Comprehensive Implementation Guide: Azure Machine Learning MLOps with Responsible AI Controls (PRJ-AZURE-AI-054)

Author: Manus AI **Date:** January 26, 2026

1. Project Overview

This project, **PRJ-AZURE-AI-054**, establishes a secure, compliant, and enterprise-scale **Azure Machine Learning (Azure ML) MLOps** foundation. Its core focus is the integration of **Responsible AI** principles and robust **Governance, Risk, and Compliance (GRC)** controls directly into the model lifecycle. This solution is designed for organizations operating in highly regulated sectors, such as finance, healthcare, and government, where data privacy, model transparency, and regulatory adherence are paramount.

The solution provides a trusted environment for the entire AI lifecycle, from experimentation and training to deployment and monitoring. It specifically addresses the challenges of integrating advanced AI services, such as **Azure OpenAI Service** and **Azure Cognitive Services**, into a secure enterprise architecture. By enforcing network isolation, strict access controls, and automated bias detection, the project ensures that AI initiatives can scale without compromising security or ethical standards.

The key features of this MLOps framework include:

- **End-to-End MLOps Pipeline:** Automated workflows for data preparation, model training, registration, deployment, and monitoring.
- **Responsible AI Integration:** Use of the `responsibleai` SDK and Azure ML capabilities for fairness assessment, interpretability, and error analysis.

- **GRC-by-Design:** Explicit mapping of technical controls to major compliance frameworks (NIST AI RMF, ISO/IEC 42001, GDPR, SOC 2).
- **Secure Foundation:** Deployment of the Azure ML Workspace with High Business Impact (HBI) settings, network isolation (Private Link), and Managed Identities.

This guide provides the detailed, step-by-step instructions required to deploy and configure this production-ready, GRC-compliant MLOps environment on Azure.

---

2. Business Context

The rapid adoption of Artificial Intelligence, particularly large language models (LLMs) and generative AI, introduces significant opportunities but also complex risks. This project is a strategic response to these challenges, transforming potential liabilities into a competitive advantage through structured governance.

The Problem: Unmitigated AI Risk

Organizations adopting AI at scale face a triad of critical risks:

1. **Compliance and Regulatory Risk:** Evolving global regulations (e.g., EU AI Act, GDPR) impose strict requirements on AI systems, particularly those classified as "high-risk." Failure to comply can result in massive fines and reputational damage.
2. **Ethical and Reputational Risk:** Uncontrolled models can exhibit bias, lack transparency, or generate inappropriate content, leading to unfair outcomes, loss of public trust, and brand damage.
3. **Security and Data Leakage Risk:** Integrating external AI services without proper network and access controls can expose sensitive data, violating data sovereignty and privacy mandates.

A lack of robust **AI governance** and a fragmented approach to MLOps are the primary causes of these unmitigated risks.

The Solution: Secure Azure AI Implementation Framework

This solution establishes a **secure Azure AI implementation framework** that embeds responsible AI controls and rigorous model governance directly into the MLOps pipeline. It leverages the native security and compliance features of Azure services to create a **Trusted AI Environment**.

Key aspects of the solution include:

- **Data Sovereignty:** All data remains within the secure Azure boundary, utilizing Azure Private Link and VNet injection to ensure network isolation.
- **Automated Governance:** Azure Policy is used to enforce organizational standards, such as requiring model cards and bias reports before model promotion.
- **Integrated Responsible AI Tools:** The MLOps pipeline is mandated to include steps for bias detection, interpretability, and fairness assessment using the Responsible AI Dashboard and SDK.

Quantified Business Value and ROI

The implementation of this GRC-focused MLOps solution delivers measurable business value:

Value Proposition	Description	Quantified Impact (Estimated)
Risk Mitigation & Compliance	Proactively meets evolving regulatory requirements (e.g., AI Act, GDPR). Reduces the likelihood of non-compliance fines and legal costs.	**ROI:** Estimated 15-25% reduction in potential regulatory fines and legal defense costs over five years.
Operational Efficiency	Automated MLOps pipelines accelerate the time-to-market for new AI models while ensuring quality and compliance checks are never skipped.	**Efficiency Gain:** 30-40% faster deployment cycle for compliant models compared to manual, ad-hoc processes.
Data Privacy & Security	Network isolation and HBI settings eliminate data exposure to the public internet and third-party services, maintaining strict data sovereignty.	**Cost Savings:** Avoidance of costs associated with data breaches (estimated at $4.45 million per breach on average [1]).
Responsible AI & Trust	Integrated bias detection and transparency tools build internal and external trust in AI outcomes, reducing the risk of model failure in production.	**Business Value:** Increased adoption rate of AI solutions across the enterprise due to higher trust and reliability.
Cost Optimization	Centralized management and automated scaling of compute resources (covered in Section 9).	**Cost Savings:** 10-20% reduction in cloud compute costs through efficient resource utilization.

---

3. GRC Mapping

Governance, Risk, and Compliance (GRC) are not add-ons but foundational pillars of this MLOps solution. The architecture explicitly maps technical controls to industry frameworks and regulatory requirements, ensuring auditability and a defensible compliance posture.

Compliance Frameworks Alignment

The solution is designed to align with the following leading AI and security compliance standards:

Framework	Focus Area	Technical Control Mapping in PRJ-AZURE-AI-054
Microsoft Responsible AI Standard	Fairness, reliability, privacy, inclusiveness, transparency, and accountability.	Integrated bias detection, Model Cards, Content Moderation on LLMs, HBI Workspace setting.
NIST AI Risk Management Framework (AI RMF)	Comprehensive AI risk management lifecycle, from design to deployment and monitoring.	Model versioning, MLOps pipeline gates for risk assessment, comprehensive AI logging.
ISO/IEC 42001 (AI Management)	Requirements for establishing, implementing, maintaining, and continually improving an AI management system.	Defined MLOps processes, documented architecture, use of Azure Policy for continuous compliance.
OWASP Top 10 for LLM	Mitigation of the most critical security risks specific to Large Language Models (e.g., Prompt Injection, Data Leakage).	Network isolation (Private Link), Content Filtering, Strict RBAC on Azure OpenAI endpoints.
SOC 2 (Trust Services Criteria)	Security, Availability, Processing Integrity, Confidentiality, and Privacy.	Access Controls (RBAC), Comprehensive AI Logging (Security), Data Encryption (Confidentiality).

Security Controls and Audit Evidence

The MLOps environment is configured to automatically generate and retain the necessary evidence for compliance audits:

Security Control Implemented	GRC Alignment	Audit Evidence Generated
Content Filtering and Moderation	Microsoft Responsible AI Standard, OWASP LLM Top 10	Content moderation logs from Azure OpenAI/Cognitive Services.
Data Encryption and Access Controls	GDPR, HIPAA, SOC 2	Azure Key Vault access logs, Azure RBAC assignment reports, HBI Workspace configuration.
Model Versioning and Governance	NIST AI RMF, ISO/IEC 42001	Azure ML Model Registry history, MLOps pipeline approval logs.
Bias Detection and Mitigation	Microsoft Responsible AI Standard, AI Act (EU)	Fairness and Bias Evaluation Reports generated by the `responsibleai` SDK and stored as pipeline artifacts.
Comprehensive AI Logging	AI Act (EU), SOC 2	AI Interaction Logs and Audit Trails captured in Azure Monitor and Log Analytics Workspace.

Regulatory Alignment Deep Dive

The architecture directly addresses key articles and requirements from major global regulations:

Regulation	Requirement/Article	Control Alignment	Implementation Detail
AI Act (EU)	High-risk AI requirements (e.g., transparency, human oversight, technical robustness).	Model Cards, Bias Detection, Comprehensive Logging.	Mandatory Model Card generation in the MLOps pipeline; Human review gate before production deployment.
GDPR	Article 22 (Automated decisions), Article 35 (DPIA), Data Minimization.	Data Processing Records, Model Transparency, Data Encryption.	Data used for training is pseudonymized; Data Processing Records are maintained for all data used in training and inference.
HIPAA	§ 164.308(a)(3) (Workforce access to AI systems), Protected Health Information (PHI) security.	Strict RBAC, Data Encryption, Audit Trails.	PHI is never stored in the ML Workspace; Access to training data and models is restricted to authorized personnel via Just-In-Time (JIT) access.
SOC 2	CC6.1 (Data access), CC7.2 (Monitoring), CC8.1 (Change Management).	Access Controls, Comprehensive AI Logging, Model Versioning.	All changes to the MLOps pipeline are tracked via Git and deployed via CI/CD; Continuous monitoring of model performance and data drift.

---

4. Prerequisites

Successful deployment requires a prepared environment and the necessary access permissions.

4.1. Azure Subscription and Permissions

- **Active Azure Subscription:** Required to provision resources.
- **Permissions:** The deploying identity (user or Service Principal) must have the `Owner` or `Contributor` role on the target Resource Group, and permissions to

create Service Principals and assign roles.

4.2. Local Development Tools

The following tools must be installed and configured on the local machine or CI/CD agent:

Tool	Purpose	Installation/Configuration
Azure CLI	Command-line interface for deploying and managing Azure resources.	Install via `curl -sL https://aka.ms/InstallAzureCLIDeb
Git	Version control for MLOps code, configuration, and model scripts.	Install via `sudo apt install git`. Configure user name and email.
Docker	Containerization for model training and inference environments.	Install Docker Engine. Ensure the user is in the `docker` group (`sudo usermod -aG docker $USER`).
Python (3.8+)	Execution environment for MLOps scripts and Responsible AI SDK.	Install via `sudo apt install python3.10 python3.10-venv`.

4.3. Service Principal Setup

For automated, non-interactive deployment (CI/CD), a Service Principal (SP) is mandatory.

1. **Create the Service Principal:**

```
az ad sp create-for-rbac --name "http://prj-azure-ai-054-sp" --role "Contributor" --scopes "/subscriptions/<SUBSCRIPTION_ID>/resourceGroups/<RESOURCE_GROUP_NAME>" --json-auth
```

2. **Output Credentials:** Save the `appId`, `password`, and `tenant` from the output. These will be used to log in to Azure in the CI/CD pipeline.

```
# Example login in CI/CD
az login --service-principal -u <appId> -p <password> --tenant
<tenant>
```

---

5. Architecture Overview

The solution architecture is a secure, hub-and-spoke model centered around the **Azure Machine Learning Workspace**. This design ensures that all components are interconnected via private endpoints, eliminating exposure to the public internet and adhering to the principle of least privilege.

Core Components

1. **Azure ML Workspace (The Hub):** The central control plane for all MLOps activities. It manages compute, data, models, and endpoints.
    - **Key Configuration:** Deployed with `hbi-workspace=true` to enable enhanced data protection and compliance features. Network access is restricted to private endpoints.

2. **Azure Key Vault (Security Core):** Securely stores all secrets, including connection strings, API keys for Azure OpenAI, and service principal credentials.
    - **Access:** Only the Azure ML Workspace's Managed Identity and the MLOps pipeline SP are granted `Get` permissions on necessary secrets.

3. **Azure Data Lake Storage Gen2 (ADLS Gen2):** Secure, tiered storage for all data assets (training data, model artifacts, logs).
    - **Security:** Access is restricted via VNet integration and RBAC. Data is encrypted at rest, and double encryption is recommended.

4. **Azure Policy & RBAC (Governance Layer):** Enforces organizational standards and least-privilege access across all resources.
    - **Policy Enforcement:** Custom policies ensure that only compliant models (e.g., those with a Model Card and Bias Report) can be registered or deployed.

5. **Azure Kubernetes Service (AKS) or Azure ML Managed Endpoint (Inference):** Provides a scalable, secure environment for real-time and batch inference.
    - **Security:** Deployed within a VNet or using Managed Endpoints with private access, ensuring the model endpoint is not publicly accessible.
6. **Azure OpenAI Service/Cognitive Services (Advanced AI):** Integrated for advanced capabilities.
    - **Security:** Deployed with network isolation (Private Endpoint) and mandatory content filtering/moderation enabled to block inappropriate inputs/outputs.

Data and Control Flow

1. **Data Ingestion:** Data is securely loaded into ADLS Gen2 via a private link.
2. **Training:** The Azure ML Compute Cluster (deployed in the VNet) accesses data from ADLS Gen2 via a private link, trains the model, and logs metrics to the ML Workspace.
3. **Responsible AI Assessment:** A dedicated step in the MLOps pipeline uses the `responsibleai` SDK to generate fairness and interpretability reports.
4. **Model Registration:** The model and its associated reports (Model Card, Bias Report) are registered in the ML Workspace Model Registry.
5. **Deployment:** The MLOps pipeline promotes the model to a staging or production endpoint (AKS/Managed Endpoint), which accesses the model artifact from ADLS Gen2 and secrets from Key Vault, all via private links.
6. **Monitoring:** Inference requests and model performance metrics are logged to Azure Monitor/Log Analytics, providing a comprehensive audit trail.

---

6. Step-by-Step Implementation

The deployment uses the Azure CLI for a scripted, repeatable, and idempotent process.

6.1. Setup Environment Variables

Define all necessary variables. Use a consistent naming convention to simplify management and auditing.

```
# Project and Location Variables
export PROJECT_ID="PRJ-AZURE-AI-054"
export RESOURCE_GROUP="rg-${PROJECT_ID}-mlops"
export LOCATION="eastus" # Choose a region with Azure ML and Azure OpenAI
support
export ML_WORKSPACE_NAME="mlw-${PROJECT_ID}"
export KEY_VAULT_NAME="kv-${PROJECT_ID}-sec"
export ACR_NAME="acr${PROJECT_ID//-}" # ACR names must be globally unique
and lowercase
export VNET_NAME="vnet-${PROJECT_ID}"
export SUBNET_ML="snet-ml-workspace"
export SUBNET_COMPUTE="snet-ml-compute"
```

6.2. Create Resource Group and VNet

Create the resource group and a virtual network (VNet) to host all resources, ensuring network isolation from the start.

```
# 1. Create Resource Group
az group create --name $RESOURCE_GROUP --location $LOCATION

# 2. Create Virtual Network and Subnets
az network vnet create \
    --name $VNET_NAME \
    --resource-group $RESOURCE_GROUP \
    --location $LOCATION \
    --address-prefix 10.0.0.0/16

# Subnet for ML Workspace and Private Endpoints
az network vnet subnet create \
    --vnet-name $VNET_NAME \
    --resource-group $RESOURCE_GROUP \
    --name $SUBNET_ML \
    --address-prefixes 10.0.0.0/24 \
    --service-endpoints Microsoft.KeyVault Microsoft.Storage

# Subnet for ML Compute Cluster (requires delegation for AKS/Compute)
az network vnet subnet create \
    --vnet-name $VNET_NAME \
    --resource-group $RESOURCE_GROUP \
    --name $SUBNET_COMPUTE \
    --address-prefixes 10.0.1.0/24
```

6.3. Deploy Azure Key Vault

Deploy the Key Vault and configure a basic access policy for the deploying user/SP.

```
az keyvault create \
    --name $KEY_VAULT_NAME \
    --resource-group $RESOURCE_GROUP \
    --location $LOCATION \
    --enabled-for-deployment true \
    --sku standard

# Note: Access policies for the ML Workspace Managed Identity will be set
after the workspace is created.
```

6.4. Deploy Azure Machine Learning Workspace

This is the most critical step. The workspace is deployed with HBI settings and network isolation enforced by setting `allow-public-access-when-behind-vnet` to `false`.

```
az ml workspace create \
    --name $ML_WORKSPACE_NAME \
    --resource-group $RESOURCE_GROUP \
    --location $LOCATION \
    --hbi-workspace true \
    --key-vault $KEY_VAULT_NAME \
    --identity system_assigned \
    --vnet-name $VNET_NAME \
    --subnet $SUBNET_ML \
    --allow-public-access-when-behind-vnet false
```

6.5. Configure Private Endpoints for Dependencies

The ML Workspace requires private links to its dependent resources (Storage, Key Vault, Container Registry).

```
# Get the ML Workspace's Managed Identity Principal ID
ML_IDENTITY_ID=$(az ml workspace show --name $ML_WORKSPACE_NAME --resource-
group $RESOURCE_GROUP --query "identity.principalId" -o tsv)

# Grant ML Workspace Managed Identity access to Key Vault secrets
az keyvault set-policy \
    --name $KEY_VAULT_NAME \
    --resource-group $RESOURCE_GROUP \
    --object-id $ML_IDENTITY_ID \
    --secret-permissions get list

# Deploy Private Endpoints for Storage and Key Vault (if not automatically
handled by the workspace creation)
# In a full VNet deployment, you would create private endpoints for the
storage account, key vault, and ACR.
# Example for Key Vault:
# az network private-endpoint create \
#     --name pe-kv-${PROJECT_ID} \
#     --resource-group $RESOURCE_GROUP \
#     --vnet-name $VNET_NAME \
#     --subnet $SUBNET_ML \
#     --private-connection-resource-id $(az keyvault show -n $KEY_VAULT_NAME
-g $RESOURCE_GROUP --query id -o tsv) \
#     --group-ids vault \
#     --connection-name kv-connection
```

6.6. Deploy Azure OpenAI Service with Content Filtering

Deploy the Azure OpenAI service and ensure content filtering is enabled by default.

```
export AOAI_NAME="aoai-${PROJECT_ID}"
az cognitiveservices account create \
    --name $AOAI_NAME \
    --resource-group $RESOURCE_GROUP \
    --location $LOCATION \
    --kind OpenAI \
    --sku S0 \
    --custom-domain "${AOAI_NAME}"

# Deploy a model (e.g., gpt-4)
az cognitiveservices account deployment create \
    --name $AOAI_NAME \
    --resource-group $RESOURCE_GROUP \
    --deployment-name gpt-4-deployment \
    --model-name gpt-4 \
    --model-version "0613" \
    --model-format OpenAI \
    --sku-name "Standard" \
    --capacity 100
```

6.7. Configure Responsible AI Policy Enforcement

While the workspace is deployed, the MLOps pipeline must be configured to enforce GRC. This is often done via Azure Policy.

```
# Pseudo-code for Policy Assignment:
# 1. Define a custom policy (e.g., 'Require Model Card and Bias Report')
# 2. Assign the policy to the Resource Group, setting the effect to 'Deny'
or 'Audit'
echo "--- Azure Policy for Responsible AI Enforcement ---"
echo "Policy Name: Require-Responsible-AI-Reports"
echo "Policy Definition: Ensures that any model registered in the ML
Workspace Model Registry"
echo "                   must have associated artifacts for 'Model Card' and
'Fairness Report'."
echo "Assignment Scope: $RESOURCE_GROUP"
echo "Effect: Deny (to block non-compliant model registration)"
echo "-------------------------------------------------"
```

---

7. Validation & Testing

Validation ensures both the infrastructure and the MLOps process adhere to the security and Responsible AI requirements.

7.1. Infrastructure Validation

Verify the core security and network configurations.

Test	Command	Expected Result
HBI Workspace	`az ml workspace show --name $ML_WORKSPACE_NAME --resource-group $RESOURCE_GROUP --query "hbiWorkspace"`	`true`
Network Isolation	`az ml workspace show --name $ML_WORKSPACE_NAME --resource-group $RESOURCE_GROUP --query "publicNetworkAccess"`	`Disabled` or `null` (implying private access only)
Key Vault Access	Attempt to retrieve a secret using the ML Workspace's Managed Identity.	Success (if policy is set correctly)
VNet Connectivity	Run a simple training job on the ML Compute Cluster.	Job succeeds, demonstrating connectivity to ADLS Gen2 and Key Vault via private links.

7.2. Responsible AI Validation

The MLOps pipeline must include a dedicated validation step using the `responsibleai` SDK.

1. **Model Training and Logging:** The training script must log the model and its associated Responsible AI artifacts.

```
# Pseudo-code for model validation script
from responsibleai import RAIInsights
from azureml.core import Run


run = Run.get_context()
# ... load data and model ...

# 1. Initialize RAIInsights
rai_insights = RAIInsights(
    model=model,
    train=train_data,
    test=test_data,
    target_feature='target',
    task_type='classification',
    # Define sensitive features for fairness assessment
    sensitive_features=['gender', 'age_group']
)

# 2. Add Responsible AI components
rai_insights.responsible_ai_tools.fairness.add(
    metric='selection_rate',
    group_feature='gender'
)
rai_insights.responsible_ai_tools.error_analysis.add()
rai_insights.responsible_ai_tools.interpretability.add()

# 3. Compute and Upload
rai_insights.compute()
rai_insights.upload_to_azureml(run)


print("Responsible AI components integrated and validated. Reports
uploaded to Azure ML.")
```

2. **Model Card Generation:** A Model Card (a standardized document detailing the model's purpose, limitations, and performance) must be generated and attached to the registered model. This is the primary evidence for transparency and human oversight.

3. **Content Moderation Test (for LLMs):** If Azure OpenAI is used, test the content filtering policy by submitting a prompt that violates the policy. The expected result is a blocked response and a log entry in Azure Monitor.

8. Troubleshooting

Common issues in a highly secure, VNet-isolated MLOps environment often revolve around networking and identity.

Issue	Potential Cause	Resolution
Deployment Fails (403 Forbidden)	Service Principal lacks necessary permissions on the Resource Group or Key Vault.	Verify the Service Principal has `Contributor` role on the Resource Group. Ensure the SP has `Get/List` permissions on Key Vault secrets/keys.
Model Training Fails (Cannot Access Data)	Compute cluster cannot access ADLS Gen2.	Check VNet and Private Endpoint configuration. Ensure the ML Workspace's Managed Identity has `Storage Blob Data Contributor` role on the ADLS Gen2 account.
Inference Endpoint Error (500 Internal)	Missing secrets (e.g., OpenAI key) in Key Vault, or the endpoint's Managed Identity lacks access.	Ensure the endpoint's Managed Identity has `Get` permission on the Key Vault secrets. Verify the secret name in the deployment configuration.
ML Studio Portal is Inaccessible	The ML Workspace is deployed with private access only (`allow-public-access-when-behind-vnet: false`).	Access the ML Studio via a jump box or a client machine connected to the VNet via Azure VPN Gateway or Azure Bastion.
Model Registration Blocked	Azure Policy is enforcing the Responsible AI requirement.	Ensure the MLOps pipeline successfully generated and attached the Model Card and Bias Report artifacts before attempting model registration.
`responsibleai` **SDK Fails**	Missing dependencies in the Conda environment.	Verify the `environment.yml` file includes `responsibleai` and its dependencies (e.g., `scikit-learn` , `pandas`).

9. Cost Optimization

Optimizing costs in an MLOps environment is crucial, especially with high-performance compute and premium services like Azure OpenAI.

9.1. Compute Scaling and Efficiency

- **Low-Priority VMs:** Configure Azure ML Compute Clusters to use **low-priority VMs** for non-critical or non-time-sensitive training jobs. This can reduce compute costs by up to 80%.
- **Auto-Shutdown Policies:** Implement aggressive **auto-shutdown policies** for compute clusters. Configure the cluster to scale to zero nodes (`min_nodes=0`) after a short period of inactivity (e.g., 15 minutes).
- **Instance Type Selection:** Right-size the compute instance. Use GPU-enabled VMs only when necessary for deep learning, and prefer cost-effective CPU VMs for data preparation and light training.

9.2. Storage Tiering and Lifecycle Management

- **ADLS Gen2 Tiering:** Utilize the tiered storage capabilities of ADLS Gen2. Move older model artifacts, raw training data, and historical logs to **Cool** or **Archive** storage tiers using **Lifecycle Management Policies**.
- **Artifact Retention:** Implement a strict retention policy for model artifacts and run logs. Delete artifacts for models that have been retired or superseded after a defined compliance period (e.g., 5 years).

9.3. Inference Endpoint Scaling

- **Aggressive Auto-Scaling:** For inference endpoints (AKS or Managed Endpoints), implement aggressive auto-scaling policies. Configure the minimum number of instances to zero (`min_instances=0`) to ensure no cost is incurred during idle periods.
- **Serverless Endpoints:** For models with infrequent or bursty traffic, consider using **Azure ML Serverless Endpoints**, which offer consumption-based pricing and automatically scale to zero.

- **Azure OpenAI Consumption:** Monitor Azure OpenAI usage closely. Utilize features like provisioned throughput only when sustained, high-volume traffic is guaranteed; otherwise, rely on pay-as-you-go consumption.

---

10. Security Best Practices

The security posture of this MLOps solution is built on a defense-in-depth strategy, focusing on network, identity, and data protection.

10.1. Network Isolation (Zero Trust)

- **Azure Private Link:** Use Azure Private Link for all core services (Azure ML Workspace, ADLS Gen2, Key Vault, Azure Container Registry, Azure OpenAI). This ensures all traffic remains on the Azure backbone network, eliminating exposure to the public internet.
- **VNet Integration:** Deploy the Azure ML Compute Cluster and inference endpoints (AKS/Managed Endpoint) directly into a dedicated VNet subnet.
- **NSG and Firewall Rules:** Apply Network Security Groups (NSGs) to subnets to restrict inbound and outbound traffic to only the necessary service tags and ports.

10.2. Data Encryption and High Business Impact (HBI)

- **Encryption at Rest and In Transit:** Enforce TLS 1.2+ for all data in transit. Ensure all data at rest in ADLS Gen2 and Key Vault is encrypted.
- **Customer-Managed Keys (CMK):** For the highest level of compliance (e.g., HIPAA), configure the Azure ML Workspace and its associated storage accounts to use **Customer-Managed Keys** stored in Azure Key Vault.
- **HBI Workspace:** The `hbi-workspace: true` setting is critical. It enforces stricter data handling, such as disabling diagnostic data collection and ensuring all data is encrypted.

10.3. Least Privilege and Managed Identities

- **Role-Based Access Control (RBAC):** Implement the principle of least privilege. Grant users and Service Principals only the minimum necessary RBAC roles.
    - *Data Scientists:* Reader on the Resource Group, specific roles for ML Experimentation.
    - *MLOps Engineers:* Contributor on the Resource Group, specific roles for MLOps deployment.
- **Managed Identities:** Use **System-Assigned Managed Identities** for all service-to-service communication (e.g., ML Workspace accessing Key Vault, Compute Cluster accessing ADLS Gen2). This eliminates the need to manage and rotate credentials.

10.4. Content Moderation and Audit Trails

- **Generative AI Moderation:** Enable and configure the built-in content filtering policies on Azure OpenAI and Cognitive Services. This is a mandatory control to prevent the generation of harmful, illegal, or inappropriate content.
- **Comprehensive Audit & Logging:** Enable Azure Monitor and Log Analytics for all resources. Configure diagnostic settings to capture:
    - **AI Interaction Logs:** Every request and response to the deployed model endpoints.
    - **Control Plane Logs:** All management operations (e.g., model registration, deployment changes).
    - **Access Logs:** All successful and failed access attempts to data and secrets.

Cleanup

To remove all deployed resources and avoid incurring further costs, execute the following command. **Warning:** This action is irreversible.

```
az group delete --name $RESOURCE_GROUP --yes --no-wait
```

References

[1] IBM. (2023). *Cost of a Data Breach Report 2023*. Retrieved from https://www.ibm.com/security/data-breach