KPMG

Future Proofing your AI Strategy with Scalable AI Infrastructure

Introduction

In today's digital landscape, the adoption of Artificial Intelligence (AI) is rapidly transforming industries, driving innovation, and unlocking new opportunities for organisations. From predictive analytics to personalised customer experiences, AI is revolutionising business operations. A research study by Tata Consultancy Services, one of the largest IT Service providers, reveals that more than **8/10 C-suite leaders have already deployed AI. According to Fortune Business Insights 2024 survey**, AI's market size is to rise from $621 billion in 2024 to **$2,740 billion by 2032**. AI is now seen as one of the key strategies, organisations are eyeing to achieve their business objectives and ambitions.

The rapid evolution of AI has led to a substantial increase in the **demand for computational resources** and marked the importance of the underlying layer that hosts such AI platforms and models. Compute limitations **(availability & cost efficiency)** remains one of the biggest challenges to overcome while deploying an AI model. The development of robust AI infrastructure is essential to support the computational, data management, scalability, deployment, and performance optimisation requirements of AI applications. As AI continues to advance, the role of AI infrastructure is pivotal in enabling organisations to harness the full potential of AI technologies.

AI Infrastructure

AI infrastructure refers to hardware and software environment designed to develop, deploy, and execute the AI workloads.

What makes AI infrastructure special is its high performance and scalable characteristics. Scalable AI Infrastructure is crucial for global businesses, as it ensures that their AI systems can handle **growing computational demands**. Many big players like Nvidia and Intel are investing heavily to build chips that can run complex AI models. Recently, Nvidia had launched a new Blackwell GPU that can enable organisations to build and run real-time generative AI on trillion-parameter large language models at up to **25x less cost and energy consumption** than its predecessor.

Components of AI Infrastructure

Irrespective of the type of organisation or industry, the core AI infrastructure components remains the same. At a high level, it includes computation resources, data management & storage, data processing frameworks and Machine learning (ML) frameworks.

Computational Resources:

The computational demand while running a machine learning or an AI task is generally intensive. It requires powerful servers and clusters that can process large quantities of data quickly, essential for complex AI models and simulations. These require specialised hardware such as **central processing unit (CPU) and graphics processing unit (GPU)**. Due to the scalability requirements, cloud environment is better suited to host the AI models as they can be scaled up and down based on the compute needs.

Data Management & Storage

AI applications require a large amount of data to provide predictions and analysis. Bigger the dataset size, better the accuracy. This historic or real-time data are stored in a data management system which performs tasks like data ingestion, data processing and data analytics. This involves storage solutions like **cloud storages, data lake, data warehouse or distributed file systems.**

Data Processing Frameworks

Data processing frameworks are critical for handling and transforming data efficiently before it can be used for model training and inference. These frameworks help with data cleaning, aggregations, and preparation.

Machine Learning (ML) Frameworks

ML frameworks provide necessary tools, libraries, and interfaces to develop, train and deploy AI models.

What to consider while deploying AI infrastructure

Deploying an AI infrastructure is a complex process that requires careful consideration of various factors to ensure optimal performance, scalability, compatibility, and cost effectiveness. AI models and their datasets are meant to drastically grow with time, and the underlying AI infrastructure should be scalable enough to support these models. The infrastructure should also be modular and upgradable to cater to emerging use cases in the AI world.

Any AI model involves huge chunks of data that need to be ingested and processed. Model training is also a crucial piece in the overall AI deployment. The machine learning algorithms should be able to process the **enormous data sets swiftly**, leading to faster model training and inference. The core AI infrastructure should be built considering the existing technology stack of the organisations. This ensures **smoother integrations** with the vast amount of historical and real-time data, hence making the model training method more mature.

Every **AI use case is different** and may require specific hardware, software, data management and integration capabilities. For example, a complex machine learning model like linear regression would require less computational power compared to the deep learning models like convolutional neural networks, which need powerful GPUs or TPUs. Real time applications like autonomous driving, real time fraud detection requires ultra-low latency and high performing infrastructure. Any batch processing tasks can still tolerate higher latencies and may use less powerful infrastructure.

Technology & Infrastructure Assessment for AI deployment

Many organisations are not sure where to begin their AI adoption journey. According to **"the 2023 State of AI Infrastructure Survey", 54% of respondents** had highlighted that they are facing infrastructure related challenges while developing and deploying their AI models. Weak infrastructure and cloud security controls can impact the integrity of the AI operating environment. A critical question before deploying AI models is to ask if the existing IT infrastructure and data ecosystems can support the AI technologies.

To overcome the above challenges, organisations can follow the below 3-phased approach to **discover** the current technology estate and **assess** the open-source AI platforms/ frameworks to **recommend** the best-fit technology:

01 Current State Discovery

To ensure a comprehensive AI implementation, conduct a current state discovery focused on specific AI use cases and business capability. Then assess the existing technology and infrastructure to identify gaps, evaluating components such as compute resources, storage solutions, data processing frameworks, security measures, data flow, application architecture, and integrations. Finally, analyse the current programming languages and frameworks to understand integration requirements.

02 Assessment & Gap Analysis

To determine the technology suitability for specific AI use cases, research and evaluate both enterprise and open-source AI technology options like Microsoft OpenAI, AWS Bedrock, PyTorch, and TensorFlow. Then engage in discussions with these AI technology partners to assess factors such as scalability, performance, security, and integration compatibility. Additionally, evaluate large language models (LLMs), the type and volume of data available for training, and the application architecture.

03 Target State Recommendation

For target state recommendation, shortlist the best-fit AI technology or service stack based on the feasibility assessments. This includes recommending appropriate AI models and frameworks, suitable data storage solutions, necessary virtual machines, or containers (GPU, CPUs) and programming languages & LLM models. Finally, define a future roadmap for implementing the recommended AI technology, outlining deployment and integration strategies.

Fully Managed AI infrastructure from Cloud Service Providers

It may not be the right business case for organisations who don't have deep pockets or the required skillset to build the AI infrastructure, LLM models, deep learning frameworks or the machine learning libraries from scratch. Cloud service providers like AWS, Azure & Google Cloud offers fully managed AI services and AI trained models that may reduce organisation's initial investments.

To name a few, there is Amazon Bedrock, Amazon SageMaker, AWS Deep Learning AMIs from AWS; Azure Machine Learning, Azure Cognitive Services, Azure Databricks from Microsoft Azure and Google Vertex AI, BigQuery ML, TensorFlow on Google Cloud from Google Cloud.

Conclusion

Artificial Intelligence has been with us over the last few years and has shown a tremendous growth year on year. With technology getting more mature and AI becoming the crucial link for businesses to grow multi-fold, all the components of an AI model needs to be addressed by organisations.

They must engage in thorough planning, starting with a comprehensive discovery of their current state to understand existing capabilities and gaps. This groundwork enables informed decisions when selecting the appropriate AI technology stack, ensuring the infrastructure is robust, scalable, future proof, and aligned with the specific needs of their AI use cases.

How can KPMG help?

KPMG has a track record of hundreds of successful global projects delivering solutions and tech infrastructure to support AI solutions for clients across the globe, demonstrating our capability to manage and deliver complex AI projects successfully. Our AI implementations have consistently led to improved operational efficiencies and enhanced decision-making processes, significantly adding value to our clients' organisation.

We have a proven, multi-phase approach that integrates and sequences the various interdependent activities that are required to successfully drive the AI journey, using our established "accelerators" to take clients from strategy to results delivery.

1. Discovery

Tailoring the program to client's ambition, initial maturity level, and latest AI developments and trends.

2. Decisioning

Defining client's AI strategy and roadmap, and high-level design for AI process, organisation design, governance, risk, technology, and people capabilities.

3. Design

Developing detailed designs across all elements of the AI business and operating model, and the detailed implementation and resourcing plan.

4. Operationalisation

Implementing the AI models, conducting skill development and delivering AI use cases.

Client Benefits

- ✓ A pragmatic AI vision and strategy, incorporating alignment to their strategy, market trends, and the current state maturity.
- ✓ Fit-for-purpose AI strategy detailing objectives and priorities holistically.
- ✓ Quantifiable success measures and practical plans for delivery.
- ✓ Accelerated design of AI operating model, technology, governance and risk frameworks.
- ✓ People capability development for key teams and roles.
- ✓ Benefit realisation through the initial use cases, and ongoing access to Insights and delivery support.

Contacts

Dishant Nagpal

Assistant Manager – Cloud Presales, KGS

✉ dishantnagpal@kpmg.com

in https://www.linkedin.com/in/dishant-nagpal-07414346

Dishant is a seasoned professional with more than 11 years of experience in developing IT and multi-cloud transformation solutions. He is the part of KGS Presales team and works on proposing disruptive cloud advisory, cloud engineering and software engineering solutions. His passion for emerging technologies has helped him to derive innovative solutions for various global clients.

Some or all of the services described herein may not be permissible for KPMG audited entities and their affiliates or related entities.

kpmg.com/uk

The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavour to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should act on such information without appropriate professional advice after a thorough examination of the particular situation.

Document Classification: KPMG Public

Create: CRT143124I | August 2024