based on ESG research, 30% of organizations view AI as one of the areas of data science and analytics in which they will make the most significant investments over the next 12-18 months.² Just how much will those investments be? Of those organizations currently utilizing AI to some extent or expecting to utilize it, 41% will spend at least \$500,000.³ And while previous investments in big data and analytics solutions may impact that spend, a majority of organizations will look to leverage those previous investments as part of their new AI/ML initiatives. Customers across industries are in various stages of their AI journey—from identifying business use cases and evaluating infrastructure solutions from both hardware and software perspectives, to ensuring data science tools and platforms properly align with AI objectives and requirements, and eventually deploying AI into production while fitting the agile, real-time needs of the business.

The Skills Gap

Due to AI's enormous potential to disrupt markets, organizations are jumping in, even though they may not currently employ the appropriate personnel who can support the company's AI initiatives. And while 39% of organizations are looking to hire internally to address AI skills gaps, many are not making the most appropriate choices for those who will work on AI initiatives. Inevitably, this leads to growing numbers of employees not being adequately trained to perform in AI-related positions.

In fact, of those organizations that plan to develop or already have a specialized AI infrastructure, ESG research shows 38% are lacking a key AI employee role—the data scientist—which has numerous ramifications. In those organizations without data scientists, 63% of employees are being asked to fulfill tasks outside of their core expertise. And in the cases where

¹ Source: ESG Master Survey Results, [2019 Technology Spending Intentions Survey](#), March 2019.

² Ibid.

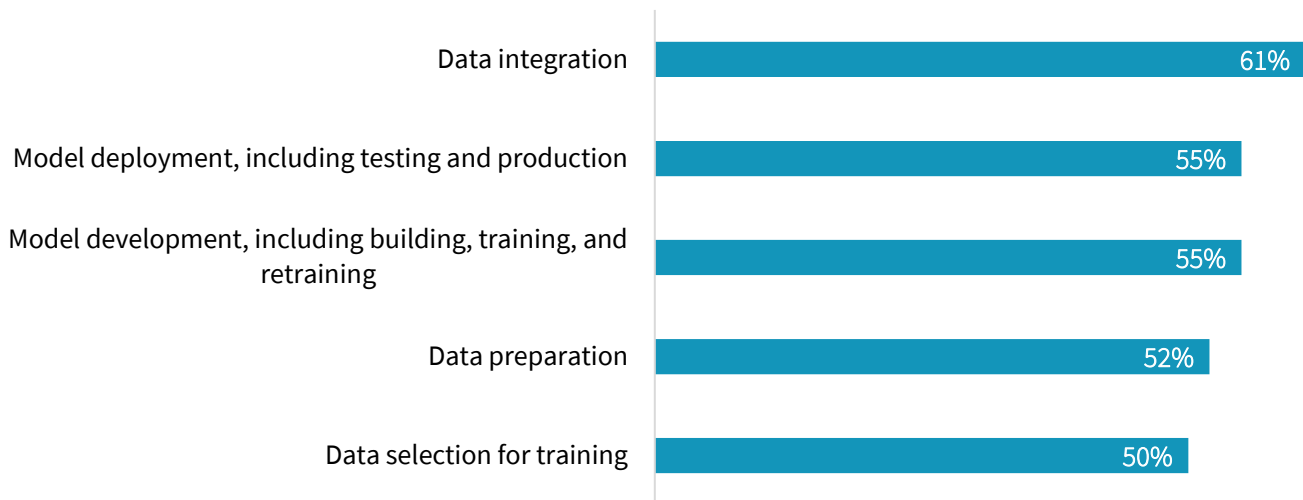
³ Source: ESG Master Survey Results, [Artificial Intelligence and Machine Learning: Gauging the Value of Infrastructure](#), March 2019. All ESG research references and charts in this white paper have been taken from this master survey results set unless otherwise noted.

organizations happen to have a data scientist on staff, they are being tasked with non-data-science tasks, such as data integration and data preparation. Situations like these can rapidly create a snowball effect where all aspects of the data pipeline are impacted—from data integration and preparation, through data science tasks such as model selection, training, and tuning, to eventual deployment into production (see Figure 1).

Due to the distributed nature of data, even with the right personnel in place, operational silos between roles create constant challenges. A great example is how data scientists aligned to a specific line of business understand the business use case and data science approach necessary to solve it, while IT will likely lack the data science or business acumen necessary to properly align infrastructure to efficiently and effectively support requirements. If the data scientists need access to new data, integrating new and different data sets may lead to different IT requirements. This is a primary reason for data integration being the most often cited challenge associated with the data pipeline.

Figure 1. Phases of the AI/ML Data Pipeline at which Personnel Are Performing Tasks Outside of Core Expertise

At what phases of the AI/ML data pipeline are individuals or teams being asked to complete or execute tasks that fall outside of their core expertise? (Percent of respondents, N=190, multiple responses accepted)




Source: Enterprise Strategy Group

Looking at the research, it is apparent organizations need assistance in their pursuit of becoming AI-driven companies. While hiring staff is a priority, tight budgets and lack of available talent in the job market make it difficult. ESG research shows that 38% of organizations are looking to work with technology vendors possessing expertise and strategic partnerships across the AI data pipeline, vendors that can help address the skills gap, ultimately offering a viable means to improve time to value and total cost of ownership.

Data Pipeline and Technology Stack Complexity

Further exacerbating the skills gap issue are the inefficiencies across the data pipeline, which impact numerous personas within an organization—individuals in IT and data-centric roles, developers, and business users. One of those inefficiencies is simply due to the number of technologies from different vendors across the pipeline, all of which must be properly integrated, optimized, and maintained.

ESG research found that, on average, organizations work with 37 different vendors across their AI data pipeline, which include a combination of both hardware and software vendors. This points to a growing need for vendor consolidation to simplify management, improve efficiency, and more easily embrace and integrate next-generation technology.

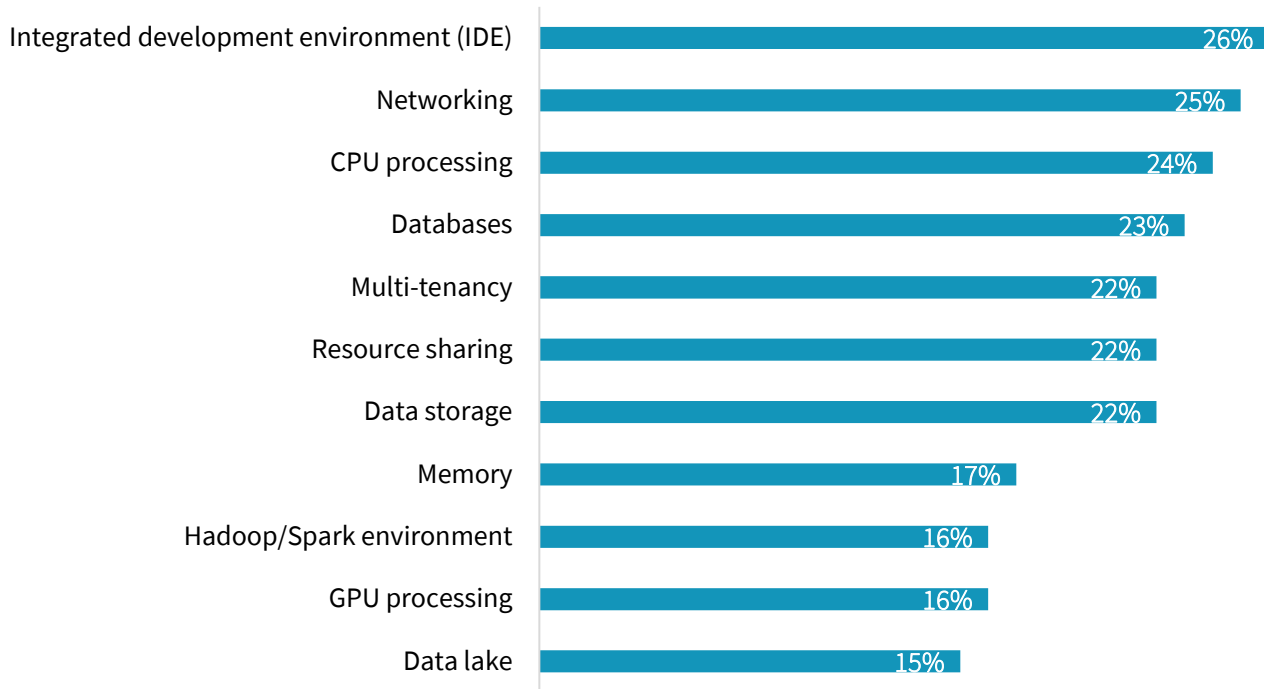


On average, organizations work with **37 different vendors** across their AI data pipeline, which include a combination of both hardware and software vendors.

When asked about the weakest links in their organization’s ability to deliver an effective AI/ML environment with their infrastructure stack, the largest percentage of respondents selected integrated development environment (developer persona), followed by networking, CPU processing (IT persona), databases (data-centric role), and multi-tenancy (business persona) (see Figure 2).

Figure 2. Weakest Links in the AI Infrastructure Stack

Which parts of the infrastructure stack do you believe are or will be the weakest links in your organization’s ability to deliver an effective AI/ML environment? (Percent of respondents, N=300, three responses accepted)



Source: Enterprise Strategy Group

Organizations want an infrastructure stack that provides the right levels of security and governance, reliability, cost savings, performance, manageability, and scale. They need help addressing the constantly evolving nature of the AI/ML infrastructure and software technology stack. They want guidance as to whether to utilize CPUs or GPUs for processing; Spark or Hadoop/MapReduce for big data processing; OpenShift or Kubernetes with Kubeflow as a container application platform; and TensorFlow or PyTorch for a deep learning framework. Organizations also want to be confident their partner can provide simplification across their AI stack, regardless of where they are on their AI journeys.

Enter Cisco.

Cisco’s Unified Architecture

A global leader and innovator, Cisco has been delivering business technology solutions and platforms for more than three decades, enabling organizations to achieve success by ensuring the efficiency, reliability, and security of an integrated IT infrastructure.

A Three-pronged Approach to AI Helps Organizations Operationalize AI at Scale

Because operationalization can be a major roadblock to AI initiatives, as well as an area where many AI projects fail, Cisco has taken a three-pronged approach with its Computing Solutions for Machine Learning. This proven methodology yields an effective means of helping organizations operationalize AI at scale. The three-pronged approach comprises:

1. Addressing the complexities of the distributed nature of data and operational silos across organizations;
2. Expanding and evolving ecosystems associated with AI from a skills and tools standpoint; and
3. Offering a unified architecture to address the diversity and growth of data.

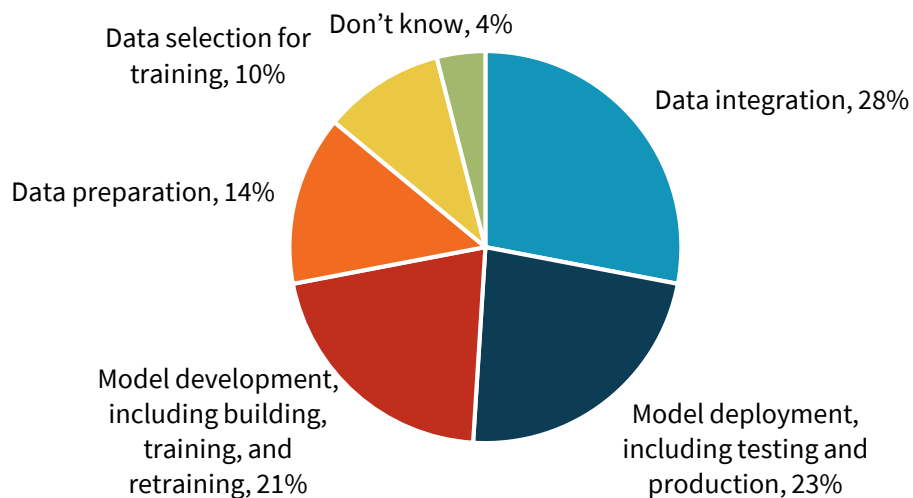
Organizations are able to leverage a full offering of products and services across an organization’s AI data pipeline—from integration and training, to test/dev and deployment/inference, including intelligent cloud-based management anchored by AI. To keep pace with the dynamic nature of data across an organization, Cisco provides the right infrastructure to accommodate all stages of the data pipeline and also works to validate independent software vendors (ISVs) on Cisco infrastructure to support customer’s data pipelines. Through proven partnerships across the AI ecosystem, organizations gain a validated solution from a proven industry leader, enabling an agile and efficient deployment process within their preferred environment.

Simplicity and Manageability Across the Data Pipeline

Today, there are growing numbers of organizations on the AI path. According to ESG research, 28% of those organizations cite data integration as their greatest challenge, while 23% cite model deployment into production, and 21% cite model development, including building, training, and tuning.

Figure 3. Most Challenging AI/ML Data Pipeline Phases

What phase of the AI/ML data pipeline do you feel creates the most challenges for your organization? (Percent of respondents, N=300)



Source: Enterprise Strategy Group

Clearly, travelling down the AI path with a trusted partner can make the process far less daunting and a lot more manageable. Recognizing challenges across the entire pipeline, Cisco has prioritized simplicity and manageability across the data lifecycle. Cisco's unified architecture supports AI initiatives with a holistic approach. Its full stack curation of solutions incorporates infrastructure and data analytics products and solutions and marries them together.

For data integration, organizations can confidently leverage pre-integrated technology, which helps aggregate data sources, and properly prepares data for training. On the training side, Cisco enables the model development, validation, and execution of data science tasks. Cisco can simplify the process of A/B testing and deployment into production for inference through containerization and unified infrastructure management.

Enabling AI through Full-stack Curation

Acknowledging that many organizations currently have data-driven initiatives underway, Cisco has partnered with a number of AI ecosystem players to assist in a seamless transition from big data and analytics to AI. Drawing upon its years of expertise in big data deployments, as well as its work with Hortonworks, SAP, Splunk, and others, Cisco is helping customers take advantage of new AI capabilities that big data vendors are adding to their offerings.

With ESG research showing that 83% of organizations are potentially restricted in their AI/ML investments due to previous investments in big data and analytics solutions, Cisco has simplified the AI/ML stack to enable rapid deployment of next-generation, AI-centric applications, bringing together the three main categories of an advanced analytics infrastructure: ingestion, compute, and storage:

- To maximize the value of data, organizations should look to leveraging tools such as Kafka, Jupyter, and Spark that enable access and analysis as soon as data is generated.
- For compute-intensive aspects, ML frameworks such as TensorFlow and PyTorch are available. Kubernetes and Kubeflow enable the rapid deployment of resources to support development, training, and tuning of models.
- For storage, Cisco partners including IBM, NetApp, Pure Storage, SwiftStack, and Scality enable organizations to tightly couple the right storage for the right business use case.

In short, depending on the type of data (structured/unstructured), volume of data, and ingestion rate of data, organizations have access to the most appropriate storage solution to fit their unique needs.

For example, if an AI initiative is centered around image analysis, using object storage would make sense. Cisco Validated Designs (CVD) allow organizations to hit the ground running. The Cloudera Data Science Workbench CVD provides a deep learning infrastructure with GPU-based nodes in a scalable Hadoop environment with Jupyter Notebook. As a leading Kubeflow contributor, Cisco is well positioned to create Kubeflow (TensorFlow running on Kubernetes), enabling rapid development and deployment for increased portability and scalability of the machine learning stack. Due to Cisco's active participation in the Kubeflow open source project along with Cisco Hyperflex, Cisco was presented with 2019's Google Cloud Partner Award for the second year in a row. Similarly, a Hortonworks-based data lake CVD integrates Hadoop with YARN scheduling across CPUs and GPUs and along with Docker, supports enables full data lifecycle for both the data scientists and IT teams.

Cisco Solutions and Services

With full-stack curation, Cisco can assist organizations on their AI journey every step of the way, including bridging the gap between IT and data scientists by delivering AI-enabling technologies specific to the role. With its holistic view, Cisco is able to address the challenges of diverse teams that require AI assistance.

Cisco offers infrastructure solutions right-sized for every phase of AI/ML projects across the organization—from test/dev, through model development, to training and inferencing. Cisco solutions enable data scientists to train and tune machine learning models in a right-sized computing platform, while supporting IT in deploying a distributed, next-generation application backed by AI.

For machine learning model training and tuning (as well as test/dev environments), Cisco offers UCS C240 servers for a wide range of storage and I/O-intensive infrastructure workloads—from big data and analytics to collaboration. For organizations requiring a more holistic infrastructure, Cisco’s hyperconverged offering, Cisco HyperFlex, offers GPU-integrated nodes that can ensure the right level of processing power, while supporting multiple use cases and workloads. When an organization is ready to deploy a model into production for inference, they will have the ability to leverage the smaller footprint Cisco HX220c server across edge locations.



UCS C240 M5



HX220c Edge M5

Deep Learning with Cisco

For those organizations already on the AI path and looking to use more advanced AI techniques in a core data center, Cisco’s UCS C480 ML can be leveraged to accelerate deep learning. Powering AI workloads at scale, the UCS C480 ML, by itself or in converged infrastructures, can assist organizations in attaining faster time to insights and enabling better business decision-making.

Because the UCS C480 ML is tightly integrated with NVIDIA NVLINK technology, organizations gain a full-stack solution, purpose-built to simplify the adoption of deep learning. The solution has been validated with the most popular AI frameworks, providing organizations with an optimal balance of performance and capacity to satisfy current requirements, while enabling future scale. Since the UCS C480 ML is part of the Cisco UCS family, organizations receive the same level of manageability and ease of use they receive from other Cisco UCS models. What’s more, Cisco offers support for the NVIDIA GPU Cloud-Ready (NGC) program, providing customers with pre-tested GPU-enabled software containers.



UCS C480 ML M5

FlashStack for AI/FlexPod for AI

Cisco has been working with its converged infrastructure partners, Pure Storage and NetApp, to deliver the UCS C480ML as part of a converged infrastructure stack. Cisco and Pure Storage have collaborated to produce FlashStack for AI workloads (combining the compute scale of Cisco’s C480ML with the storage scalability of Pure Storage FlashBlade). Cisco and NetApp have partnered to expand the FlexPod product line with FlexPod for AI, offering a reliable platform for AI and ML enterprise applications and next-generation compute, storage, and fabric.

Systems Management with Intersight Unified Management Platform

Cisco recognizes the distributed nature of growing organizations, and offers Cisco Intersight, a cloud-based, unified management platform to address the growing complexity of operational silos across locations. Organizations are able to centrally manage all deployed infrastructure from a single pane of glass, whether a HyperFlex cluster in a remote office, or a UCS C480 ML used for deep learning in a core data center. And since Cisco Intersight is a SaaS-based solution, organizations can benefit from Cisco's intelligent automation and proactive guidance. With its continued focus on innovation, Cisco makes extensibility a top priority.

The Bigger Truth

For more than three decades, Cisco has been enabling companies to leverage technology that enables the adoption of new business models, gain new efficiencies, and discover new opportunities. As organizations look to effectively embrace AI, they must turn to a vendor that offers the right level of guidance required to ensure success. With a unified approach across the complete lifecycle of AI data, Cisco can help demystify the AI stack for early adopters, as well as provide the next-generation hardware and integrations required to advance usage from machine learning to deep learning.

All trademark names are property of their respective companies. Information contained in this publication has been obtained by sources The Enterprise Strategy Group (ESG) considers to be reliable but is not warranted by ESG. This publication may contain opinions of ESG, which are subject to change from time to time. This publication is copyrighted by The Enterprise Strategy Group, Inc. Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of The Enterprise Strategy Group, Inc., is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact ESG Client Relations at 508.482.0188.



Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides actionable insight and intelligence to the global IT community.

© 2019 by The Enterprise Strategy Group, Inc. All Rights Reserved.