ENTERPRISE MACHINE & DEEP LEARNING WITH INTELLIGENT STORAGE

SUMMARY

Fueled by data, infrastructure advances, and the ubiquity of machine learning and deep learning (ML/DL) toolkits, artificial Intelligence (AI) solutions are fast becoming a mainstay in the enterprise data center. AI turns data into insights across a broad swath of enterprise verticals as diverse as automotive, healthcare, life sciences, finances, technology, retail, and beyond. Data is now a competitive advantage in industries such as insurance − where predictive AI removes risks from underwriting, finance − where real-time deep-learning recognizes fraud as it happens, and even data center management − where patterns are analyzed to predict failures and scalability issues.

Artificial Intelligence and especially deep learning bring new demands to how data is served to the compute engines that consume it. The new realities of deploying artificial intelligence in the data center change the demands of density, throughput, concurrency and even scale-out data architecture change. IT must think differently about marrying storage and compute to deliver on the promise of AI for the enterprise.

This paper describes how deep learning and artificial intelligence in the enterprise bring new workflows and challenges to data center architecture. It also addresses how solutions can be constructed from infrastructure architectures specifically designed to bring scale-out compute and storage closer together.

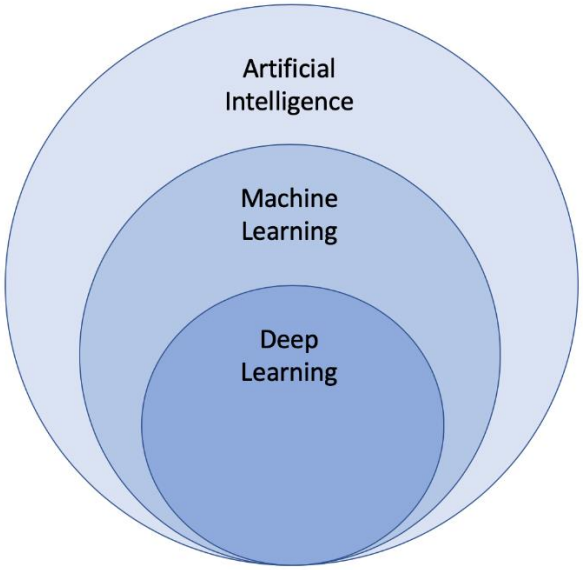
Deep learning requires large amounts of data to be fed into the processor without making the processors wait for that data. Properly marrying compute with the right storage technology, such as the Dell EMC Isilon series, allows data to be fed into the machine learning pipeline at the speed of the processor. Properly balanced systems accelerate innovation and deliver flexibility and agility to both IT organizations and the data scientists who rely on them.

DEEP LEARNING IS CHANGING THE ENTERPRISE

It seems by now that everyone has seen the famous magazine cover of The Economist, set against the backdrop of an oil rig, declaring that data is “The World’s Most Valuable

Resource."[1] Advances in machine learning (ML) and deep learning (DL) have indeed brought new importance to every organization's data – *data has become a competitive differentiator*.

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



Source: Moor Insights & Strategy

Artificial intelligence describes a general class of technologies where computers make decisions or provide insights that one typically associates with human intelligence. A simple example of AI is the retail recommendation engine, such as you might find at a retailer's on-line website that makes product recommendations based on both your past shopping history and what you're currently browsing for.

Machine learning is a type of artificial intelligence in which algorithms parse data, learn from that data, and apply that learning in real-world environments to make decisions. A machine learning system to detect email spam, for example, might be trained on samples from the millions of emails that are shunted to the spam folder by users every day.

Deep learning, one of the more interesting and active areas of artificial intelligence, is a sub-set of machine learning. Deep learning uses algorithms called *neural networks* to

[1] The Economist, The World's Most Valuable Resource is No Longer Oil, But Data. May 6, 2017.

continually refine its predictions from data as it's encountered. Deep learning is at the heart of autonomous vehicles, sentiment analysis that recognizes human moods, and most other AI techniques that deal with real world data to make dynamic decisions or recommendations. This paper deals most directly with DL, though architectural similarities carry across when implementing any type of data-intensive AI system.

AI use cases are many and diverse, from AI-driven chatbots and voice response systems, to customer behavior prediction, to optimized supply-chain management. The era of the intelligent enterprise has arrived and many are daunted by determining how to exploit machine learning technologies to best attain competitive advantage in their industry.

According to a recent survey[2] of over 1,300 IT professionals, more than 60 percent of respondents work for organizations planning to spend at least 5 percent of their IT budget on artificial intelligence. One-fifth of those respondents work for organizations that are planning to spend a staggering 20+ percent of their IT budget on AI.

Artificial Intelligence using deep learning techniques impacts every business, often in unexpected ways. Just a few examples of how machine learning, and its more specialized offshoot deep learning, impacts the modern enterprise follow:

- The **media and entertainment (M&E) industry** utilizes machine learning to bring intelligence to a number of tasks. Sentiment analysis is used to classify audience reaction to screenings of movies and television shows. The M&E industry also relies on image recognition trained by deep learning algorithms to provide automated metadata generation on countless amounts of video content.
- Modern **manufacturing** operations across industries rely on artificial intelligence and machine learning for many aspects of their operations. Image recognition systems analyze products on manufacturing lines to identify defects. Machine learning systems also assist in predictive failure analysis by analyzing sensors throughout a factory to recognize and identify patterns that could lead to failures if not addressed. Machine learning is also used to drive supply chain decisions that keep a just-in-time operation optimized with intelligent procurement and logistics.
- The **automotive and transportation** industries are using deep learning to change the way that the world thinks about cars. Deep learning techniques are fueling the race towards the ultimate goal of delivering autonomous vehicles.

[2] O'Reilly Media, The State of Machine Learning Adoption in the Enterprise, 2019.

> Along the way, we are seeing real world applications in the form of intelligent and adaptive cruise control systems, semi-automatic driving, predictive failure analysis, and even driver monitoring, to ensure a driver is aware of what's happening in the vehicle. None of these things would be possible without the on-going advances in machine learning and deep learning.

While deploying AI technologies into the enterprise is very impactful, it remains a new story for most organizations. It is important to simplify and look at the common building blocks before starting an AI project. Almost all DL solutions, whether supporting image detection, image classification, segmentation, natural language processing and/or predictive analytics, utilize a common set of core technologies. These techniques are deployed on platforms which natively support and are fine-tuned for common software packages, such as TensorFlow, PyTorch and Caffe2, which are ubiquitous in implementing these use cases.

Most traditional IT practitioners do not have the skills required to efficiently architect and deploy AI solutions for widely varying use cases. Machine learning and deep learning are technologies that arrive with new challenges and require innovative ways of thinking about data.

The same survey cited above shows that a lack of understanding of how to deploy deep learning, coupled with infrastructure that isn't ready for these workloads, as a significant inhibitor to adoption.

The most critical task for any competitive IT organization is to close that gap and build the skillset required to deploy deep learning, supported by flexible, future-proof analytics architectures.

ARCHITECTING FOR DEEP LEARNING IN THE DATA CENTER

Understanding the multi-dimensional impact of deep learning on storage architecture requires a high-level understanding of a typical learning workflow. Each stage in the learning pipeline places different demands on the underlying infrastructure. This is illustrated in Figure 2.

FIGURE 2: TYPICAL MACHINE LEARNING/DEEP LEARNING PIPELINE

	INGEST IOT, Logs, Sensors, Users, Etc	DATA PREP CPU-intensive Servers	REFINE GPU-enabled Server & Workstations	TRAIN High Performance GPU-based Servers	DEPLOY CPU or Inference Accelerated Edge, Client, or Server	RETENTION Long-term Storage
Access Pattern	Sequential	Sequential or Random	Random	Random	Random	Sequential
Access Type	Write	Read & Write	Read	Read	Read	Write
Concurrency	Variable	Low	Moderate	High	Low	Low
Performance	High	High	High	High	Moderate	Low
Storage	Block, File, or Object	Block or File	File or Object	File or Object	Block or Memory	Block, File, or Object
Scale	MB-GB	MB-TB	TB-PB	TB-PB	KB-MB	TB-PB

Source: Moor Insights & Strategy

These steps are summarized as follows:

- **Data Ingest** − Data arrives from an external source (or multiple sources), such as edge devices, log files, voice or video streams or customer relationship management systems. The data arrives and is stored. The storage solution only needs to be as performant as the incoming data requires.
- **Data preparation** − The data is cleansed and transformed for training. This critical step ensures that the data is consistent, outliers are identified, and the data set is optimized for the training algorithms. Some types of machine learning, such as supervised learning, require that data be labeled during this phase.
- **Data Discovery and Visualization** − Data scientists work with the data to optimize the training algorithms and parameters. This is a very iterative process, though one requiring only modest amounts of storage and compute.
- **Model Training and Development** − The bulk of the work occurs at this stage. The cleansed data is fed into a cluster of GPUs, or other high-performance compute engines, where it iterates for what is often very long periods. Training

requires high-throughput storage, optimized for high-concurrency random read operations.

- **Model Deployment** or **Production Inference** − The models generated by the training phase are deployed against data in the real world. The characteristics of this phase are highly dependent upon the type of deep learning being deployed. Image recognition, for example, may occur on a client device such as a smart camera, with little interaction with storage systems, while more advanced applications may run within an enterprise data center.
- **Data Retention** − Data used in training the model or used in repetitive inference for deep learning is retained for archival or re-use purposes. This is a critical step. Archiving the data ensures that models can be recreated and that the data can be mined for future insights.

These steps are backed by a set of design principles that must be considered when implementing an infrastructure to support deep learning:

- **Performance and scale:** Performance cannot degrade with scale. Each component, whether compute, storage or networking, should scale linearly and independently so the system can seamlessly grow with the workload in order to avoid compute, IO and networking bottlenecks.
- **Flexibility:** AI systems are built around data. The reality inherent in this dynamic is that software, analytic techniques and use cases will inevitably change as the AI ecosystem evolves, but an organization's data remains relatively constant. Systems should support long-lasting data storage while maintaining the flexibility to evolve with the changes in business needs.
- **Enterprise data management:** Data utilized by deep learning, despite its non-traditional use, is enterprise data and should be managed as such. Security, data protection, regulatory compliance, and other traditional data management concerns apply to deep learning data as well. Storage solutions deployed into these environments must integrate well with the existing policies and procedures for data management in the enterprise.

While discussions of machine learning and deep learning naturally gravitate towards compute, it's clear that these solutions force new ways of thinking about data. Deep learning requires thinking differently about how data is managed, analyzed and stored.

DATA IN A DEEP LEARNING ENVIRONMENT

The characteristics of data in a deep learning workflow are different from most other IT applications:

- **Data is mainly unstructured**, consisting of images, audio, free text or even streams of time-series data. Storage architecture for a deep learning environment must be optimized for unstructured data. Storage should also support multiple data access protocols such as SMB, NFS, HDFS, S3, and HTTP to deliver the utmost in operational flexibility.
- The **scale of data** is increasing dramatically with video and edge sensors, in particular, with higher resolution content generating many terabytes of data for analysis over concise periods. Retaining this data for later analysis or retraining can lead to petabytes of storage needs. Extracting reliable insights from DL requires a deep historical record of data to analyze. Storage solutions in this environment should have the ability to scale-out simply and non-disruptively.
- **Data usage varies significantly**, having different needs for each stage of the learning pipeline. Cleaning or labeling data, for example, has very different performance demands than the processes feeding that same data into a cluster for training or for real-time inference. One end of this pipeline can be satisfied by traditional local storage, DAS, or mid-tier storage. The other end of the pipeline requires throughput and enterprise features that can keep up with modern processing technology.
- **Data arrives from everywhere**. Deep learning applications have very diverse sources of data. Data for analysis or model generation may arrive from the edge, cloud-native applications, voice services, and even server log aggregation applications. Storage must be architected to ingest data from a variety of sources.
- **The lifecycle of data models.** AI requires a consistent set of management tools that span the gamut of high-performance to deep archival storage to keep data alive in a storage architecture aligned with the overall AI workflow of an enterprise. Similarly, turning existing data into the inputs for new AI capabilities requires data management tools that allow an IT organization to deploy new solutions against existing storage.

These high-level characteristics translate into real considerations when choosing a data management solution for deep learning. It's important to point out that this data is still

"enterprise data" and needs to be protected against hardware and software failures, secured against breaches, and managed efficiently.

The type of deep learning that an organization deploys also impacts the storage architecture supporting those workflows. Image recognition, for example, which is heavily used in industries such as media & entertainment, manufacturing and the automotive industry, is based around the application of convolutional neural networks (CNN) and deep neural networks (DNN).

CNN is a type of neural network that learns to classify and recognize images through a number of highly repetitive steps. The data access patterns for CNNs during both training and recognition require a storage architecture that is tuned for a very high number of small block read accesses to the underlying storage array.

Putting this into real-world perspective, in benchmarking performed by Dell EMC and NVIDIA, a Dell EMC Isilon F800 storage system was paired with NVIDIA DGX-1 servers comprised of multiple NVIDIA Tesla V100 GPUs. Each GPU executed more than five thousand parallel threads, which equates to an average of 703 concurrent file reads per GPU[3]. It's critical that the storage system paired with a deep learning system have the ability to serve data at scale and extreme concurrency without causing the processing elements to stall waiting for data.

That's just one example. Other deep learning systems have differing requirements. Intelligent systems that provide real-time pattern recognition for financial fraud detection, for example, may require very high-performance block storage. Applications with these constraints may be better served by high-throughput, low-latency block storage arrays, such as the Dell EMC PowerMax series.

Similar considerations with block sizes, file I/O patterns and scale exist. The critical takeaway is that serving up data for machine learning and deep learning is very different from any other enterprise workload. Managing data for deep learning requires deploying solutions that are built for high concurrency and multi-dimensional performance at scale with tiering across a single namespace and simple management through a consistent set of tools.

[3] Whitepaper: Dell EMC Isilon and NVIDIA DGX-1 Servers for Deep Learning, https://www.dellemc.com/en-us/collaterals/unauth/white-papers/products/storage/Dell_EMC_Isilon_and_NVIDIA_DGX_1_servers_for_deep_learning.pdf

DELL EMC: DELIVERING STORAGE FOR DEEP LEARNING

The power of AI can only be realized through efficient and performant delivery of data, leaving several factors to consider when designing storage solutions for machine learning and deep learning applications, where different phases of the learning pipeline have different requirements for performance, scale and concurrency.

At the same time, it makes sense to deploy storage architectures that seamlessly tier and scale to meet the requirements of all phases of a deep learning workload.

The Dell EMC Isilon family provides a solid base from which to deliver storage capabilities in supporting the full life-cycle of enterprise deep learning. This follows the workflow from training, learning, deployment and, ultimately, to long-term archival needs.

DELL EMC ISILON ONEFS

The power of any storage system is in its underlying operating system software. Dell EMC Isilon OneFS operating system provides the intelligence behind Dell EMC Isilon scale-out NAS storage solutions.

OneFS's powerful features and capabilities optimize and simplify data storage at the core of every artificial intelligence workflow. The software provides seamless tiering while providing a single namespace, managing data placement, optimizing and tuning the performance of each array based on detected traffic patterns, and providing for non-disruptive and linear storage scaling. The Dell EMC Isilon OneFS operating system delivers to each of these capabilities.

Simplicity of managing storage allows data scientists to focus their efforts on managing the machine learning process, without have to worry about the details of the underlying storage infrastructure. This simplicity also allows IT administrators to deploy the right mix of flexible and efficient storage solutions that span the needs of machine learning and deep learning.

- **Consolidated data lake** – Consolidates data across the analytics workflow in one place to simplify data analytics pipelines.
- **Multi-protocol support** − Enables analytics to come to the data to support a "store once, use many" methodology to improve agility.

- **Enterprise Data Governance** – Protects data with native resiliency and security features.
- **Seamless tiering** – Tiers storage between all-flash, hybrid, and archive nodes in the same cluster to allow for economic petabyte scaling and access to larger data sets.
- **Intelligent caching** − Provides the capability to dynamically tune the storage system's caching characteristics based on the workloads that are consuming data. The Isilon OneFS caching targets concurrent read performance, which is a crucial performance characteristic in deep learning workflows.
- **Linear Scalability** − Allows Isilon systems to maintain consistent performance while servicing the highly-concurrent parallel workloads characteristic of deep learning implementations.
- **DevOps and As-a-Service support out of the box** – Enables enterprises to carve out dev, test and production data environments or provide multiple production data environments with clear tenant separation through multiple access zones within same Isilon cluster.

The software manages the overall experience and intelligence inherent in the Dell EMC Isilon series. The combination of simple manageability with the array's solid performance and scalability characteristics makes Isilon an attractive platform for deep learning workloads.

DELL EMC ISILON: A PLATFORM DESIGNED FOR MACHINE LEARNING AND DEEP LEARNING

The top tier of the Dell EMC Isilon storage family is the Isilon F800 All-Flash Scale-out NAS. According to Dell[4], the F800 delivers performance and capacity that sits near the top of the industry. The F800 can perform up to 250,000 IOPS with 15GB/second aggregate throughput in a single 4U chassis and up to 15.75M IOPS and 945GB/second in a full 252 node cluster.

Looking at capacity, the Isilon F800 starts at 10s of Terabytes of storage and can non-disruptively scale-out to 10s of Petabytes in a single namespace. Isilon delivers up to 85% storage efficiency plus offers deduplication and compression technology that can

[4] Dell EMC Isilon F800 Specifications: https://www.dellemc.com/en-us/collaterals/unauth/data-sheets/products/storage/h15963-ss-isilon-all-flash.pdf

reduce data storage capacity requirements up to a 3:1 ratio, increasing the effective capacity of the solution.

The Isilon F800 is capable of keeping deep learning compute nodes well fed. Equipped with 60 high-performance SSDs and eight 40Gbps ethernet connections, these machines are architected to deliver consistent performance across the high-levels of concurrency required by deep learning. Beyond simply providing consistent performance, the Isilon F800 can be tiered with both Isilon Hybrid and Isilon Archive nodes to deliver easy to manage petabyte scalability.

Nowhere is this performance more demonstrable than in the jointly-developed Dell EMC reference architectures marrying the capabilities of the Isilon F800 with the NVIDIA Tesla V100 GPU accelerated servers like the PowerEdge C4140, DSS 8440 and NVIDIA DGX-1. Benchmarks of these solutions showed the performance of the ResNet-50 benchmark with up to 72 GPUs achieving linear image per second performance from 8 to 72 GPUs with GPU utilization at 97%[5].

These benchmark numbers demonstrate that in one of the highest-performing deep learning computers available today, the processor is the bottleneck, while the Dell EMC Isilon F800 keeps it fully fed with the data.

DELL EMC POWERMAX: HIGH PERFORMANCE BLOCK STORAGE

There are some steps in the AI workflow and specific ML and DL algorithms that require very low-latency block storage for real-time response rates during data ingest, data prep and production inference.

The Dell EMC PowerMax series of block storage solutions, as one of the top performing storage architectures currently available, are well architected to support these scenarios. The PowerMax is built on end-to-end NVMe, delivering latencies under 300ms at between 1.7 and 10M IOPs (for the PowerMax 2000 and PowerMax 8000, respectively) and with up to 13TB per brick[6].

[5] Dell EMC Whitepaper. Dell EMC Isilon and NVIDIA DGX-1 servers for deep learning. November 2018. https://www.dellemc.com/en-us/collaterals/unauth/white-papers/products/storage/Dell_EMC_Isilon_and_NVIDIA_DGX_1_servers_for_deep_learning.pdf
[6] Dell EMC PowerMax Specification Sheet: https://www.emc.com/collateral/data-sheet/h16739-powermax-2000-8000-ss.pdf

Dell has positioned the PowerMax to support the most demanding real-time AI workloads that are being deployed in enterprises today.

DELL EMC: FULL STACK DEEP LEARNING

Storage and compute are intertwined in deep learning environments. A well-architected infrastructure for deep learning, with all of the associated complexities of managing data, comes down to balance, interoperability, performance and flexibility. Despite high levels of similarity across implementations, there is no one right way. Every deployment and each environment differs slightly.

There is a myriad of options in deploying machine learning and deep learning workloads. Different phases require not just different data access but also different compute solutions. AI practitioners can choose to run workloads on bare metal servers, in virtual machines, or even in Docker-like containers.

Beyond simply delivering individual elements into a deep learning infrastructure, Dell EMC works to enable solutions that can be quickly deployed by IT practitioners. Dell EMC simplifies architectural decisions and shortens deployment times with Ready Solutions and reference architectures (RA) that combine elements to solve the problem at hand. Dell EMC provides solution configuration guidelines that help enterprises size and scale their data analytics and AI solutions to align with their specific workload requirements.

The Ready Solutions and RA's blend the right-sized Dell PowerEdge servers with Dell EMC network switches, Isilon storage and a software stack optimized for the solution. The Ready Solutions are validated and orderable hardware and software stacks optimized to accelerate AI initiatives, shortening the time to architect a new solution by 6-12 months. Bringing additional power and benefit to the Dell EMC Ready Solutions for AI are Dell Technologies consulting, support, financing and deployment services. These services all work together to ensure a frictionless solution deployment.

Reference architectures are tested and validated stacks targeted at Dell's customers and solution partners. While Ready Solutions can be ordered directly from Dell, RAs are aimed at helping IT practitioners build their own best of breed solutions based on Dell Technologies' proven products.

TABLE 1: EXAMPLES OF SOME OF THE AVAILABLE READY SOLUTIONS AND REFERENCE ARCHITECTURES

Type	Solution	Key Elements	Key Partners
Ready Solutions for AI	Deep Learning with Intel	Isilon H600 PowerEdge R740xd PowerEdge C6420	Intel
	Deep Learning with NVIDIA	Isilon F800 PowerEdge R740xd PowerEdge C4140	NVIDIA
	Machine Learning with Hadoop	Isilon H500/H600 PowerEdge R640	Hortonworks
Reference Architectures for AI	Dell EMC Isilon and NVIDIA DGX-1 for Deep Learning	Isilon F800 NVIDIA DGX-1	NVIDIA
	Dell EMC Isilon and PowerEdge C4140 for Deep Learning	Isilon F800 PowerEdge C4140	NVIDIA
	Dell EMC Isilon and DSS 8440 for Deep Learning	Isilon F800 DSS 8440	NVIDIA
	Dell EMC Isilon and PowerEdge R940 for Algorithmic Trading	Isilon F800 PowerEdge R940	Intel

Source: Moor Insights & Strategy

CONCLUSION

Data has become many organizations' most strategic and differentiating asset. AI techniques are revolutionizing the way that data is interpreted and utilized. Enterprises are heavily investing in building knowledge and deploying infrastructure to support this reality.

At the same time, artificial intelligence, whether machine learning or deep learning, requires IT organizations to think about data and storage architecture differently from those supporting more traditional enterprise workloads. The attributes of the data are different. The complexity of the analytics is different. The needs of the consumers of that data are different. The ability to keep accelerated compute nodes fed with data is paramount. The Dell EMC Isilon-based AI solutions are designed for precisely these needs.

Deploying deep learning solutions requires careful thought; it requires partnering with technology providers who understand the demands of this new world, delivering the breadth of targeted solutions required to ease the pain of IT practitioners living in that world.

Dell EMC is a great example of such a partner. Deep learning puts data first, and Dell EMC is one of the world's leaders in managing the data from data centers, private and public clouds, and edge networks. The breadth of the Dell EMC AI portfolio uniquely positions them to help design the best environment possible to meet customer needs. Dell EMC has an expansive storage portfolio managing and protecting customer data, along with and services and solutions optimized for success with AI.

Learn more at Dell EMC's dedicated website: https://www.dellemc.com/en-us/solutions/artificial-intelligence/index.htm

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR
Steve McDowell, 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 Dell. 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.