GlobalData.

Telstra

Recommendations for Building Agile Digital Infrastructure in the AI Era

Businesses are actively navigating through the uncertainties of new AI-fuelled innovative technologies. GlobalData research finds that 58% of businesses believe AI has already significantly disrupted their business and 48% says this disruption has already started in a tangible manner.[1]

Put another way, most industries and companies have entered an era of hyper connected digital interactions and AI-enabled workloads and workflows. Consequently, IT and business leaders must advance their data strategy. Focus is upon the creation of rigorous analytic capabilities and platforms for real-time decisioning and predictive insights; whatever the form, location, or source of the consumed data.

Seismic Trends of the AI Era

There are four major themes that are moulding digital business in the AI era. These shifts include:

1. **Remote, Hybrid and Distributed Workforces:** Research shows a shift in employees expressing a clear preference of working within flexible hours and locations. Technology is a major enabler for this shift, and so in turn, employers have also invested in new cloud and collaboration technologies to enable workplace transformation.[2]
2. **Multi-Cloud adoption:** GlobalData estimates over 80% of enterprises are adopting a multi-cloud infrastructure approach. Moreover, 60% of enterprise have further increased this mix slightly or significantly within the past year.[3] Cloud locations and availability zones are increasing to meet business resiliency and/or regulatory requirements.
3. **Vertical-specialised Infrastructure and Capabilities:** Technology is no longer confined to individual companies; it typically extends to collaborations across vertical ecosystems. Partnerships are essential to increased visibility, and traceability of data across supply chains and shared end-customer experiences. Emphasis is upon building APIs - often referred to as the solution glue - allowing disparate business systems to exchange information in real-time. There is also a rise in Industry Clouds, which comprise composable building blocks that are specific to the workload, security, data privacy and/or regulatory requirements of certain verticals.
4. **Human-to-Machine and Machine-to-Machine Communications:** There are over 1,000 scaled IoT projects tracked by GlobalData in the APAC region alone, and we anticipate machine-generated data and use cases will continue to proliferate. Natural language processing is facilitating communications between humans and machines, including language, syntax, context, semantic, content and translation. A 2023 National Bureau of Economic Research working paper predicts Generative AI could raise productivity by 14% in customer service roles.

[1] GlobalData. 25 February 2024. Thematic Intelligence: Tech Sentiment Polls Q1 2024 (n=355)
[2] Based on GlobalData's Thematic research on Future of Work. Within the past 6 months as of writing this report, GlobalData tracked 87.7K new hires, 1.5K filings and 700 deals related to the Future of Work theme.
[3] May 2023. GlobalData. ICT Customer Insights Survey – 2023 (n=2,428).
[4] Erik Brynjolfsson et al., "Generative AI at Work," NBER Working Paper No. 31161, 2023, April 2023, https://www.nber.org/system/files/working_papers/w31161/w31161.pdf

AI and Data Requirements are Fuelling a Cloud Reset

Data takes many forms and originates from multiple sources. As an overgeneralisation, data is only as valuable in relation to the direct or indirect insights it can drive to making informed business decisions.

In the context of AI, large language models (LLMs), or neural networks, are trained from huge data sets from a variety of sources to understand inputs from natural human language to deliver outputs such as text, audio, video, and images. While companies have a strong interest in deploying own or using third-party LLMs, each model will have a bare minimum of one billion unique parameters, plus a significant compute, memory, and storage footprint.[5] These massive data streams and sets are never static, but continuously evolve to improve speed and accuracy.

Transitioning to the AI-enabled era is driving IT leaders to re-calibrate – and in some respects – to even reset their overall cloud strategies. Recent GlobalData studies have highlighted an increase in the repatriation of workloads from public clouds to private clouds or metro-hosted premise solutions in 2024.[6]

Considerations of AI-enabled Workloads and Workflows

- **Data Gravity:** As companies digitise, the intensity of data use amplifies. Large data sets tend to attract additional data sets, applications, services, and business logic, akin to how mass attracts a gravitational pull proportionate to size. The proliferation of LLMs combined with the phenomenon of data gravity means models will continue to grow by volume, size, variety, and overall complexity. These models and data sets will be harder to move. Consequently, applications and services are more likely to be deployed closer to where the data resides. This is to reduce latency, improve performance, enhance security, or lower data mobility transfer costs.
- **Non-IT Factors:** A report by MIT Technology Review Insights identified non-IT factors such as regulatory, compliance, and data privacy environments as a leading barrier to rapid AI adoption.[7] The macro landscape is constantly evolving, so it is critical to embark upon IT strategies that can navigate changing non-IT requirements in an agile manner.
- **FinOps:** An industry FinOps study[8] finds that reducing waste and unused resources is the single highest priority in 2024. The study spanned major public cloud platforms, and all spend brackets.
- **Data Security and Privacy Protection:** Business must safeguard all corporate assets, shared ecosystem data, and end-customer personal identifiable information (PII) – whether that is data in motion, data at rest, or AI-enhanced synthetic data.

[5] A Large Language Model has over 100 billion parameters by GlobalData definitions and taxonomy. A Medium Language Models range from 5 to 100 billion parameters; a Small Language Model will have less than 5 billion parameters.
[6] Barclays CIO Survey program, 2024
[7] Generative AI: Differentiating disruptors from the disrupted, MIT Technology Review Insights, 2024
[8] State of FinOps Survey Reducing Waste and Managing Commitments Top Key Priorities for FinOps Practitioners, FinOps Foundation, 2024

Building Scalable and Agile Digital Infrastructure in the AI Era

The successful transition towards AI-enabled digital businesses will be underpinned by well-defined digital infrastructure strategies. Hyperscale modalities - while having ample computational resources - will not always be the most economically viable or best aligned to business mandates. A simple, yet compelling industry framework that GlobalData discovered highlight the importance of:

- **Right Fit:** The next five years may be more disruptive in data than the previous 10 or 15. Legacy environments will likely not be able to support the growth of data stemming from AI investments. Additionally, inflexible contracts, poor interoperability with different cloud providers and/or integration with premise-based solutions results in complexity and negative impact to businesses. It is important to identify the most suitable technology stack to run AI-enabled workloads. Consolidating estates or moving to new compute paradigms is advised. Organisations must also consider cybersecurity, regulatory and governance models when assessing and optimising their cloud deployments.
- **Right Size:** Digital Infrastructure is not a one-size fits all. It is important for businesses to match ICT resources to workloads and workflows in the most effective and economic way possible. IT leaders need to consider the scale of data, as well as the volume of computations that AI workloads will demand. GlobalData believes it is important for businesses to move away from proprietary billing systems to formal financial management frameworks that better align to modern corporate and ecosystem workload and workflow fiscal objectives.
- **Right Locate:** Many AI algorithms and data sets will have business or technical requirements for residing closer to their respective data sources. These could entail remote edge locations, near-site locales or within proximity data centres. Enterprises need to weigh up considerations such as latency sensitivity, data residency requirements, with compute, and storage costs to create the best business outcomes.

Beyond frameworks, GlobalData asserts that modern AI-requirements are driving enterprises to re-calibrate their overall infrastructure strategies. Businesses should consider the case for repatriating workloads to non-public cloud environments, where it makes sense. One, main consideration is cost and the need to minimise egress charges, especially as data sets grow. Another consideration is to maintain better control and oversight of datasets used for large language models. Thirdly, to ensure the optimal performance of workloads, especially for latency-sensitive use cases.

About Telstra International

Telstra International is the global arm of leading telecommunications and technology company Telstra. We empower enterprise, government, carrier, and OTT customers around the world with innovative technology solutions, including data and IP networks and network application services.

For more information, please visit: www.telstrainternational.com/digitalinfra

About GlobalData

GlobalData employs 3,500 developers, data scientists, analysts, award-winning journalists, editors and researchers, working in 23 offices worldwide and serving 4,500 clients in over 160 countries.

For more information, please visit: www.globaldata.com