

# Cloud-Based Infrastructure for Artificial Intelligence and Machine Learning Workloads: A Scalable Framework for Saudi Enterprises

Mohammed Vaquar Ahmed Haseeb

0009-0000-0379-705X

Published October 2026

## Abstract

In Saudi companies artificial intelligence and machine learning are moving from being mere experiments to becoming operational capabilities, which is leading to an infrastructure issue that goes well beyond the need for computing power. The processes involved in training, fine-tuning, inference, feature preparation, model monitoring, governance and retraining all have diverse requirements for accelerators, storage, networking, data locality, elasticity and reliability. This review draws on research carried out between 2020 and 2025 in order to look at how cloud architecture can support these types of workloads on an enterprise level while taking into account the Saudi situation regarding rapid digital transformation, the adoption of cloud services, data governance and the modernisation of various sectors. The review applies a structured evidence-synthesis approach focused on infrastructure architecture, workload orchestration, MLOps, cloud-edge distribution, security, reliability, cost control and organisational readiness. The evidence shows that scalable AI infrastructure should be viewed as a coordinated platform rather than as a set of separate services; effective designs incorporate elastic pools of CPU and GPU resources, container orchestration, tiered data services, automated model pipelines, observability, policy controls, workload-aware scheduling and selective edge execution. The Saudi Enterprise AI Cloud Framework proposed in this review combines these capabilities into six layers and includes a closed operational loop that links demand prediction, placement, execution, telemetry, policy evaluation and lifecycle learning. The review maintains that scalability must be assessed together with availability, cost predictability, data sovereignty, operational resilience and discipline in managing the model lifecycle. It ends with an implementation roadmap and a research agenda for businesses that wish to industrialise AI as part of their digital transformation aligned with Vision 2030.

## 1 Introduction

Saudi companies are shifting from ordinary digitalization to systems in which data, predictive models and automated decisions are incorporated into day-to-day operations. Banks are using machine learning to detect risk and fraud signals; industrial companies apply predictive analytics to maintenance and process optimisation; healthcare organisations are becoming more reliant on cloud-based data services; and the modernisation of the public sector has led to greater demand for scalable digital platforms. This growth has altered the infrastructure issue. The main problem is no longer whether an enterprise can get a virtual machine on which a model can run. It is now whether the organisation can maintain a reliable environment in which many AI and machine-learning workloads can coexist, scale, recover, remain under control and deliver value without incurring uncontrolled costs.

Since AI workloads place unusual stress on the infrastructure due to their highly variable resource requirements, training a model may demand continuous access to GPUs and high-throughput storage, while online inference might require only short bursts of computing power with strict limits on latency. Feature engineering can be both data- and memory-intensive, and experimentation often involves the creation of a large number of temporary environments. Deep-learning clusters also show a wide variety of job sizes together with queueing effects and utilization issues which make the use of conventional static allocation inefficient [12]. As a result, research into cloud resource management focuses on prediction, workload-aware scheduling and dynamic allocation rather than on fixed provisioning [10,11]. These findings are directly applicable to businesses that wish to scale their AI without having to set up permanently overprovisioned infrastructure.

The cloud offers elasticity, access to managed services and geographic distribution, but simply adopting the cloud does not eliminate operational complexity. Since AI systems generate model artifacts, training datasets, feature stores, experiment metadata, deployment endpoints and monitoring signals which have to stay in sync, MLOps research has consequently redefined production machine learning as an integrated lifecycle that links development, deployment, monitoring and governance [2]. AI-as-a-Service also illustrates how cloud platforms can make reusable intelligence capabilities available while passing on the responsibility to service design, transparency, assurance and provider governance [3]. The conclusion is that enterprise AI infrastructure should combine with cloud engineering and ML engineering rather than regard them as separate fields.

Saudi Arabia is a particularly appropriate context in which to consider this kind of integration. Empirical research into the adoption of cloud technology within organizations in Saudi Arabia has shown that organizational readiness, security, governance, cost and institutional factors are key determinants of adoption [22-25]. Likewise, more recent studies concerning business performance in Saudi Arabia have pointed out that technological readiness and data quality are important prerequisites for realizing the value of AI [26]. Hence, the infrastructure architecture has to be both technically scalable and institutionally deployable. A framework that fails to take account of governance, skills, data controls or operating models might be technically advanced but will be difficult to maintain.

The review fills this gap by examining the way Saudi companies can organise their cloud-based infrastructure for use with AI and ML as a unified platform. It looks at six related questions: how computing power should be pooled and scheduled; how data should be stored and transferred; how the lifecycles of models should be automated; how cloud and edge resources should work together; how reliability, security and observability should be incorporated; and how enterprises should manage concerns about cost, sovereignty and organisational readiness. The outcome is a comprehensive framework which turns recent research into a practical enterprise architecture without supposing that a single cloud, a single model family or a single sector will be the main force in future adoption.

### **1.1 Aim and objectives**

This review aims at creating a cloud infrastructure framework that is based on evidence and can be scaled for use with AI and ML workloads in businesses in Saudi Arabia. Different from previous studies which have only looked at algorithmic performance or at cloud adoption on an isolated basis, this review regards the infrastructure as the operating system for the entire model lifecycle.

The aims are to: (1) bring together evidence concerning cloud, edge, scheduling and MLOps techniques that affect the scalability of AI workloads; (2) determine the infrastructure features required to ensure reliable training, inference and ongoing lifecycle operations; (3) examine the constraints specific to Saudi Arabia relating to readiness, data quality, security, governance and resilience; (4) suggest a multi-layered enterprise architecture that incorporates compute, data, orchestration, lifecycle automation and governance; and (5) establish specific, measurable priorities for implementation and identify research gaps for companies that are transitioning from pilot AI projects to large-scale production of AI applications.

## 2 Conceptual foundations

Cloud-based AI infrastructure may be viewed as a continuous range of computing, storage, networking, and lifecycle services that are put together in order to turn data into deployed and monitored models. The cloud is appealing since it decouples demand from the need to own fixed hardware. However, AI increases the significance of resource heterogeneity. CPUs are still useful for preprocessing, for orchestration and for general services, while GPUs and other accelerators take on a major role in many training tasks and in high-throughput inference tasks. Evidence from large-scale GPU data centres shows that job duration, the number of accelerators requested and user behaviour vary enough to make it worthwhile to use scheduling and energy-aware management based on historical data [12]. Research into resource planning also indicates that due to the rapidly changing nature of AI workloads, it is advantageous to use precomputed or predictive allocation strategies rather than purely reactive ones [13].

Another key element is the cloud-edge continuum. Research into edge intelligence separates the use of AI for optimising edge systems from the actual execution of AI at the edge [4,5]. By bringing certain inference and preprocessing tasks nearer to the sources of the data, latency and bandwidth consumption can be reduced, although central cloud resources are still beneficial for heavy training, for the aggregation across different sites and for long-term storage. Further studies on federated learning demonstrate that distributed learning can maintain data locality while coordinating improvements to the model across various nodes [6-9]. For Saudi enterprises which have operational sites, branches, plants, hospitals or logistics facilities located in different geographical areas, this results in an architecture in which the cloud and the edge act as complementary rather than competing destinations.

The third pillar is lifecycle engineering. MLOps treats production machine learning as a workflow that can be repeated and which involves data preparation, experimentation, model registration, deployment, monitoring and retraining [2]. It is necessary to move from DevOps to MLOps since the behaviour of a model depends not just on changes to the code but also on changes to the data, feature drift and statistical degradation. From the point of view of the platform, this implies that the infrastructure must offer artifact lineage, reproducible environments, automated validation and model-specific observability. A cloud platform that merely provides computing capacity fails to address these lifecycle risks.

The fourth foundation is intelligent resource management. Research into machine-learning-assisted cloud resource management includes work on workload prediction and on container placement and scheduling [10]. By using integrated deep-learning methods it is possible to predict workloads and resource usage, which allows for proactive scaling rather than merely reacting to demand after the fact [11]. With regard to inference, serverless research shows that adaptive batching and automatic

model partitioning can reduce cost and latency when traffic is intermittent or the models are too large to fit into a single function [14-16]. Likewise, research on Kubernetes scheduling highlights the importance of gang scheduling, resource awareness and workload-specific policies for complex distributed applications [18]. All of these studies indicate that scalability results just as much from orchestration decisions as it does from raw capacity.

### 3 Review methodology

A structured integrative review approach was adopted since the research question involves cloud computing, machine learning systems, resource scheduling, MLOps, edge intelligence and Saudi technology adoption. The aim was to carry out an analytical synthesis rather than a statistical meta-analysis. The evidence window was restricted to the period 2020–2025 in order that the review should reflect current cloud-native, accelerator-based and MLOps practices. Two of the attached papers were used as structural and thematic references: the first offers evidence regarding AI-enabled load balancing in a cloud environment associated with Saudi Arabia, and the second shows an example of a structured review method in the area of Saudi digital technology research. Only literature cited within the 2020–2025 period specified in the request was included in the manuscript.

The search concepts were divided into four categories: cloud infrastructure and resource management; AI/ML training, inference and MLOps; edge-cloud and distributed intelligence; and Saudi cloud and AI adoption. In order to be eligible, studies had to provide empirical evidence, a review synthesis, a validated architecture or a substantial conceptual framework that was relevant to enterprise AI infrastructure. Those studies which concentrated only on algorithm accuracy and not on the infrastructure aspect were excluded. Preference was given to sources that included reports on operational metrics, architectural components, workload behaviour, adoption determinants or governance requirements.

Thematic coding was employed instead of counting the votes. The evidence was sorted into the following categories: compute elasticity, accelerator scheduling, data architecture, orchestration, model lifecycle automation, observability, edge distribution, security, governance, cost, and organizational readiness. The findings were subsequently compared with those in the relevant technical and Saudi-specific literature. This method prevents the erroneous treatment of different experimental settings as if they were directly comparable while at the same time enabling the identification of recurring design principles.

**Table 1.** Evidence domains used in the review synthesis.

| Evidence domain           | Primary question                       | Typical evidence                              | Framework implication                |
|---------------------------|----------------------------------------|-----------------------------------------------|--------------------------------------|
| Cloud resource management | How should capacity follow demand?     | Prediction, scheduling, autoscaling           | Elastic heterogeneous resource pools |
| AI/ML lifecycle           | How are models industrialized?         | MLOps, registries, monitoring                 | Reproducible lifecycle automation    |
| Cloud-edge intelligence   | Where should workloads execute?        | Edge AI, federated learning                   | Selective distributed execution      |
| Saudi adoption            | What constrains enterprise deployment? | Readiness, governance, security, data quality | Policy-aware operating model         |

---

## 4 Compute architecture and elastic resource pools

A necessary first step in order to achieve scalable enterprise AI is to have a heterogeneous compute fabric. It should not be the case that training, batch inference, streaming inference and feature engineering are all required to use the same instance profile; rather, companies should set up logical resource pools which provide access to CPUs, memory-rich nodes, GPUs and specialized accelerators via standard scheduling interfaces. With this kind of design it is less likely that expensive accelerators will be reserved for tasks that can efficiently run on general-purpose resources.

AI workload traces illustrate the importance of having workload awareness. Since deep-learning data centres have jobs of varying sizes and different completion times, past behaviour can be used to enhance scheduling and energy efficiency [12]. In real-time situations, resource demand may change when the set of active AI tasks changes; adaptive resource planning can reduce energy consumption without compromising the timing requirements [13]. Therefore, the findings support the idea of a control plane that predicts demand and assigns jobs based on resource intensity, urgency, data location, and service objectives.

Container orchestration provides a practical way of dealing with such a control plane. While Kubernetes provides features including scheduling, health checking, management of replicas, service discovery and recovery, its default scheduling is not guaranteed to be optimized for distributed AI. The literature on Kubernetes scheduling points out continuing problems concerning resource heterogeneity, gang scheduling and domain-specific policies [18]. Gang scheduling is important in the case of model training since distributed jobs may need to use multiple accelerators at the same time; if only a subset is allocated then resources will be left idle without any useful work being done. With regard to inference, autoscaling must take into account request rate, batch size, model loading time and accelerator memory rather than relying solely on CPU utilization.

Enterprises must therefore divide up the three scaling mechanisms. For stateless inference services, horizontal scaling involves adding replicas. Vertical scaling consists of altering the resource allocations whenever there is a change in memory or accelerator requirements. Queue-aware scaling determines the number and type of training workers according to the backlog and the expected duration. Historical workload models can be used to support the third mechanism, while telemetry can avoid overreacting to temporary spikes. The architecture should also provide preemptible or cheaper capacity for fault-tolerant experiments, while setting aside reliable capacity for important services.

The main design principle is to maximize the amount of useful work done by models per unit of scarce resource, rather than simply focusing on server utilization. A GPU that seems to be highly utilized by an inefficient task may in fact provide less business value than a shared or partitioned accelerator that is serving multiple models. It is therefore necessary for infrastructure metrics to link utilization to throughput, queue time, model latency, training progress and service-level objectives.

## 5 Data, storage and network architecture

AI infrastructure is usually limited by the movement of data long before it is restricted by computing power. Training pipelines tend to read large datasets repeatedly, produce intermediate files and write out checkpoints. Inference services need fast access to features, while governance demands a durable lineage that links source data, features, model versions and predictions. For this reason a scalable design must use a tiered storage system rather than a single repository.

Object storage is suitable for use with raw and curated datasets, model artifacts and checkpoints since it provides durability and scalability. For active training datasets which need to feed the accelerators at a high throughput, high-performance local or distributed storage is useful. Transactional databases and feature stores are used for low-latency online features, while metadata repositories keep record of the experiment lineage and deployment state. The architecture must clearly specify the movement between these different levels so that data is not copied repeatedly without proper policy.

The network design should correspond with the workload topology since distributed training can cause heavy east-west traffic between the workers and data ingestion leads to north-south traffic from enterprise systems or edge sites. Research involving the cloud and the edge has shown that shifting some of the processing to the edge can reduce bandwidth usage and response time [4-9]. This is especially important in cases where Saudi enterprises have remote industrial sites or branches that are sensitive to latency. However, moving all AI processing to the edge may result in management fragmentation and poor utilization; the better approach is therefore selective edge execution combined with centralized lifecycle governance.

The AI study mentioned above is a helpful illustration of the need to take data and resource coordination into account. The proposed method for load balancing combines AI techniques with cloud-enabled data processing and looks at both throughput and energy usage, thus supporting the idea that the quality of service depends on how the computational demand is distributed among the available resources [1]. Even though this study focuses on a healthcare scenario within an IoE environment, the fundamental lesson applies generally in that data-intensive AI services require the simultaneous optimization of compute, network and energy rather than having them tuned separately.

Saudi companies should also classify their data according to how sensitive it is, the requirements relating to residency, the level of latency they can tolerate and its reuse value. Such classifications can be used to guide placement policies; sensitive datasets should stay in controlled areas or private environments, while derived features or anonymised outputs can make use of wider cloud services. Implementing placement policies that are aware of these factors reduces the conflict between scaling up AI and keeping governance in place.

## 6 MLOps and lifecycle automation

Infrastructure only ceases to be merely infrastructure and becomes an enterprise AI platform when models are able to move safely from the experimental stage to production. Research in the field of MLOps highlights the necessity of reproducible pipelines, the management of artifacts, automated deployment, and monitoring [2]. Likewise, the concept of AI-as-a-Service stresses that reusable intelligence services require well-defined service boundaries and governance [3]. These concepts lend support to a lifecycle architecture in which each model is regarded as a versioned operational asset.

The pipeline must start with data that can be reproduced and with the application of feature transformations. Training needs to keep record of the version of the code, the version of the dataset, the environment, the hyperparameters and the resulting metrics. The model registry then has to enable a controlled move from candidate to approved artifact. Deployment pipelines should test not only the behaviour of the software but also the model-specific conditions such as schema compatibility,

baseline performance and resource requirements. Thus a direct link is established between model governance and infrastructure automation.

Monitoring during production should involve two types of signals. The system signals are latency, throughput, memory usage, GPU utilization, error rates, and saturation, while the model signals consist of input drift, changes in feature distribution, prediction distribution, confidence behaviour, and the quality of outcomes once the labels become available. Retraining should not take place merely because a scheduled date has come around; instead, it should be triggered by actual degradation, the availability of new data, or policy decisions. This helps to avoid unnecessary computing costs and enhances auditability.

Role clarity is also necessary in the case of MLOps. It is the data scientists who optimise model behaviour, the platform engineers who look after the shared infrastructure, the application teams who integrate the model services, the security teams who set the controls, and the business owners who decide on the level of acceptable risk and the service objectives. If this separation does not take place, production machine learning tends to become dependent on individual practitioners. Those Saudi organisations which have been studying the adoption of AI have found that technological readiness and data quality affect the business performance that is actually achieved [26], thus underlining the need to regard skills, process and infrastructure as part of a single operating model.

A scalable framework must provide standardized templates for the various types of workload such as batch training, distributed training, batch inference, real-time inference and edge deployment. The use of standardised templates reduces configuration drift while at the same time permitting specialised models to ask for different resources. Moreover, it is possible to incorporate cost, security and reliability policies once at the platform level rather than having to reconstruct them for each project.

## **7 Reliability, resilience and observability**

Since AI services are increasingly taking part in decisions concerning revenue, safety, customer experience, and operations, platform reliability needs to go beyond simply keeping the servers running. An AI service that is reliable must keep on giving acceptable predictions, be able to recover from failures of its infrastructure and have to provide sufficient evidence in order to diagnose both software and model problems. Redundancy is therefore needed at many different levels.

Container orchestration is able to replace failed replicas and reschedule the workloads at the infrastructure level. At the service level, having multiple inference replicas together with traffic management helps to avoid single points of failure. At the data level, durable storage and versioned artifacts allow for recovery. At the model level, a controlled rollback enables the restoration of a previous model version if the new deployment performs poorly. Moreover, studies into serverless inference show that platform-level batching and partitioning can assist in meeting latency objectives when demand varies [14-16].

It is observability that links these different layers together. Logs provide an explanation of events, metrics give an overview of the system's behaviour and traces illustrate the path of a request as it moves between various services. In the case of AI, the traditional signals must be correlated with the version of the model, the source of the features and the deployment configuration. Otherwise, a spike in latency might be observed without knowing whether it is due to increased input volume, slower data retrieval, competition for accelerators or greater model complexity. The research in the field of

AIOps takes this concept further by using analytics to identify abnormal infrastructure conditions and to prioritise the necessary remedial actions [20,21].

Organizational resilience is just as important. Saudi research has associated cloud capabilities with enterprise resilience and has recorded concerns regarding security, governance and operational control [22,28]. It therefore indicates that resilience should be considered as a governance property as well as a technical one. Businesses must have incident ownership, recovery objectives, tested runbooks and lessons learned after an incident. A model endpoint which has an automatic restart but no responsible owner is not operationally resilient.

The framework in question therefore involves the use of closed-loop operations: telemetry is gathered, assessed in relation to the technical and model objectives, turned into scaling or remediation actions, and kept as learning evidence. Although human approval is still required in cases involving high-impact changes, routine recovery should be automated where policy permits.

## 8 Security, sovereignty and governance

The attack surface is increased by AI infrastructure since it involves code repositories, datasets, model artifacts, APIs, container images, credentials and automated pipelines. As a result, security measures should be based on the entire lifecycle and not merely concentrate on the production endpoint.

In studies carried out in Saudi Arabia on cloud adoption, security, privacy, governance and organizational readiness have repeatedly been identified as factors affecting adoption [22-25]. The significance of this evidence lies in the fact that enterprise AI often involves data which is sensitive from a regulatory, commercial or strategic point of view.

Federated learning provides a method for situations in which the raw data should stay distributed. Surveys carried out during the review period refer to federated learning as a means of coordinating model training while keeping the data at the participating nodes, although they also point out communication efficiency, heterogeneity and security as challenges that have not been resolved [6-9].

Governance should also take into account the various decisions involved in the model's lifecycle; each model in production should have an owner, a statement of its intended use, approved data sources, evidence of validation, a record of its deployments, and retirement criteria.

It is also necessary to take into account the concentration of cloud providers and the issue of portability. Although containerization and open interfaces can help to avoid unnecessary lock-in, complete portability is generally unrealistic since managed AI services are different.

## 9 Cost, sustainability and capacity economics

Cloud elasticity has the ability to lower capital expenditure, but AI can lead to fluctuating operating costs since accelerator instances, high-performance storage and data transfer are costly. Cost governance should therefore be incorporated into scheduling and architecture; the appropriate unit of measurement is not merely the monthly amount spent on the cloud but rather the cost per successful training run, the cost per thousand predictions, the cost per model version and the cost per business outcome.

There are several mechanisms offered by research into resource management. Predictive workload models are able to reduce idle capacity [11]. Historical job characterization helps to improve cluster

utilization [12]. Adaptive planning can be used to match resources with changing AI tasks [13]. Serverless designs are attractive for intermittent inference since the capacity reduces when it is not being used, although attention must be paid to model startup time and memory limitations [14-16]. Further research into Kubernetes scheduling shows that policy design has an impact on fragmentation and overall efficiency [18].

Firms ought to adopt a hierarchy of capacity with reserved or committed resources providing support for the steady baseline demand, elastic on-demand resources taking on the bursts, interruptible resources being used for retryable experimentation and edge resources looking after latency-sensitive local inference. This kind of approach is more robust than choosing a single purchasing model for all AI workloads. Chargeback or showback must be used in order that project-level consumption may be identified so that teams can link their design decisions to cost.

Sustainability must be considered together with cost; efficient scheduling is able to decrease the number of active nodes and enhance accelerator utilization, while right-sizing the model can stop unnecessarily large architectures from using premium hardware. Furthermore, research into distributed AI emphasizes the need to carry out computation at locations where data and resources can be made use of effectively [19]. For Saudi companies, since digital growth is taking place on a national scale, capacity efficiency contributes to both economic and environmental goals even though sustainability metrics are not currently available everywhere.

The principle is simple: elasticity is valuable only when it is governed. Unbounded autoscaling can turn technical resilience into financial instability. Cost policies should therefore set quotas, budgets, anomaly alerts<sup>10</sup>. The context of enterprise adoption in Saudi Arabia, leaving headroom for critical services.

## 10 Saudi enterprise adoption context

The situation in Saudi Arabia introduces a strategic aspect to the technical architecture. Research into cloud adoption in the healthcare, higher education and government sectors indicates that adoption is influenced by technical, organizational and environmental conditions, not just by the performance of the infrastructure [22-25]. When it comes to the adoption of community-clouds, the emphasis is on a combination of factors such as perceived usefulness, institutional influence, organizational capability and environmental fit [23]. Likewise, research carried out on healthcare organizations in Saudi Arabia has also identified a number of factors that affect decisions regarding cloud adoption [22]. It therefore follows that enterprise AI platforms should also have an adoption model and a reference architecture.

It is just as important to be ready for AI. Evidence from businesses in Saudi Arabia shows that AI and ML capabilities can lead to improved performance, but the extent of that improvement is influenced by technological readiness and the quality of the data [26]. In practice, companies cannot make up for poor data governance just by buying more GPU capacity; data ownership, data stewardship, the quality of integration, and the redesign of business processes must all move forward as the infrastructure is deployed.

Sectoral differences are also important: in the financial services sector, security, low latency and explainability may be given priority; industrial organizations might need edge inference close to their operational technology and reliable processing when connectivity is intermittent; in the healthcare sector, privacy and controlled data access may be stressed; and public-sector organizations might

have strong requirements in the area of sovereignty and be subject to procurement constraints. Because of this, the proposed framework is deliberately made up of modular components, with the same layers being kept but the policy, placement and service objectives differing from sector to sector.

Research into artificial intelligence and digital finance in Saudi Arabia shows that the country's programme of modernisation is leading to a greater number of applications for intelligent automation and data-driven services [27]. Meanwhile, studies of cloud resilience within Saudi companies show that cloud capabilities can help ensure organizational continuity [28]. Taken together, these findings point to a shift from a situation where AI infrastructure is developed on a project-by-project basis to one involving shared enterprise platforms that can be used by various business units.

The most effective way to carry out adoption is in a step-by-step manner. Businesses should start by setting up a governed cloud landing zone, using shared data services and having a small number of accelerator resources; then introduce standard MLOps templates; afterwards move on to multi-team scheduling and observability; and finally add edge, federated or multi-cloud approaches only if it is justified. By following this sequence, operating maturity is achieved before architectural complexity becomes too great.

**Table 2.** Saudi enterprise workload classes and infrastructure priorities.

| Workload class              | Dominant requirement           | Preferred platform pattern            | Key control                       |
|-----------------------------|--------------------------------|---------------------------------------|-----------------------------------|
| Distributed training        | Accelerator throughput         | Queued GPU pools with gang scheduling | Budget and dataset lineage        |
| Real-time inference         | Low latency and availability   | Autoscaled replicated endpoints       | SLO and rollback                  |
| Batch analytics / inference | Cost efficiency                | Elastic or interruptible compute      | Quota and scheduling window       |
| Edge inference              | Local response / data locality | Managed edge with central lifecycle   | Model version and device identity |

## 11 Proposed Saudi enterprise AI cloud framework

The framework consists of six layers, the first of which is the data and integration layer. This includes enterprise sources, object storage, streaming ingestion, feature stores and metadata. Instead of functioning as a one-way pipeline, the framework works on a closed loop basis.

## Saudi Enterprise AI Cloud Framework

A layered platform for scalable AI and ML workloads

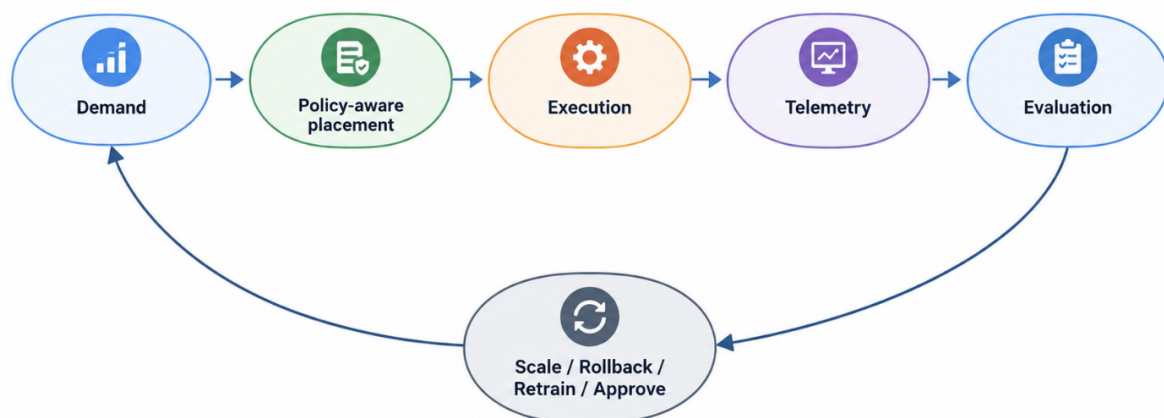


**Figure 1.** Layered Saudi Enterprise AI Cloud Framework.

Demand is introduced via experiments, scheduled training jobs or inference requests. The architecture brings together a number of different lines of evidence and includes predictive and machine-learning-based resource management to support proactive allocation [10,11]. Intentionally, the framework does not specify a single vendor and enterprises can implement the various layers using managed public-cloud services, private cloud, sovereign environments or a combination of both.

## Closed-Loop Operation of the AI Cloud Platform

Operational outcomes become evidence for future scheduling and governance



**Figure 2.** Closed-loop infrastructure operation linking workload demand, policy, telemetry and lifecycle actions.

---

## 12 Implementation roadmap and evaluation metrics

The implementation must take place in four stages, the first of which involves setting up the foundations such as cloud landing zones, identity, network segmentation, encrypted storage, data classification and basic cost controls.

The balanced scorecard should be used to measure each stage, and the scalability metrics should be queue time, jobs completed per accelerator-hour, inference throughput and scaling response.

The metrics must be considered as a whole; it is not always the case that a decrease in queue time resulting from permanently doubling the GPU capacity is an efficient one.

Saudi companies should also keep a close eye on various organizational indicators, including the percentage of teams that are using standardized pipelines, the extent of reuse of shared components, the level of staff capability coverage, and the ratio of production models that have designated business owners. Such measures will show if the platform is developing into an enterprise capability rather than remaining a centrally managed technical island.

## 13 Research gaps and future directions

There are still some gaps. First, a great deal of the technical literature assesses scheduling or serving mechanisms in controlled conditions, whereas enterprise environments involve governance constraints, a variety of different models, and integration with legacy systems. There is a need for more longitudinal research into how infrastructure patterns perform across a range of models rather than just on individual benchmarks. Second, although MLOps research has developed conceptually, empirical evidence regarding organizational adoption, lifecycle failure modes, and cost remains relatively scarce. The challenges outlined in recent MLOps and AIOps research indicate that technical automation by itself is not sufficient to overcome operational and business obstacles [21].

Third, research tailored to Saudi conditions should shift from a general focus on technology adoption to one that aims to achieve specific, measurable outcomes regarding AI infrastructure. Although the current studies carried out in Saudi Arabia offer useful evidence on the factors that determine the use of cloud services, AI readiness and applications in the digital sector [22-28], there are fewer investigations that quantify the extent to which accelerators are used, the time it takes to deploy models, the reliability of inference, the trade-offs involved with data locality, or the economics of AI platforms within Saudi companies. Obtaining such evidence would enable infrastructure recommendations to be adjusted to the actual conditions under which they operate.

Fourth, there needs to be a more in-depth investigation into data sovereignty and distributed learning. Although federated and edge methods can decrease the need to move data to a central location, they do bring with them new challenges relating to coordination, security and observability [6-9]. Research should examine whether the extra complexity is worth it when compared with the security offered by centralised processing. Fifth, sustainability metrics for AI infrastructure must be practically standardised so that companies can compare model performance with energy and resource usage.

In the end, the new generative models place greater demands on memory, the accelerators, and the economics of inference. Although the framework presented here is still applicable since it is based on workload rather than being specific to a model, further research should look into model compression,

the sharing of accelerators, retrieval infrastructure and policy controls concerning large generative workloads in Saudi enterprise environments.

## 14 Conclusion

The process of scaling artificial intelligence within enterprises in Saudi Arabia is at its core an infrastructure and operating model issue. Although cloud platforms offer elastic capacity, the value that enterprises derive depends on how compute resources, data, orchestration, model lifecycles, reliability, security and governance are integrated. Evidence from the period 2020 to 2025 indicates that workload prediction, accelerator-aware scheduling, cloud-edge distribution, federated learning, serverless inference and MLOps each address part of the problem.

The six-layer architectural structure together with a closed operational loop is how the proposed Saudi Enterprise AI Cloud Framework groups these capabilities, and it stresses the use of heterogeneous resource pools, policy-aware scheduling, tiered data services, automated model lifecycles, correlated observability, resilient deployment and integrated governance. The fact that Saudi Arabia has adopted the scheme shows that technological readiness, data quality, security and organizational aspects all need to develop in line with the platform.

For businesses, the main concern is not achieving the highest possible level of architectural complexity, but rather disciplined scalability—this means being able to add capacity and introduce automation while at the same time keeping service reliability, data control, cost visibility and accountability. Companies which set up these basic elements are then in a position to move from individual AI projects to offering reusable enterprise-level intelligence services; on the other hand, those who enlarge their models without maintaining proper lifecycle and governance practices run the risk of turning experimental successes into operational fragility. That is why the framework sees cloud infrastructure not as a place for hosting but as a managed and measurable foundation for sustainable AI adoption in Saudi Arabia.

## References

- [1] Aqeel I, Khormi IM, Khan SB, Shuaib M, Almusharraf A, Alam S, et al. Load balancing using artificial intelligence for cloud-enabled Internet of everything in healthcare domain. *Sensors*. 2023;23:5349. <https://doi.org/10.3390/s23115349>
- [2] Kreuzberger D, Kühl N, Hirschl S. Machine learning operations (MLOps): Overview, definition, and architecture. *IEEE Access*. 2023;11:31866–31879. <https://doi.org/10.1109/ACCESS.2023.3262138>
- [3] Lins S, Pandl KD, Teigeler H, Thiebes S, Bayer C, Sunyaev A. Artificial intelligence as a service. *Business & Information Systems Engineering*. 2021;63:441–456. <https://doi.org/10.1007/s12599-021-00708-w>
- [4] Deng S, Zhao H, Fang W, Yin J, Dustdar S, Zomaya AY. Edge intelligence: The confluence of edge computing and artificial intelligence. *IEEE Internet of Things Journal*. 2020;7(8):7457–7469. <https://doi.org/10.1109/JIOT.2020.2984887>
- [5] Mendez J, Bierzynski K, Cuéllar MP, Morales DP. Edge intelligence: Concepts, architectures, applications, and future directions. *ACM Transactions on Embedded Computing Systems*. 2022;21(5):48. <https://doi.org/10.1145/3486674>
- [6] Xia Q, Ye W, Tao Z, Wu J, Li Q. A survey of federated learning for edge computing: Research problems and solutions. *High-Confidence Computing*. 2021;1(1):100008. <https://doi.org/10.1016/j.hcc.2021.100008>
- [7] Abreha HG, Hayajneh M, Serhani MA. Federated learning in edge computing: A systematic survey. *Sensors*. 2022;22(2):450. <https://doi.org/10.3390/s22020450>
- [8] Zhang C, Xie Y, Bai H, Yu B, Li W, Gao Y. A survey on federated learning. *Knowledge-Based Systems*. 2021;216:106775. <https://doi.org/10.1016/j.knosys.2021.106775>

- [9] Bao G, Guo P. Federated learning in cloud-edge collaborative architecture: Key technologies, applications and challenges. *Journal of Cloud Computing*. 2022;11:94. <https://doi.org/10.1186/s13677-022-00377-4>
- [10] Goodarzy S, Nazari M, Han R, Keller E, Rozner E. Resource management in cloud computing using machine learning: A survey. In: *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*; 2020. p. 811–816. <https://doi.org/10.1109/ICMLA51294.2020.00132>
- [11] Bi J, Li S, Yuan H, Zhou MC. Integrated deep learning method for workload and resource prediction in cloud systems. *Neurocomputing*. 2021;424:35–48. <https://doi.org/10.1016/j.neucom.2020.11.011>
- [12] Hu Q, Sun P, Yan S, Wen Y, Zhang T. Characterization and prediction of deep learning workloads in large-scale GPU datacenters. In: *International Conference for High Performance Computing, Networking, Storage and Analysis (SC '21)*; 2021. <https://doi.org/10.1145/3458817.3476223>
- [13] Nam SA, Cho K, Bahn H. Adaptive resource planning for AI workloads with variable real-time tasks. *Computers, Materials & Continua*. 2023;74(3):6823–6833. <https://doi.org/10.32604/cmc.2023.035481>
- [14] Ali A, Pinciroli R, Yan F, Smirni E. BATCH: Machine learning inference serving on serverless platforms with adaptive batching. In: *International Conference for High Performance Computing, Networking, Storage and Analysis (SC20)*; 2020. <https://doi.org/10.1109/SC41405.2020.00073>
- [15] Yu M, et al. Gillis: Serving large neural networks in serverless functions with automatic model partitioning. In: *2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS)*; 2021. p. 138–148. <https://doi.org/10.1109/ICDCS51616.2021.00022>
- [16] Jarachanthan J, Chen L, Xu F, Li B. AMPS-Inf: Automatic model partitioning for serverless inference with cost efficiency. In: *50th International Conference on Parallel Processing (ICPP '21)*; 2021. <https://doi.org/10.1145/3472456.3472501>
- [17] Li B, Patel T, Gadepally V, Gettings K, Tiwari D. DASH: Scheduling deep learning workloads on multi-generational GPU-accelerated clusters. In: *2022 IEEE High Performance Extreme Computing Conference (HPEC)*; 2022. <https://doi.org/10.1109/HPEC55821.2022.9926390>
- [18] Carrión C. Kubernetes scheduling: Taxonomy, ongoing issues and challenges. *ACM Computing Surveys*. 2022. <https://doi.org/10.1145/3539606>
- [19] Janbi N, Katib I, Albeshri A, Mehmood R. Distributed Artificial Intelligence-as-a-Service (DAIaaS) for Smarter IoT and 6G Environments. *Sensors*. 2020;20(20):5796. <https://doi.org/10.3390/s20205796>
- [20] Dong W. AIOps architecture in data center site infrastructure monitoring. *Computational Intelligence and Neuroscience*. 2022;2022:1988990. <https://doi.org/10.1155/2022/1988990>
- [21] Díaz-de-Arcaya J, et al. A joint study of the challenges, opportunities, and roadmap of MLOps and AIOps: A systematic survey. *ACM Computing Surveys*. 2024. <https://doi.org/10.1145/3625289>
- [22] Ayadi F. Critical factors affecting the decision to adopt cloud computing in Saudi health care organizations. *Electronic Journal of Information Systems in Developing Countries*. 2022;88(6):e12231. <https://doi.org/10.1002/isd2.12231>
- [23] Aldahwan NS, Ramzan MS. Quadruple theories based determinants and their causal relationships affecting the adoption of community cloud in Saudi HEI. *BioMed Research International*. 2022;2022:2382535. <https://doi.org/10.1155/2022/2382535>
- [24] Alrebdi N, Khan N. Core elements impacting cloud adoption in the government of Saudi Arabia. *International Journal of Advanced Computer Science and Applications*. 2022;13(6). <https://doi.org/10.14569/IJACSA.2022.0130633>
- [25] Alshahrani H, et al. Investigating the factors influencing the use of cloud computing. *Journal of Information Security and Cybercrimes Research*. 2022;5(2):104–115. <https://doi.org/10.26735/RNOA9602>
- [26] Alarefi M. The impact of artificial intelligence on business performance in Saudi Arabia: The role of technological readiness and data quality. *Engineering, Technology & Applied Science Research*. 2024;14(5):16802–16807. <https://doi.org/10.48084/etasr.7871>
- [27] Al-Baity HH. The artificial intelligence revolution in digital finance in Saudi Arabia: A comprehensive review and proposed framework. *Sustainability*. 2023;15(18):13725. <https://doi.org/10.3390/su151813725>
- [28] Afandi W. Cloud computing and organizational resilience: An empirical investigation in the context of Saudi enterprises. *International Journal of Computer Applications*. 2021;183(30):47–58. <https://doi.org/10.5120/ijca2021921693>

- 
- [29] Felemban H, Sohail M, Ruikar K. Exploring the readiness of organisations to adopt artificial intelligence. *Buildings*. 2024;14(8):2460. <https://doi.org/10.3390/buildings14082460>
- [30] Rajagopal A, et al. Machine learning operations in health care: A scoping review. *Mayo Clinic Proceedings: Digital Health*. 2024;2(3):421–437. <https://doi.org/10.1016/j.mcpdig.2024.06.009>