stack8s – Managed Kubeflow as a Service

G-Cloud 15 Service Definition

1. Service description

stack8s Managed Kubeflow is a fully managed MLOps platform for building, deploying, and operating machine learning (ML) workflows on secure, sovereign, multi-cloud infrastructure. It provides a production-grade Kubeflow environment that includes pipeline orchestration, notebook-based development, experiment tracking, model registry, and scalable model serving, all exposed through an intuitive web UI and APIs that require no infrastructure coding. The service is delivered as Infrastructure Software as a Service (iSaaS) under Lot 2a and is designed to run across public, private, community, and hybrid cloud deployments while maintaining strict data residency and sovereignty boundaries.

The platform abstracts the complexity of provisioning and operating Kubeflow and Kubernetes for ML workloads. stack8s is responsible for setting up and managing clusters, GPU and CPU resources, storage, networking, and security controls across approved cloud and on-premise environments, while customers focus on building ML pipelines and models. Access is provided via a secure web management portal and programmatic APIs/Terraform, integrated with enterprise identity providers and role-based access control (RBAC).

Managed Kubeflow is built for teams who need to industrialise ML: it standardises how data scientists, ML engineers, and DevOps teams collaborate, with an emphasis on reproducibility, auditability, and operational resilience. The platform is suitable for public sector organisations and regulated industries that require sovereign hosting, strong isolation, and clear shared responsibility between platform provider and ML teams.

---

2. Service features and benefits

2.1 Service features

- **End-to-end ML pipeline orchestration** The service provides Kubeflow Pipelines for designing, running, and managing complex ML workflows from data ingestion through training, validation, and deployment. Pipelines can be defined using Python SDKs or YAML and scheduled for recurring runs, enabling robust, repeatable ML processes with lineage and metadata captured at every step.
- **Jupyter Notebooks and Visual Studio Code integration** Data scientists can develop and run experiments using managed Jupyter notebooks directly within the platform, with access to cluster resources and data sources. Integration with Visual Studio Code allows engineers to work from their preferred IDE while connecting to Kubeflow-backed environments for execution, debugging, and collaboration.

CONFIDENTIAL - DO NOT DISTRIBUTE WITHOUT PERMISSION

- **Experiment tracking and model registry** Managed Kubeflow includes experiment tracking and model registry capabilities to capture parameters, metrics, artefacts, and versions across runs. Teams can compare experiments, promote models through environments, and maintain an auditable history of what was trained, when, and with which data and hyperparameters.
- **Distributed training with TensorFlow and PyTorch** The platform supports distributed training workloads using popular frameworks such as TensorFlow and PyTorch, leveraging Kubernetes to orchestrate multi-node training jobs. Users can configure resource requirements (GPU/CPU, memory) and distribution strategies, and the platform handles scheduling, scaling, and fault tolerance.
- **Scalable model serving and deployment** Managed Kubeflow provides scalable model serving components for deploying models as APIs or batch inference services. Autoscaling policies can be configured to adjust replicas based on demand, ensuring models remain responsive during peak usage while optimising resource consumption during quieter periods.
- **Hyperparameter tuning and optimisation** The service offers tools for automated hyperparameter tuning, allowing users to define search spaces and optimisation strategies for experiments. This enables efficient exploration of model configurations at scale, improving model performance without manual trial-and-error.
- **Integrated DevOps and MLOps** The platform is designed to integrate with CI/CD pipelines and infrastructure-as-code workflows, enabling automated deployment of ML pipelines and services. APIs and Terraform support allow teams to embed Kubeflow operations into existing DevOps tooling, unifying application and ML lifecycle management.
- **GPU and CPU auto-scaling** Managed Kubeflow automatically scales compute resources (both GPU and CPU) based on workload demands and predefined quotas. This ensures efficient utilisation of resources, enabling cost-effective training and inference while maintaining performance targets.
- **Secure multi-tenant workload isolation** The service enforces logical and physical isolation between tenants using Kubernetes namespaces, RBAC, quotas, and network policies. Where required, dedicated clusters or environments can be provided to fully isolate workloads and guarantee performance for specific customers or projects.
- **No-code UI for infrastructure setup and management** A web-based console allows users to configure Kubeflow environments, clusters, and ML workflows without writing infrastructure code. This reduces the barrier to entry for organisations that do not have specialist Kubernetes or MLOps skills, while still allowing advanced users to interact via APIs and Terraform.

2.2 Service benefits

- **Accelerate ML model development and deployment** By providing an integrated environment for pipelines, notebooks, training, and serving, the service significantly shortens the path from idea to deployed ML model. Teams can focus on feature engineering and model quality rather than provisioning and maintaining infrastructure.
- **Accelerated time-to-value** Pre-integrated components and managed operations allow organisations to start running ML pipelines quickly, reducing the time and effort required to build a production MLOps platform from scratch. This is particularly valuable for departments seeking to deliver ML-backed services within fixed budgets and timelines.
- **Reduced infrastructure management complexity** stack8s manages cluster lifecycle, upgrades, capacity, and security baselines, removing the need for dedicated in-house Kubernetes/MLOps

platform teams. Customers can consume a standardised, hardened platform rather than assembling and operating diverse tools themselves.

- **Elimination of vendor lock-in** The platform is built on open-source Kubeflow and Kubernetes, deployed across multiple sovereign cloud providers and on-premise environments. Customers retain the ability to export configurations, workloads, and artefacts, minimising lock-in to a single vendor or infrastructure provider.
- **Access to UK-based specialist expertise** Support and platform operations are delivered by UK-based engineers with expertise in Kubernetes, multi-cloud infrastructure, and MLOps. This supports public sector assurance requirements and provides direct access to specialists who understand sovereign cloud and regulated environments.
- **Reproducible and auditable ML experiments** Experiment tracking, pipeline metadata, and audit logs enable reproducibility and provide a clear record of how models were developed and deployed. This is essential for explaining model decisions, passing internal and external audits, and complying with emerging AI governance requirements.
- **Assured service continuity (SLA)** The service is backed by an SLA that covers the stack8s control plane and network mesh, with defined response and resolution targets for incidents. This assurance enables teams to rely on Managed Kubeflow for production workloads with clear expectations on availability and support.
- **Collaborative ML workflows and team productivity** Shared workspace support, integrated tools, and consistent identity and access control allow data scientists, ML engineers, and operations staff to collaborate effectively. This reduces friction between teams and improves productivity across the ML lifecycle.
- **Global scalability with local performance** Multi-cloud and hybrid support enable deployments close to data sources and users, improving performance while maintaining sovereignty. Auto-scaling ensures the platform can handle spikes in training and inference workloads without manual intervention.
- **No more infrastructure and cloud integrations headaches** The platform encapsulates integration with multiple cloud providers, on-premise environments, and networking constructs behind a unified control plane. Customers are relieved from managing complex cloud integrations, VPNs, and cluster provisioning scripts themselves.

---

3. Use cases

Managed Kubeflow supports a wide range of ML and AI workloads, including:

- **End-to-end MLOps for data science teams** Providing a managed environment where data scientists can design pipelines, run experiments, and deploy models without needing to understand Kubernetes or cloud infrastructure internals.
- **Production ML for citizen-facing services** Powering ML models behind citizen-facing applications (for example, eligibility triage, demand forecasting, document classification) that require stable, scalable model serving with strict sovereignty and audit controls.
- **Hybrid and multi-cloud ML platforms** Enabling organisations to run training and inference across multiple approved cloud providers and on-premise clusters under a single control plane, while maintaining jurisdictional boundaries for data and workloads.

CONFIDENTIAL - DO NOT DISTRIBUTE WITHOUT PERMISSION

- **Research and innovation environments** Providing a standardised platform for research units to prototype and test ML approaches with GPU access and experiment tracking, with a clear path to production without replatforming.
- **AI/ML platform for regulated sectors** Supporting regulated environments that require reproducibility, clear audit trails, strong access controls, and separation of responsibilities between platform operations and data/model ownership.

---

4. Pricing overview

Pricing for stack8s Managed Kubeflow is provided in the stack8s UK G-Cloud 15 Pricing Document within the Digital Marketplace. That document details charging models for the control plane, cluster resources (CPU/GPU, storage), support tiers, and any optional professional services.

---

5. Service constraints

- **Planned and emergency maintenance windows** Planned maintenance typically occurs weekly on Sundays between 00:00 and 05:00 GMT, with at least 7 days' advance notice to customers. Emergency maintenance may be performed with a minimum of 4 hours' notice where urgent security or stability issues must be addressed.
- **Scope of support and shared responsibility model** Support is limited for customer-deployed Kubernetes objects, unsupported hardware, and non-native runtimes. Under the shared security model, stack8s secures and operates the platform (control plane, network mesh, managed services), while customers remain responsible for the security and compliance of their data, models, and applications running on the platform.
- **SLA coverage boundaries** The SLA covers the stack8s control plane and managed network mesh, but does not extend to the availability of third-party IaaS providers used to host clusters and resources. Incidents caused solely by underlying provider outages may be outside the direct control of stack8s, although mitigations such as multi-region or multi-provider deployments can be configured.
- **Data ingress/egress responsibilities** Customers are responsible for their own data ingress and egress costs with underlying cloud providers. The stack8s managed private network mesh is engineered to minimise egress charges by optimising routing and locality of workloads and data where possible.
- **Professional services scope** A billable Professional Services team is available for ad-hoc support, custom integrations, and features not covered by the standard managed service. Such work is out of scope for the base service and is provided under separate commercial arrangements.

CONFIDENTIAL - DO NOT DISTRIBUTE WITHOUT PERMISSION

6. Technical requirements

- **Client and identity prerequisites**
  - Latest stable web browser with HTTPS and JavaScript enabled for accessing the management portal.
  - An active stack8s account with appropriate security access and role assignments.
  - Customer identity provider must support SAML 2.0 or OAuth 2.0 for federated authentication.
- **Application packaging and runtime requirements**
  - All applications and workloads deployed to the platform must be packaged as OCI-compliant container images.
  - Customers are responsible for building container images that comply with their internal security baselines and policy requirements.
- **Network and connectivity requirements**
  - Outbound HTTPS (port 443) access is required from customer environments to the stack8s control plane for API and portal access.
  - Customer VPC/VNet must support VPN peering or equivalent secure connectivity to integrate with the stack8s managed private network mesh.
  - Customer security policies must permit the necessary managed infrastructure connectivity for cluster operations, monitoring, and updates.
- **Commercial and security prerequisites**
  - An active stack8s billing agreement is required for provisioning and running resources under the service.
  - Customer security policies must allow deployment and management of infrastructure components by stack8s in approved environments.
  - Customers must install endpoint protection on all host environments under their control that interact with or host workloads managed via the service.

CONFIDENTIAL - DO NOT DISTRIBUTE WITHOUT PERMISSION