451 Research®

PATHFINDER REPORT

The Demands and Imperatives of a Comprehensive Inference System for Artificial Intelligence

COMMISSIONED BY

IBM

MARCH 2020

About this paper

A Pathfinder paper navigates decision-makers through the issues surrounding a specific technology or business case, explores the business value of adoption, and recommends the range of considerations and concrete next steps in the decision-making process.

ABOUT THE AUTHORS

NICK PATIENCE

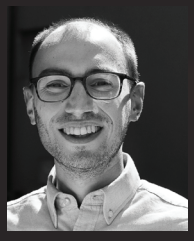
FOUNDER & RESEARCH VICE PRESIDENT, SOFTWARE

Nick Patience is 451 Research's lead analyst for AI and machine learning, an area he has been researching since 2001. He is part of the company's Data, AI & Analytics research channel but also works across the entire research team to uncover and understand use cases for machine learning. Nick is also a member of 451 Research's Center of Excellence for Quantum Technologies.

JOHN ABBOTT

FOUNDER & DISTINGUISHED ANALYST, 4SIGHT

John Abbott covers systems, storage and software infrastructure topics for 451 Research, and over a career that spans more than 25 years has pioneered specialist technology coverage in such areas as Unix, supercomputing, system architecture, software development and storage.



JEREMY KORN

ASSOCIATE ANALYST

Jeremy Korn is an Associate Analyst for the Data, AI & Analytics Channel at 451 Research, where he covers artificial intelligence and machine learning in the enterprise.

Executive Summary

With artificial intelligence (AI) becoming a more pervasive feature in the enterprise, there is plenty of discussion about how to implement the technology in a sustainable and successful way. One important consideration is the IT infrastructure necessary to implement an AI initiative at scale, particularly as inference workloads grow in magnitude and complexity. Although inference is the value-added step of the AI process, it can be especially difficult given the demands of real-time decision-making and the multitude of venues where inference can occur. It is, therefore, important for AI adopters to augment their infrastructure with comprehensive inference systems that deliver the speed, flexibility and software support integral to the next generation of inference workloads.

Benefits and Considerations of an AI Initiative

AI will be one of the most important technological advances in the history of the enterprise, and its proliferation has the means to impact, if not outright transform, a variety of enterprise processes. The core of AI is machine learning (ML), a technique by which computer systems are exposed to historical data in order to learn how to perform a task. Although ML has been refined by academics for decades, the recent convergence of large, digitized datasets, algorithmic advances including deep neural networks, and high-performance compute have pushed the technology out of the lab and into the enterprise. AI is front of mind for business leaders because of the diverse benefits the technology can generate by automating and optimizing a variety of enterprise processes. Figure 1 illustrates some of these benefits.

Figure 1: Reported benefits of AI

Source: 451 Research's Voice of the Enterprise: AI & Machine Learning, Use Cases 2020
Q: What are the most significant benefits your organization has realized or expects to realize from its use of machine learning? Please select up to 3.
Base: All respondents (n=991)

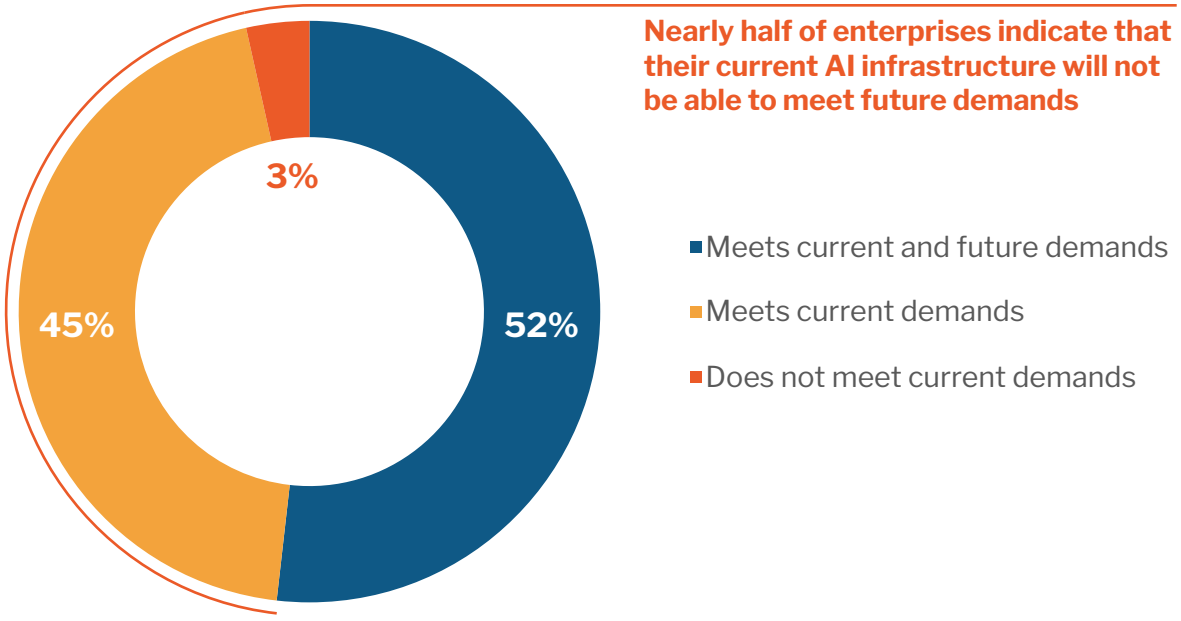
A plurality of AI adopters – 34% – cited reduced costs as a benefit of implementing the technology, but this is just the tip of the iceberg. A range of other benefits, including improved workforce productivity, increased sales and faster business agility, are all rewards of AI implementation. Depending on the application, the technology can even improve enterprise security posture or adherence to compliance protocols.

The breadth of benefits does not mean that AI adoption is easy. Successful implementation necessitates bringing together a multitude of stakeholders, skilled talent and innovative thinking. It also requires enterprises to rethink and retool their infrastructure. Organizations rushing to adopt AI often overlook a key point: an AI initiative is also an infrastructure initiative.

An IT environment for AI must be fast, robust and flexible in order to handle the volume of compute, the variety of workload venues and the critical nature of AI applications. As Figure 2 indicates, 48% of AI adopters say their current infrastructure for AI cannot scale to match future demands, and the 52% who believe in the adequacy of their current systems are probably not fully accounting for the growth and impact AI workloads will have.

Figure 2: Status of AI infrastructure

Source: 451 Research's Voice of the Enterprise: AI and Machine Learning, Infrastructure 2019
Q: How capable is your organization's current IT environment of meeting expectations for AI/ML workloads?
Base: All respondents (n=492)



Therefore – regardless of the stage of AI adoption – it is important for organizations to both reflect on the infrastructure they will need to support this technology and develop a strategy for implementation. According to 451 Research data, only 27% of AI adopters have done both of these steps, leaving a lot of room for improvement. Considering the infrastructure to support AI initiatives now will save enterprises a lot of time and pain in the future.

Three Stages of AI: Data, Training, Inferencing

In order to better appreciate the demands AI will place on enterprise infrastructure, it is important to build a solid understanding of the AI and ML process. There are three stages to the AI process: data, training and inferencing.

- **Data:** In this step, data is prepared for the model training process. It is perhaps the most integral of the three stages to the success of an AI initiative – as the adage goes, junk in, junk out. There are many components of wrangling the dataset necessary to build a first-class AI system, from identifying the right assets and integrating them to normalizing the data or enhancing it via feature engineering.
- **Training:** The next step in the AI process is training a custom model. This works by optimizing the parameters of a generic algorithm to a curated dataset. As the algorithm is exposed to more data points, it adapts to the underlying data, mimicking the process of learning. This trained model is then verified against a testing dataset to assess its accuracy. In most cases, data scientists will train a series of models and select the one with the highest accuracy.
- **Inference:** Inference is where the value of AI is realized. A trained model is not much use unless it is making novel predictions on new data, which is exactly what happens at the inference stage. The output of each inference – whether it be a binary classification or a translation from one language to another – can be leveraged by business professionals to automate or improve a business process.

The paradigm of data, training and inference provides a simplistic, but useful, overview of the AI and ML process. An important caveat is the interrelated, non-linear nature of these steps. There is almost never a straightforward linear flow from data to training to inference. Instead, the process is iterative, circling back on itself as data scientists seek to continually improve the predictive power of the model.

The complex nature of the ML process outlined above means organizations must develop a comprehensive infrastructure strategy that considers the unique demands of data, training and inference. The failure to account for one of these steps or the failure to build an interoperable system could lead to a faulty AI infrastructure, one that jeopardizes the ultimate success of an AI initiative.

Growth of Inference

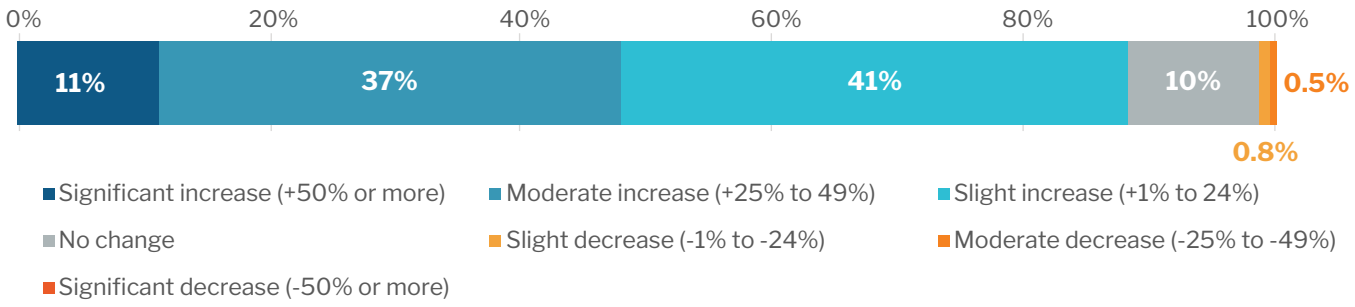
In many ways, it is still early days for enterprise AI. According to 451 Research data, 28% of organizations are still at the proof-of-concept stage, and an additional 14% have plans to adopt the technology in the next year. At this experimental phase, many organizations are focused on their most promising ML use case, iterating through the data and training stages in preparation for deployment and inference. However, many organizations are already at the inference stage. Twenty-nine percent organizations have deployed the technology in production, and even at this early stage of adoption, inference already poses an infrastructure challenge – 27% of AI adopters said inference is the most demanding stage of the AI process on their infrastructure.

Inference is likely to become a more significant challenge as AI applications grow in magnitude and complexity, leading inference to constitute a larger share of AI workloads. The growth in deployed AI models will be dramatic. Inference is where AI initiatives realize their return on investment, so it behooves organizations to deploy more models to improve and automate more processes. According to 451 Research data, the typical enterprise adopter of AI has 1-20 AI applications in production and an average of 27 applications in the proof-of-concept stage. Assuming a healthy success rate, the number of deployed applications at early adopters could double in the near term. In addition, more organizations with AI only at the proof-of-concept stage will move to production, further augmenting the number of ML models doing inference.

A second component of the equation is model complexity. Larger models increase computational density at the inference stage, and one indicator that model size will increase is the expansion of data used at the training state, as demonstrated in Figure 3.

Figure 3: Anticipated changes in data volumes for training

Source: 451 Research's Voice of the Enterprise: AI and Machine Learning, Infrastructure 2019
Q: Looking ahead 12 months, do you think you'll see an increase, decrease, or no change in the amount of data used to train your AI models? Please indicate the percentage of change in data usage over the next year.
Base: All respondents (n=390)

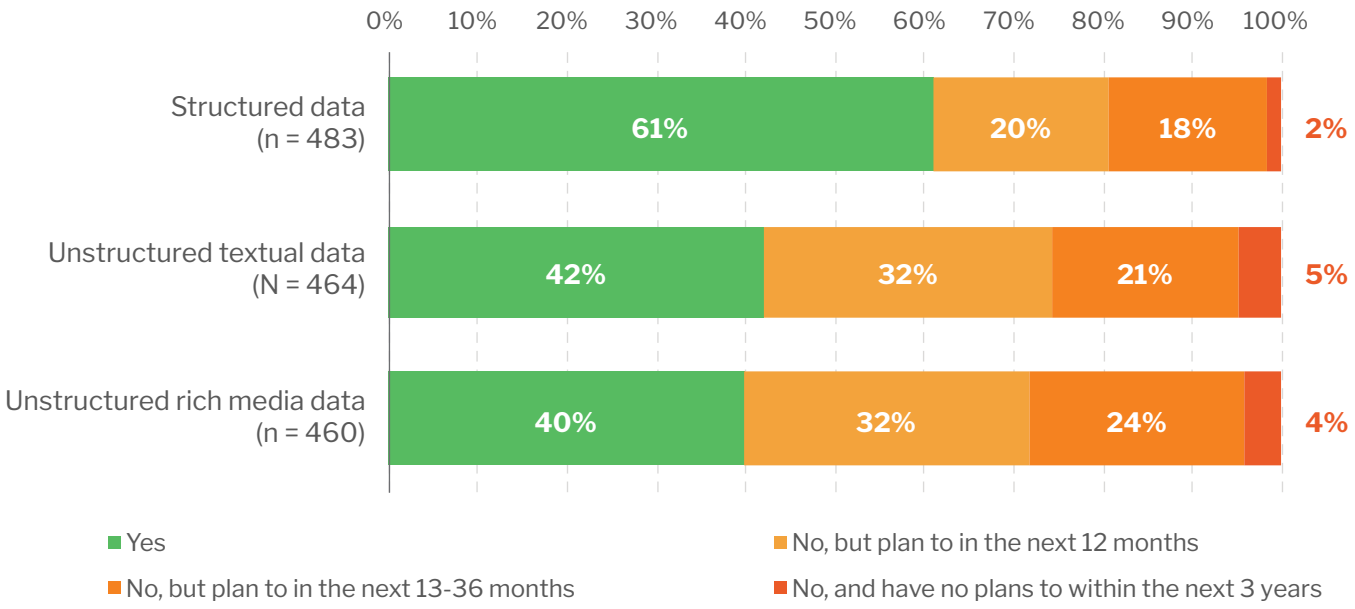


As indicated in Figure 3, a whopping 89% of AI adopters said they anticipate that the volume of data used to train their models will increase over the coming year; only 1.3% anticipate a decrease. This growth is expected to be significant, as 48% said this increase will be 25% or greater. The growth of data volumes for model training obviously will directly impact the infrastructure demands for model training but, as stated above, simultaneously indicates that models are growing in complexity, which will expand infrastructure requirements for inference.

Another indicator of the growing complexity of AI initiatives is the proliferation of data types feeding AI models, particularly unstructured data, which is illustrated in Figure 4.

Figure 4: Data types feeding AI workloads

Source: 451 Research's Voice of the Enterprise: AI and Machine Learning, Infrastructure 2019
Q: Do the following types of data feed your AI/ML workloads?
Base: All respondents



Structured data describes any information stored in a tabular format, as in a relational database. Unstructured data includes everything else: images, video, word documents, audio files. As Figure 4 indicates, structured data is primary today – 61% of AI adopters said they are using it for their AI models. On the other hand, unstructured textual data and unstructured rich media data lag, in use at 42% and 40% of AI adopters, respectively. This disparity will not last for long: integration of unstructured sources will begin to catch up to structured data in the next year, and within three years' time, AI adopters said they expect to leverage each of these data sources. Integrating more data types, especially of the unstructured variety, will make AI projects more complicated, putting increased demands on each stage of the AI process, especially inference.

In summary, the evidence suggests that models are growing in both number and complexity, driven by a desire to realize the benefits of AI technology while expanding it to new, diverse use cases. Infrastructure for data and training will be required for this AI experimentation, but inference will be required to turn these experiments into ROI. The growth of inference workloads should be top of mind for enterprises; otherwise, they run the risk of hampering their AI initiatives.

Challenges to Inferencing at Scale

The critical nature of AI systems necessitates that the underlying infrastructure be robust. However, it is not just the growing magnitude of inference workloads that present a challenge to scaling enterprise-grade AI systems. Two elements of inference – the need for real-time predictions and the diversity of inference venues – create additional challenges for assembling the optimal AI infrastructure.

Challenge 1: Real-Time Inference

One significant concern for organizations adopting AI is the imperative for real-time, low-latency inference. Model predictions are time-sensitive because they are calculated based on the status quo. If the underlying situation changes, the prediction may no longer add much value. Worse, it could lead to an action that increases risk overall. Therefore, it is critical to ensure predictions are served in a time window appropriate for the application. For example, speech models like BERT must typically complete inference within 10ms for a conversation to feel natural to a human.

AI adopters are struggling to do real-time inferencing, as evidenced by Figure 5. Only 19% said that new data for inference is processed in real time, whereas 81% do their inference as batch post-processing.

Figure 5: Status of real-time inference

Source: 451 Research's Voice of the Enterprise: AI and Machine Learning, Infrastructure 2019
Q: In your organization, which best describes how new data for making predictions from a trained AI/ML model (i.e., inferencing) is processed?
Base: All respondents (n=378)

This is not by choice. Of AI adopters not doing real-time inference, only 21% said that it is because their application does not require it. Instead, they reported infrastructure challenges: 40% blamed lackluster compute resources and 34% suggested that storage is not fast enough. In order to get to a place where deployed AI applications can deliver their full value through low-latency predictions, organizations are going to need to upgrade their infrastructure so it can handle the demands of real-time inference.

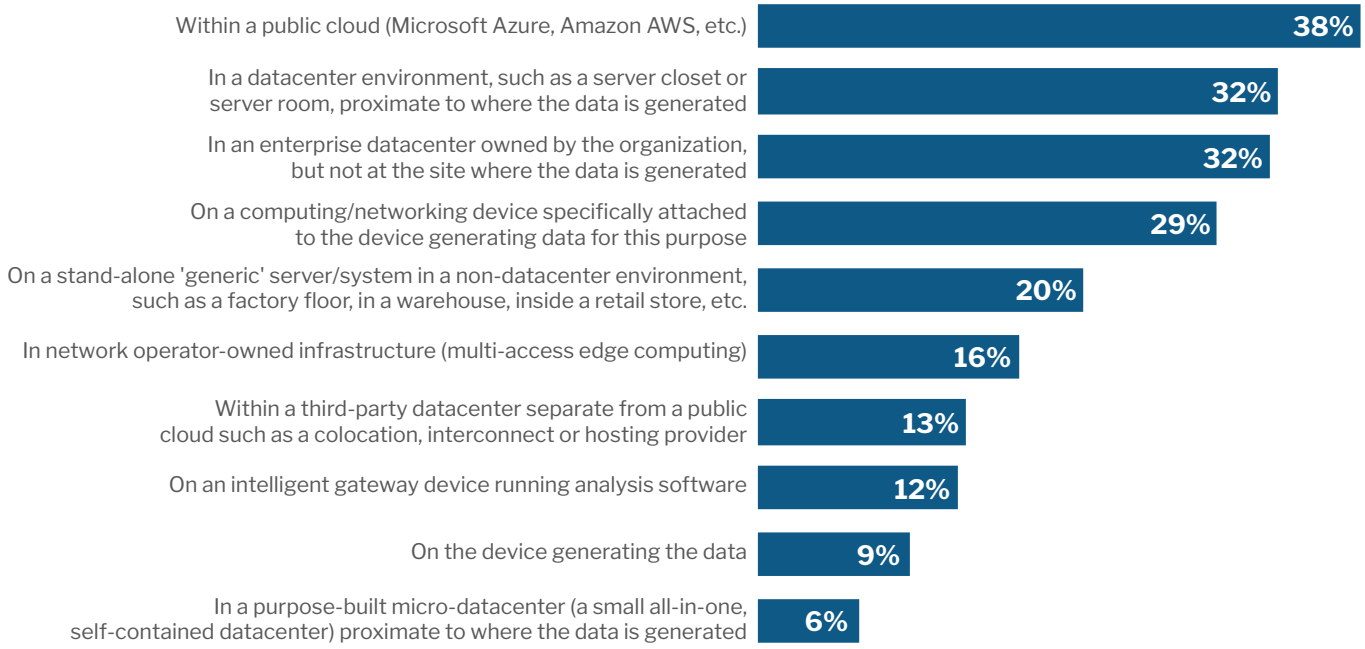
Challenge #2: Diverse Inference Venues

Another prominent concern for scaling inference is the need to support a diverse set of compute venues. In the paradigmatic ML process, while data is centralized for training, the model can be deployed to various nodes in the enterprise IT network, from core to edge, depending on the needs of a particular use case. Bringing a model closer to data inputs can help to improve latency, although it is not always feasible given the model size and complexity.

Figure 6 provides a window into the diversity of venues where AI adopters are currently performing inference. At this stage of enterprise AI adoption, datacenter environments are the most common localities for inference. Thirty-eight percent of adopters said they are performing inference in the public cloud, 32% are using traditional enterprise datacenters and 32% are doing inference in datacenter-like environments proximate to the data source. True 'edge' environments are less frequently used – only 20% are using stand-alone servers for inference, and 9% reported inference occurring on the data-generating device.

Figure 6: Venues for inference in the enterprise IT network

Source: 451 Research's Voice of the Enterprise: AI and Machine Learning, Infrastructure 2019
Q: Where do you make predictions from your trained ML models (i.e., inferencing)?
Base: All respondents (n=370)



It is not easy to predict how the workload distributions depicted in Figure 6 will change over time. There will likely be growth in inference at the edge as the number of low-latency AI applications increases, but it would be naïve to assume this growth will come at the expense of inference in the datacenter. The future of inference venues looks to be more diverse as companies try to balance the benefits and costs of various inference localities across the portfolio of their AI applications. It will not be an issue of *which* venues but instead a question of *how many*.

Inferencing System of the Future

To recap, this paper has established several key aspects of inference in an enterprise AI setting:

- Inference workloads will grow as AI applications expand in both quantity and complexity.
- Ensuring that inference can be done in real time is critical to the success of AI initiatives, and many organizations are struggling with this issue.
- Inference workloads will continue to be located at diverse localities in the enterprise IT network.

With these facts in mind, what will next-generation inference systems look like? What features and capabilities will they need to address the challenges and opportunities addressed above?

Speed

In order to produce value, predictions must be delivered within a certain window of time. Inference systems, therefore, need to deliver consistent predictions as quickly as any given AI use case necessitates. One aspect of speed is computational power. With AI models growing in complexity, a next-generation inference system will need to handle massive computational demands.

Computational accelerators such as graphical processing units and field programmable gate arrays are increasingly important components of AI infrastructure and could be one element of computational accelerators in the future. In addition to computation, memory and storage are integral to delivering in-time predictions. Inference systems must have enough memory to house complex models, and connections to storage must be robust enough to maintain throughput.

Flexibility

While AI can enhance almost any enterprise process, no two use cases are identical, and the flexibility of a next-generation inference system must reflect the variability of AI initiatives. An AI system can consist of multiple model types – support vector machines, k-nearest neighbors, random forest – and be built using many different frameworks such as Keras, TensorFlow, Torch and Scikit-learn. A next-generation inference system must be compatible with all the permutations of an AI system. Otherwise, developers will be handicapped and implementation slowed.

Flexibility is also critical to ensuring an inference system can operate at diverse venues within an enterprise IT network. As has already been explained, it might make sense to perform inference at various nodes depending on the latency requirement and cost. A next-generation inference system should not lock a vendor into a certain implementation. The needs of the AI use case, not the inference system, should dictate the implementation.

Software Support

As AI proliferates, the management of intelligent systems will become increasingly complex, and organizations will turn to software offerings to fill this gap. Software is necessary for load balancing and failover support, functionalities that guarantee the availability and robustness of inference systems.

It is also important that software integrate well with the chosen hardware stack. Aspects of model management, including monitoring and governance, require a tight integration between software and hardware, and a next-generation inference system will come with the requisite software and be compatible with open source model management tools. In the future, inference will be even more dynamic and just as important as it is today. To handle the demands of increasingly large and complex workloads, enterprise inference systems must be fast and flexible and come with software support.

Conclusions/Recommendations

As stated in the outset of this report, AI is set to be one of the most transformative technologies in history, with the potential to add value to numerous enterprise processes. AI adoption is not easy, and organizations looking to begin their AI journey or expand on their initial investments should make the following considerations.

1. **AI can provide significant value, and the time to start thinking about the technology and the necessary infrastructure is now.**

   Early adopters are already using the technology in numerous business processes and are accruing a variety of benefits such as lowered costs, improved productivity and increased sales. The time to make strategic plans around AI is now. Organizations that delay AI adoption risk ceding its varied benefits – and market share – to their competitors.

   To adopt AI strategically, of course, requires a defined use case. It also requires an investment in IT infrastructure. AI is a compute-heavy workload that will put significant demands on an enterprise IT environment, broadly defined. An AI strategy that does not include an infrastructure strategy risks failing at the start line.

2. **While data and training are often the focus of AI adopters, it is important to plan for the inferencing stage of the AI process.**

   Discussions of AI infrastructure often revolve around the data and training steps of the ML process. And, while ensuring enterprise IT is prepared for the demands of these stages, organizations shortchange inference at their own risk.

   Not only is inference where the value of AI systems is realized, but it also is a growing component of AI workloads, an increase driven by both more adoption and more complex models. Furthermore, inference has unique demands in terms of real-time processing and the diversity of compute venues. Understanding the nuances of inference is critical to developing a long-term, viable AI infrastructure strategy.

3. **A next-generation inference server needs to have speed, flexibility and software support.**

   Speed, flexibility and software support are critical components of a best-in-class inference system. Speed allows the system to perform computations at the scale necessary to deliver in-time predictions. Flexibility is necessary to handle the variety of models, frameworks and compute venues that compose an AI system. Software support ties the system together, allowing it to reach optimal performance while providing management and governance capabilities. Bringing all of these capabilities together is integral to successful inference, and organizations should prioritize inference systems with these components.

IBM offers the fastest, most efficient, most secure journey from data to insight. AI projects are growing across the enterprise, pushing the capabilities of existing solutions. As IT leaders face new challenges in an unfamiliar space and budgets shrink, enterprises want to make wise decisions with this significant investment. Enterprise AI on Power deploys at massive scale and reduces time and money spent delivering insights through unique load balancing and model optimization technologies delivered by cutting-edge IBM lab research.

IBM helps enterprises drive greater confidence in business decisions with more accurate model results and ensures AI is built for success with a trusted advisor to guide you from ideation to implementation. You can maximize infrastructure ROI and operational efficiency and lower TCO with dynamic, industry-tested and validated tools and co-optimized hardware and software. All of this is backed by IBM Power Systems servers, ranked highest in server reliability for the 11th year in a row (ITIC).[1]

Just like over 500 leading enterprises worldwide, with Enterprise AI on Power, you can deliver business insights at all stages of AI by leveraging the industry's highest data throughput and unique productivity-enhancing capabilities fueled by IBM research.

For more, visit *ibm.biz/EnterpriseAI*.

CONTENT PROVIDED BY:

IBM

1. ITIC 2019 Global Server Hardware, Server OS Reliability Survey *https://www.prlog.org/12761860-ibm-lenovo-hpe-and-huawei-top-itic-sever-and-server-os-reliability-poll.html*

About 451 Research

451 Research is a leading information technology research and advisory company focusing on technology innovation and market disruption. More than 100 analysts and consultants provide essential insight to more than 1,000 client organizations globally through a combination of syndicated research and data, advisory and go-to-market services, and live events. Founded in 2000, 451 Research is a part of S&P Global Market Intelligence.

NEW YORK
55 Water Street
New York, NY 10041
+1 212 505 3030

SAN FRANCISCO
One California Street,
31st Floor
San Francisco, CA 94111
+1 212 505 3030

LONDON
20 Canada Square
Canary Wharf
London E14 5LH, UK
+44 (0) 203 929 5700

BOSTON
75-101 Federal Street
Boston, MA 02110
+1 617 598 7200

451 Research®