Applied AI Technologies Studio

Technical Specifications Summary

Q1 2024

The Solution

The AI Studio facilitates the delivery of and access to enterprise level AI technology in a managed and fully secured architecture. AI Studio users experience powerful AI models while remaining in compliance with corporate and regulatory, information and security protocols.

The AI Studio Solution is -

Secure - 100% secured behind a corporate firewall

Ensuring no data or transactions of any kind are transmitted to 3rd party AI engines or solutions. Models, vectoring solutions, data stores and all technology can be run inside a private cloud or even in an on-premise solution.

Model agnostic

Any local LLM can be deployed to enable users to engage with various models and to obtain "best in breed" solutions for targeted needs. The fully managed solution ensures that new models are tested and deployed on demand as they become available. Users can select their preferred model based on results.

Flexible

Private or public cloud or on-premise solution flexibility. The Studio can be deployed on hardware in any corporate environment ensuring compliance with corporate security and regulatory requirements. More details on hardware are provided below.

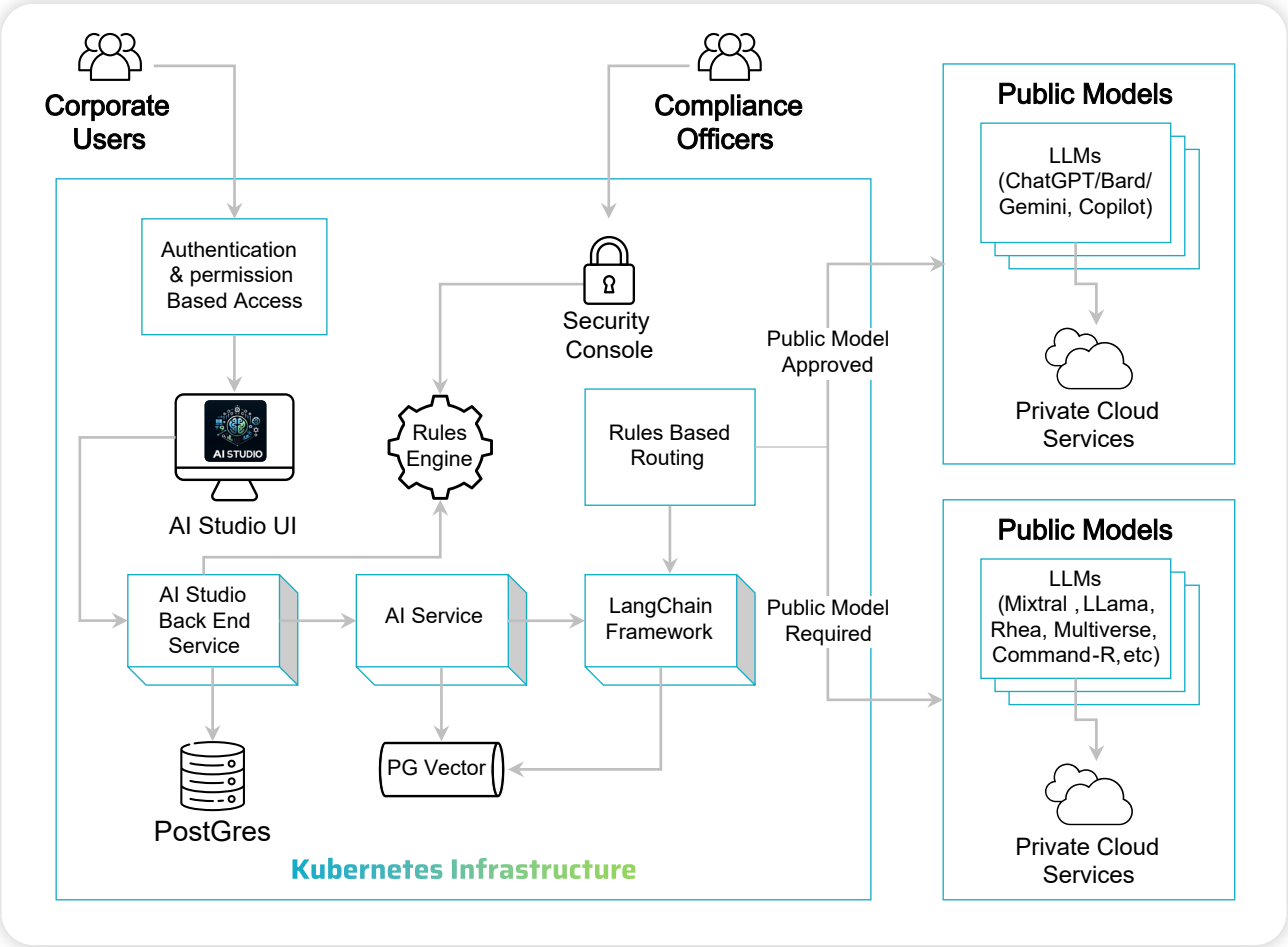
Simple and customizable

The AI Studio interface can be white-labelled and fitted for secure corporate architectures while still maintaining unparalleled ease of use by non-technical users. Simple chat BOT solutions and more sophisticated custom and "managed" BOTs can be created, tested, and deployed in a managed environment. The AI Studio empowers both entry level and advanced users the capacity to leverage the full power of deployed models and allows corporate branding and customization where required.

Supported by a world - class AI team

[s]Cube delivers industry expertise to supplement corporate AI initiatives, enabling clients to focus on their core business instead of managing AI resources. Managed services support options include implementation, cloud deployment, UI customizations, prompt engineering, technical support, training, new model testing and release management, all fully directed by IT and in compliance with mandated policies and procedures.

Below is an architecture diagram of the AI Studio illustrating the various components:



AI Studio Infrastructure Requirements

Deploying the AI Studio across various environments such as AWS, Microsoft Azure, Google Cloud, and on-premises requires a detailed understanding of the infrastructure components necessary to support its operations. This includes the requirements for storage, application deployment within a Kubernetes environment, and the GPU cluster needed for model execution.

Cloud & On-Premises Deployment Specifications

1. Kubernetes Environment

- **Cluster Configuration:** A Kubernetes cluster with at least 8 vCPUs and 32GB of RAM for minimal deployments. Production environments may require clusters with significantly more resources based on the workload.
- **Namespace:** Dedicated namespaces for AI Studio services to ensure isolation from other services.
- **Persistent Volume Claims (PVCs):** For stateful components like databases, ensure PVCs are adequately sized and backed by high -performance storage classes.
- **Ingress Controller:** An ingress controller configured to manage external access to the AI Studio services.

2. Storage Requirements:

- **Cloud Environments (AWS S3, Azure Blob Storage, Google Cloud Storage):** High-performance object storage is necessary to store the data ingested by the AI Studio. The storage solution must support scalability, data redundancy, and high availability.
- **On-Premises:** NFS or a SAN (Storage Area Network) capable of handling high I/O throughput and large storage capacities is recommended. Ensure PostgreSQL is configured with sufficient storage for the vector database operations.

3. GPU Cluster for Model Execution:

Note: the exact specifications may vary depending on the model used

- **AWS:** Utilize p3 or g4dn instances equipped with NVIDIA V100 or T4 GPUs respectively.
- **Azure:** NCas T4 v3 or NCv3 instances are recommended for their NVIDIA Tesla T4 and V100 GPUs.
- **Google Cloud:** Use A2 VMs or Compute Engine VMs with attached T4 or V100 GPUs.
- **On-Premises:** NVIDIA Tesla T4 or V100 GPUs, ensuring CUDA and cuDNN libraries are installed for optimal performance.

4. Vector Database

- **Database Setup:** PGVector requires PostgreSQL 13 or newer with PGVector extension installed. This setup enables efficient storage and querying of vector data for AI operations

Implementation Considerations

Scalability

Auto-scaling groups in cloud environments or Kubernetes horizontal pod autoscaler should be configured to handle varying loads.

Security

Implement robust security measures, including network policies in Kubernetes, secure access to storage, and encrypted data transmission.

Backup & Recovery

Regularly scheduled backups of both application data and PostgreSQL databases, including PGVector data, to ensure business continuity.

877-437-2823
info@scubeenterprise.com
www.scubeenterprise.com

AI STUDIO