Deloitte.

Is your Data Infrastructure Ready for AI at the Point of Action?

Is your Data Infrastructure Ready for AI at the Point of Action?

Artificial intelligence (AI) is increasingly a competitive necessity for many businesses—but not all AI and compute capabilities are the same. To activate the most powerful and differentiating AI applications, many organizations are missing some key technology pieces in their data architecture puzzle.

There are a variety of approaches for adopting AI, and companies have looked to cloud-based AI and software-as-a-service to automate and transform business functions. The economics of this approach make sense, but when it comes to data-intensive initiatives and applications that require high performance computing (HPC) and high bandwidth or ultralow latency, changes to business data infrastructure may be needed to avoid excessive computing cost.

In many data applications, information is sent to the cloud where third-party-managed AI reveals knowledge that improves processes, forecasts, planning, and other business priorities that can function with some latency. Solutions for lower impact use cases like "Post Transactional Reporting" or "Pre Transaction" planning / forecasting have existed for quite some time now. In searching for topline and bottom line yields, the marketplace is looking for AI to make positive impact in the middle of transactional and operational flows ***"at the speed of business."***

AI optimized for HPC technologies can consume huge amounts of data and return insights in real time, allowing operations to be the direct consumer of data. This is only possible, however, if the computation happens at the point of action (i.e., the edge). This approach can unleash differentiating applications that lead to entirely new services and products or new business models altogether.

For many organizations, the existing data infrastructure may lack some components needed to enable real-time computation and insights at the edge. These include edge applications, edge AI infrastructure, GPUs and specialized compute capabilities, and accelerated algorithms.

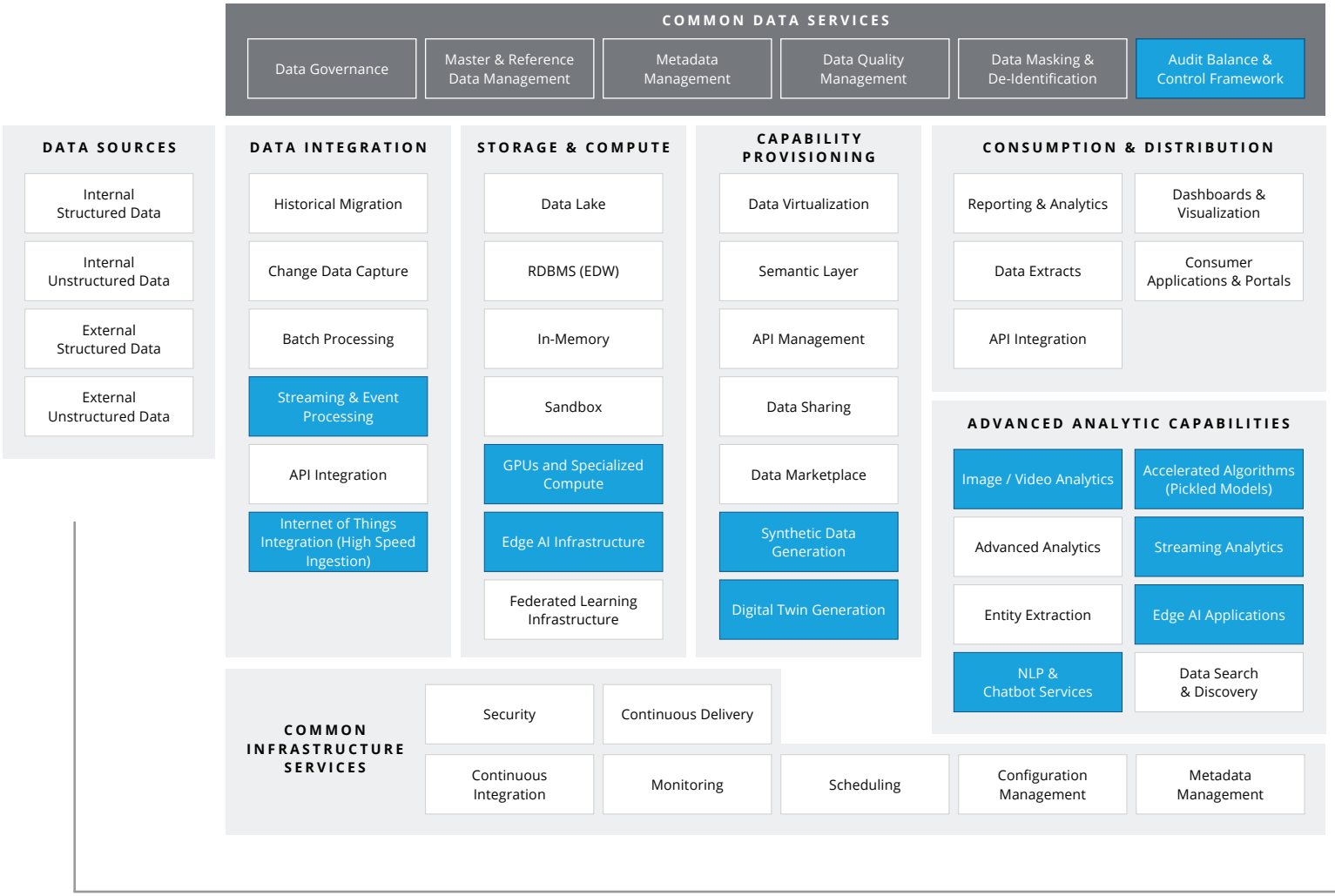
Next gen architecture from a business service view

Ultimately, every robot is owned by a business function because they are automating, replicating or enhancing what would otherwise be a human worker. When a human is taken out of the loop in this way, the business function still requires data and insights computed at the point of action. This is what it means to shift from post-transactional reporting to data attenuating business process real-time AI.

If we look at traditional data architectures, there are specific areas where a complementary suite of technologies can permit a model where operations are the primary data consumers.

Knowing what the **missing pieces** are is only part of the challenge.

The next question becomes, where do these technologies fit into the current data architecture and what does that mean for operations and business functions?



Edge Applications

Software and other smart applications may often be acquired via collaboration services with third parties, akin to how an organization may use data collaboration and composable infrastructure as a service. Edge applications typically are Edge AI powered applications which are deployed in devices throughout the physical world, performing with the equivalent of human cognition in real world. While edge applications may be found in the marketplace, the most novel and differentiating applications may be created in house.

Accelerated Algorithms

The algorithms fuel the applications, and in the data architecture, they can be grouped alongside data and AI business services, which include other automations, AI engines, and real-time data analysis. The algorithms may use a Pickle module, which allows you to save and reuse the AI model repeatedly, as opposed to retraining the model after each use, which would be prohibitively inefficient.

GPUs and Specialized Compute

The accelerated algorithms run on GPUs and HPC hardware that are optimized to run AI at the edge. GPUs enable massive parallel processing in timeframes that could never be replicated with general purpose CPUs. AI-optimized GPUs and the associated hardware allow the kind of real-time analysis needed to move away from post-transactional reporting.

Edge AI Infrastructure

Computing at the edge requires specific hardware, sensors, and services. Sensors include things like cameras, audio feeds, radar, and industrial asset monitoring. The infrastructure also includes edge storage, wired and wireless connectivity, and additional hardware for real-time compute.

With these component elements, the data architecture can begin to accommodate edge AI and unleash the potential in differentiating capabilities that require low latency, large data flows, and real-time compute.

Researchers in life sciences have succeeded in using GPUs to reduce the time needed for the analysis and sequencing of patients' genomes from days to a few hours. This provides the ability to identify and diagnose patients faster.

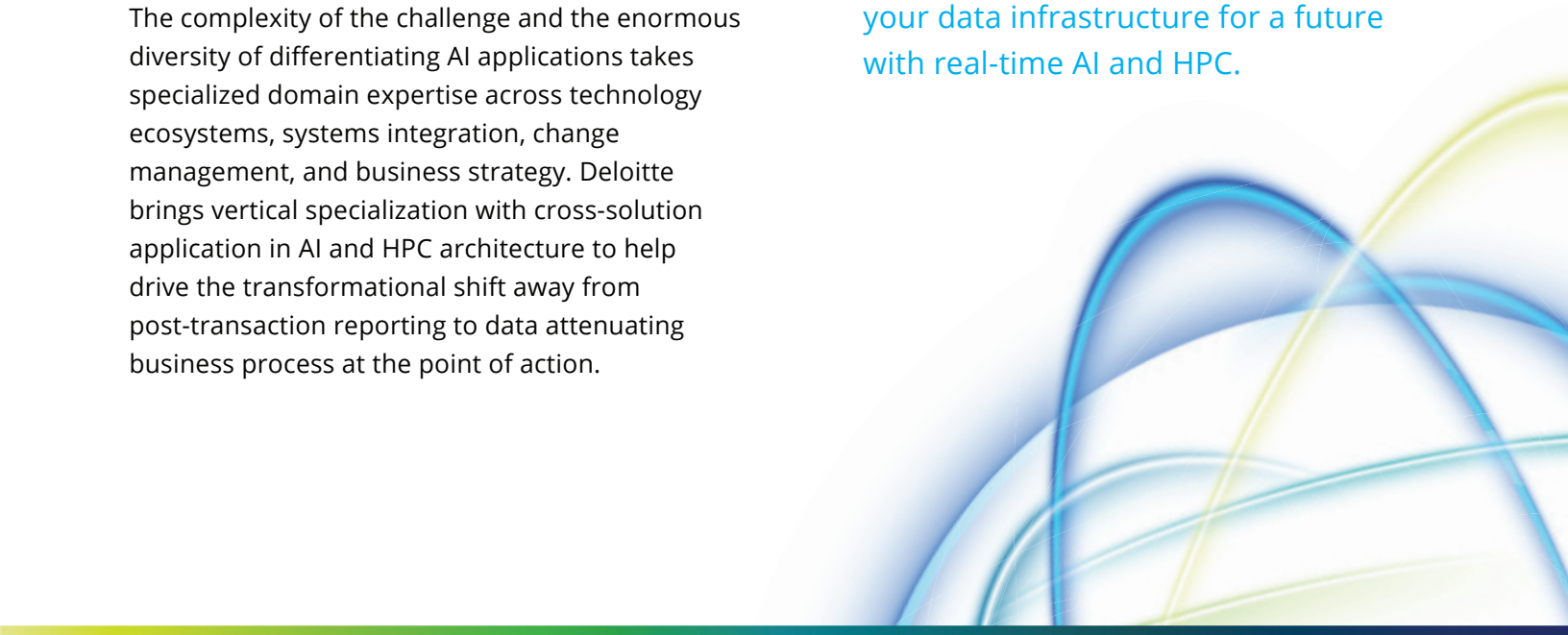
Within the automotive industry, the positioning of Edge devices in vehicles and developing the infrastructure to support the data consumption and processing enables vehicles to learn while driving and take real-time actions.

Setting up your data infrastructure for real-time compute

Once a vision is set for how the data architecture can be enhanced for HPC and edge AI, the next step is to identify which pieces can be bought, which are better used as a service, and which could be built by the enterprise. These are not just technology considerations but instead they impact the wider business strategy and spending. Indeed, shifting the data architecture is a business decision.

The complexity of the challenge and the enormous diversity of differentiating AI applications takes specialized domain expertise across technology ecosystems, systems integration, change management, and business strategy. Deloitte brings vertical specialization with cross-solution application in AI and HPC architecture to help drive the transformational shift away from post-transaction reporting to data attenuating business process at the point of action.

With our ecosystem of technology partners, we can help you **identify** the right hardware and infrastructure that aligns with business strategy and goals, and we then work with you to **implement** the right tools to **prepare** your data infrastructure for a future with real-time AI and HPC.



Get in touch

Christine Ahn
Principal
Deloitte Consulting LLP
chrisahn@deloitte.com

Brandon Cox
Principal
Deloitte Consulting LLP
brandoncox@deloitte.com

Goutham Belliappa
Managing Director
Deloitte Consulting LLP
gbelliappa@deloitte.com

Tanuj Agarwal
Snr. Manager
Deloitte Consulting LLP
tanuagarwal@deloitte.com