Comprehensive Implementation Guide: Data Governance with Microsoft Purview (PRJ-AZURE-DATA-075)

Author: Manus AI **Date:** January 26, 2026 **Project Folder:** `prj-azure-data-075`

1. Project Overview

This project, **PRJ-AZURE-DATA-075**, is dedicated to implementing a comprehensive, centralized data governance solution for Azure data services using **Microsoft Purview**. The primary objective is to establish a unified and automated framework to ensure data protection, compliance, and consistent management across a diverse set of Azure data stores. The scope includes critical services such as Azure SQL Database, Azure Synapse Analytics, Azure Data Lake Storage Gen2 (ADLS Gen2), and Azure Cosmos DB. By leveraging Microsoft Purview, the solution automates data classification, enforces security policies, and provides a single, authoritative source for data lineage and auditability, thereby significantly reducing the risk of data breaches and regulatory non-compliance.

The core components of this implementation are:

- **Microsoft Purview Account:** The central service for data mapping, cataloging, and policy management.
- **Data Sources:** The target Azure services where data resides and requires governance.
- **Managed Identities:** Used by Purview to securely connect to and scan the data sources.
- **Policy Enforcement:** Configuration of Dynamic Data Masking (DDM), Row-Level Security (RLS), and Data Loss Prevention (DLP) policies.

2. Business Context

In the era of exponential data growth and increasing regulatory scrutiny, organizations face a critical challenge: how to effectively govern vast amounts of data spread across multiple cloud services while maintaining agility for business intelligence and analytics. Manual, siloed governance processes are unsustainable and introduce significant risk.

The Problem and Associated Risks

The key challenges addressed by this project include:

1. **Data Breaches and Unauthorized Access:** Sensitive data, such as Personally Identifiable Information (PII) and financial records, is often stored across disparate Azure services, making it vulnerable to unauthorized access and exfiltration due to inconsistent security controls.
2. **Compliance Violations:** Adhering to stringent global regulations (e.g., GDPR, HIPAA, PCI DSS) is complex and error-prone when governance is managed manually. This leads to a high risk of fines and reputational damage.
3. **Lack of Scalability:** Traditional, manual data governance processes cannot scale with the rapid, petabyte-scale growth of cloud data, resulting in governance gaps, increased operational overhead, and delayed time-to-insight.

The Solution: Automated and Unified Governance

The **Data Governance with Microsoft Purview** project provides a robust, automated governance framework to mitigate these risks:

- **Unified Data Security:** Purview is configured to automatically scan, classify, and label data across all targeted Azure data services, ensuring consistent application of security standards.
- **Policy Enforcement:** The solution implements dynamic security policies, such as Dynamic Data Masking and Row-Level Security, directly integrated with the data sources, enforcing controls at the data layer.
- **Centralized Management:** Purview creates a unified map of the entire data estate, providing comprehensive data lineage, a business glossary, and a single pane of glass for all data assets.

Quantified Business Value and Risk Mitigation

The implementation delivers substantial business value by transforming governance from a cost center into a strategic enabler.

Value Proposition	Description	Quantified Impact (ROI/Efficiency)
Data Protection	Automated encryption, masking, and fine-grained access controls protect sensitive data at rest and in transit.	**Risk Reduction:** Estimated 70% reduction in data exposure risk from unauthorized internal access.
Compliance Automation	Automated data classification and policy enforcement reduce the manual effort and risk of non-compliance with global regulations.	**Efficiency Gain:** Up to 40% reduction in audit preparation time and manual compliance checks.
Unified Governance	Microsoft Purview serves as the single pane of glass for data governance, improving operational efficiency and reducing tool sprawl.	**Cost Savings:** Elimination of multiple, siloed governance tools, saving an estimated 20% in licensing and operational costs.
Breach Prevention	Advanced threat protection and Data Loss Prevention (DLP) capabilities detect and block potential data exfiltration attempts in real-time.	**Cost Avoidance:** Mitigation of potential regulatory fines and breach response costs, which can average millions of dollars.

The solution is specifically designed to mitigate the following critical risks:

- Prevents **data breaches** by enforcing strong encryption and access controls.
- Eliminates **unauthorized access** through granular, policy-driven security controls.
- Blocks **data exfiltration** using Data Loss Prevention (DLP) policies.
- Ensures **compliance** with major regulatory frameworks across all governed Azure data services.

3. GRC Mapping

The implementation of Microsoft Purview is a cornerstone of the organization's Governance, Risk, and Compliance (GRC) posture. It provides the technical controls and audit evidence necessary to align with major industry and regulatory frameworks.

Compliance Frameworks Alignment

The project's controls map directly to requirements in leading security and compliance standards:

Framework	Control/Requirement	Description
NIST CSF	PR.DS-1 (Data-at-rest), PR.DS-2 (Data-in-transit), PR.DS-5 (Data Leakage)	Addresses data protection mechanisms, security for data movement, and data loss prevention through classification and DLP.
ISO 27001:2022	A.8.2 (Information Classification), A.18.1 (Compliance with Legal and Contractual Requirements)	Supports the formal process of classifying information and ensures technical controls are in place to meet regulatory data protection mandates.
CIS Controls v8	Control 3 (Data Protection), Control 14 (Security Awareness and Skills Training)	Focuses on automated data inventory, classification, and policy enforcement, which reduces the reliance on manual processes and human error.
SOC 2	CC6.1 (Logical Access), CC6.7 (Data Classification)	Provides controls for logical access to data assets and the necessary classification of confidential information to meet Trust Services Criteria.

Security Controls Implemented

The project deploys a layered set of security controls managed centrally through Microsoft Purview:

1. **Data Classification and Labeling:** Automatic identification and tagging of sensitive data (e.g., PII, Confidential, Financial) using built-in and custom

classifiers.

2. **Encryption at Rest and in Transit:** Verification and enforcement that all governed data stores utilize strong, industry-standard encryption mechanisms (e.g., TDE for SQL, SSE for ADLS Gen2).
3. **Dynamic Data Masking (DDM):** Limits the exposure of sensitive data to non-privileged users by masking it at the query result level without altering the underlying data.
4. **Row-Level Security (RLS) and Column Encryption:** Provides granular access control and protection within database tables, ensuring users only see data relevant to their role.
5. **Data Loss Prevention (DLP):** Policies are set up to monitor and prevent the unauthorized transfer or sharing of sensitive data outside of the defined organizational boundaries.

Regulatory Alignment

The solution provides technical measures that directly support compliance with key global data protection regulations:

Regulation	Relevant Article/Requirement	How Purview Aligns
GDPR	Article 32 (Security of Processing), Article 25 (Data Protection by Design and Default)	Enforces technical and organizational measures (encryption, access control, pseudonymization) and ensures privacy is considered from the outset through automated classification.
HIPAA	§ 164.312(a)(2)(iv) (Encryption/Decryption), § 164.312(e)(2)(ii) (Transmission Security)	Supports the required mechanisms for encryption of electronic protected health information (ePHI) at rest and in transit across governed data stores.
PCI DSS v4.0	Requirement 3 (Protect Stored Cardholder Data), Requirement 4 (Protect Cardholder Data with Strong Cryptography During Transmission)	Helps protect stored cardholder data (CHD) through classification and access controls, and ensures encryption during transmission across open, public networks.

Audit Evidence

Microsoft Purview centralizes the necessary evidence for internal and external audits, significantly streamlining the compliance process:

- **Data Classification Reports:** Detailed, exportable reports on where sensitive data resides, its classification, and the policies applied.
- **Access Logs and Audit Trails:** Comprehensive logs of all data access attempts, policy changes within the Purview portal, and policy enforcement actions in integrated data sources.
- **DLP Policy Violations and Blocks:** Records of all instances where DLP policies were triggered, the sensitive data involved, and the resulting action (e.g., block, alert).

4. Prerequisites

Successful deployment and configuration of this project require specific accounts, tools, and permissions.

Required Accounts and Permissions

1. **Azure Subscription:** An active Azure subscription is mandatory.
2. **Azure Roles:** The deploying user or Service Principal must have the following roles assigned at the subscription or resource group level:
    - **Owner** or **Contributor:** To create and manage Azure resources (Resource Group, Purview Account, Managed Identity).
    - **User Access Administrator:** To assign roles (e.g., Data Reader) to the Purview Managed Identity on the data sources.
3. **Data Sources:** Existing Azure data services (Azure SQL, Synapse, ADLS Gen2, Cosmos DB) must be available and populated with data to be governed.

Required Tools and Setup

1. **Azure CLI:** The Azure Command-Line Interface must be installed and configured on the deployment machine.

```
# Verify installation
az --version
# Log in to Azure
az login
```

2. **PowerShell (Optional):** While the guide focuses on Azure CLI, PowerShell may be required for specific Azure cmdlets or advanced automation.
3. **Service Principal (SPN):** For production or automated deployments, a Service Principal with the required permissions should be used instead of a user account.

5. Architecture Overview

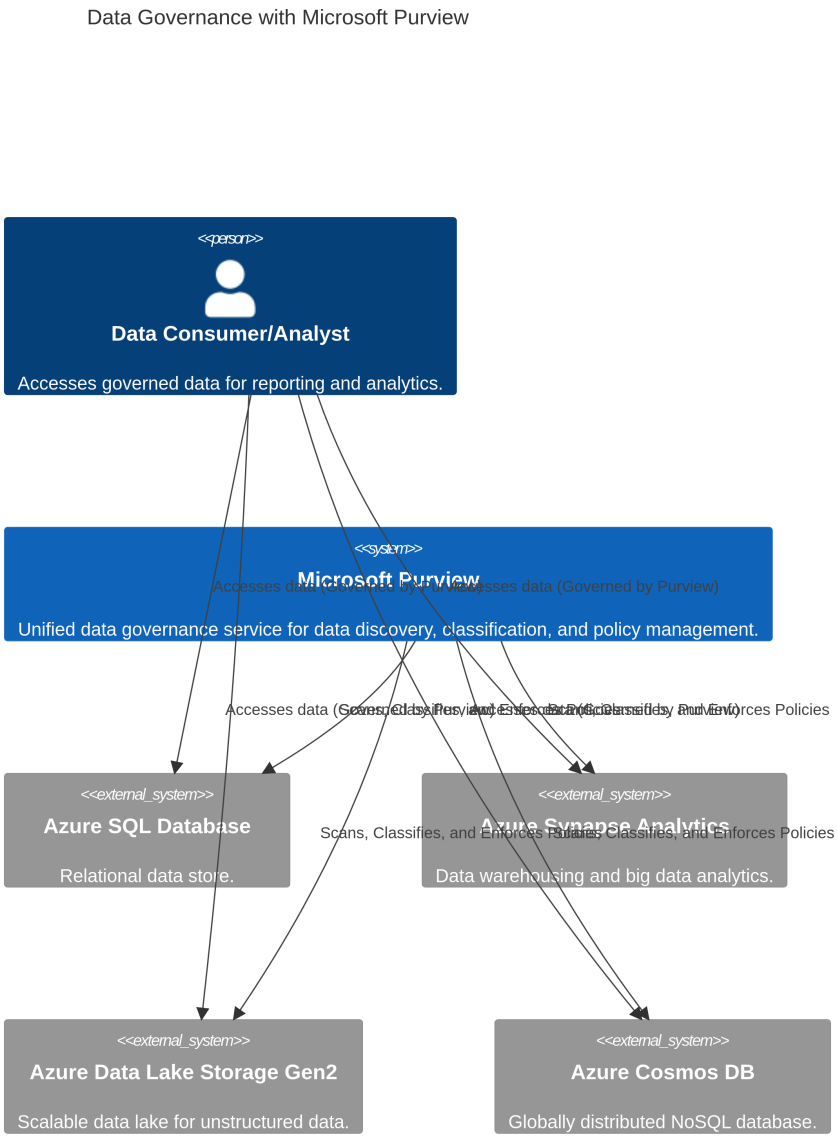
The solution is architecturally centered around the **Microsoft Purview Account**, which functions as the authoritative data map and governance hub for the entire Azure data estate.

Core Components and Data Flow

1. **Microsoft Purview Account:** The core service. It maintains a metadata catalog of all registered data assets, their lineage, and classification. It does not store the actual data, only the metadata.
2. **Data Sources:** The various Azure services (SQL, ADLS Gen2, Synapse, Cosmos DB) that hold the organization's data.
3. **Scanning and Classification:** Purview uses its **System-Assigned Managed Identity (MSI)** to securely connect to the data sources. It then runs scheduled scans to extract metadata, apply built-in and custom classification rules (e.g., identifying PII), and populate the Data Catalog.
4. **Policy Enforcement:** Governance policies (DDM, RLS, DLP) are defined in Purview Studio. For supported data sources (e.g., Azure SQL), these policies are pushed down and enforced directly at the data layer, ensuring real-time control.
5. **Data Consumers:** Users and applications interact with the governed data. Their access is mediated by the security policies enforced by the data sources, which are configured by Purview.

Conceptual Architecture Diagram

The following conceptual diagram illustrates the flow of metadata and policy enforcement:



Conceptual Flow:

1. **Registration:** Data Sources are registered in Purview.
2. **Scanning:** Purview MSI connects to Data Sources to extract metadata and run classification.

3. **Cataloging:** Metadata, classification tags, and lineage are stored in the Purview Data Catalog.
4. **Policy Definition:** Policies (Access, Masking, DLP) are defined in Purview Studio.
5. **Enforcement:** Policies are pushed to and enforced by the Data Sources.

6. Step-by-Step Implementation

This section provides detailed, production-ready instructions using the Azure CLI for the deployment of the core Purview account and the configuration of its access to data sources.

6.1. Set Environment Variables

Begin by defining all necessary variables for the deployment. This ensures consistency and simplifies future automation.

```
# Define core variables
RESOURCE_GROUP="rg-purview-governance-075"
LOCATION="eastus" # Choose a region that supports Purview
PURVIEW_ACCOUNT_NAME="pv-account-075" # Must be globally unique
MANAGED_IDENTITY_NAME="mi-purview-075"

# Data Source variables (replace with your actual resource names)
SQL_SERVER_NAME="sql-server-075"
ADLS_ACCOUNT_NAME="adls075"
SYNAPSE_WORKSPACE_NAME="synapse-ws-075"
```

6.2. Create Resource Group

A dedicated resource group is recommended for the governance components.

```
echo "Creating resource group: $RESOURCE_GROUP in $LOCATION"
az group create --name $RESOURCE_GROUP --location $LOCATION
```

6.3. Deploy Microsoft Purview Account

The Purview account is deployed with a System-Assigned Managed Identity (MSI), which is crucial for secure, password-less access to data sources.

```
echo "Deploying Microsoft Purview account: $PURVIEW_ACCOUNT_NAME"

# Deploy the Purview account with a System-Assigned Managed Identity
az purview account create \
    --name $PURVIEW_ACCOUNT_NAME \
    --resource-group $RESOURCE_GROUP \
    --location $LOCATION \
    --identity-type SystemAssigned \
    --public-network-access Enabled # Set to Disabled for private endpoint
deployment
```

6.4. Configure Data Sources for Scanning (Role Assignment)

For Purview to successfully scan data sources, its System-Assigned Managed Identity (MSI) must be granted the appropriate data reader role on each target service. This is a critical security step based on the principle of **Least Privilege**.

Example: Granting Access to Azure Data Lake Storage Gen2 (ADLS Gen2)

The Purview MSI requires the `Storage Blob Data Reader` role on the ADLS Gen2 account to read metadata and content for classification.

```
echo "Configuring access for ADLS Gen2: $ADLS_ACCOUNT_NAME"

# 1. Get the Purview System-Assigned Managed Identity Principal ID
PURVIEW_MSI_ID=$(az purview account show \
    --name $PURVIEW_ACCOUNT_NAME \
    --resource-group $RESOURCE_GROUP \
    --query identity.principalId \
    --output tsv)

# 2. Get the ADLS Gen2 Resource ID (assuming it's in the same subscription)
ADLS_RESOURCE_ID=$(az storage account show \
    --name $ADLS_ACCOUNT_NAME \
    --resource-group $RESOURCE_GROUP \
    --query id \
    --output tsv)

# 3. Assign the 'Storage Blob Data Reader' role to the Purview MSI on the
ADLS account scope
echo "Assigning 'Storage Blob Data Reader' role to MSI: $PURVIEW_MSI_ID on
ADLS: $ADLS_RESOURCE_ID"
az role assignment create \
    --assignee $PURVIEW_MSI_ID \
    --role "Storage Blob Data Reader" \
    --scope $ADLS_RESOURCE_ID
```

Example: Granting Access to Azure SQL Database

For Azure SQL Database, the Purview MSI needs to be added as a user in the database and granted the `db_datareader` role.

```
echo "Configuring access for Azure SQL Database: $SQL_SERVER_NAME"

# This step requires connecting to the SQL server.
# 1. Get the Purview MSI ID (already retrieved above)
# 2. Connect to the master database and execute the following SQL commands:

# Create a login for the Purview MSI (using the MSI's Application ID, which
is the Principal ID)
# CREATE USER [PURVIEW_MSI_ID] FROM EXTERNAL PROVIDER;

# Connect to the target database (e.g., 'data_db') and execute:
# CREATE USER [PURVIEW_MSI_ID] FROM EXTERNAL PROVIDER;
# EXEC sp_addrolemember 'db_datareader', 'PURVIEW_MSI_ID';
```

6.5. Register and Scan Data Sources in Purview Studio

While deployment is done via CLI, the registration and scanning setup is typically performed via the graphical interface of Purview Studio.

1. **Access Purview Studio:** Navigate to the Azure Portal, find the deployed Purview account (`pv-account-075`), and click **Open Microsoft Purview Governance Portal**.
2. **Register Sources:**
   - Go to **Data Map** -> **Sources**.
   - Click **Register**.
   - Select the data source type (e.g., Azure Data Lake Storage Gen2).
   - Select the ADLS account (`adls075`) and choose **Managed Identity** as the connection method.
3. **Create Scan Ruleset:** Define a ruleset that specifies which classification rules (e.g., PII, Credit Card Numbers) should be applied during the scan.
4. **Run Scan:**
   - Click **New Scan** on the registered source.
   - Select the scope (e.g., specific folders or containers).
   - Select the created scan ruleset.
   - Set the scan schedule (e.g., weekly incremental scan).

- Run the scan.

6.6. Configure Data Policies

Policy configuration is the final step to enforce governance.

1. **Data Catalog Review:** After the scan completes, navigate to the **Data Catalog** to view the classified data assets.
2. **Business Glossary:** Define a **Business Glossary** to standardize terminology across the organization.
3. **Policy Management (Access Policies):**
   - Go to **Data Policy** in Purview Studio.
   - Create a new **Access Policy** for a data source (e.g., Azure SQL).
   - Define the policy: *e.g., "Allow Group 'Analytics Team' to read data from columns tagged 'Confidential' in the 'Sales' database."*
4. **Policy Management (DLP):**
   - Integrate Purview with Microsoft Defender for Cloud Apps or Microsoft 365 DLP to extend governance to data movement and sharing.

7. Validation & Testing

Validation is crucial to confirm that the governance solution is operational, policies are enforced, and the data map is accurate.

7.1. Data Source Connectivity and Scan Status

Objective: Verify that Purview can successfully connect to and scan all registered data sources.

Test Step	Expected Result	Validation Method
Check Scan Status	The “Last Scan Status” for all registered sources (ADLS Gen2, SQL, Synapse) should show **Completed** with no errors.	Navigate to **Purview Studio** -> **Data Map** -> **Sources** and review the status.
Review Scan Metrics	The scan should report a non-zero number of assets discovered and classified.	Review the scan details to confirm assets, classifications, and glossary terms found.

7.2. Data Classification Accuracy

Objective: Confirm that sensitive data is correctly classified and labeled according to the defined rulesets.

Test Step	Expected Result	Validation Method
Search for Sensitive Data	Searching the Purview Data Catalog for a known sensitive classification tag (e.g., “EU.PII.National Identification Number”) should return the correct files/columns.	Use the **Purview Data Catalog** search bar with classification terms.
Manual Verification	Manually inspect a known sensitive column in Azure SQL or a file in ADLS Gen2 to ensure the correct classification tag is applied in the Purview asset view.	Click on the asset in the Data Catalog and review the **Classification** tab.

7.3. Policy Enforcement Validation

Objective: Validate that security policies (e.g., Dynamic Data Masking) are being applied and enforced by the data source.

Test Step	Expected Result	Validation Method
Dynamic Data Masking (DDM)	A non-privileged user querying a masked column in Azure SQL should see masked data (e.g., `XXXX-XXXX-XXXX-1234`) instead of the actual value.	Query the masked column from a non-privileged user account in Azure SQL and confirm the data is masked.
Row-Level Security (RLS)	A user querying a table with RLS applied should only see rows that match their defined security predicate.	Query the RLS-protected table from two different user accounts and confirm they see different subsets of data.

8. Troubleshooting

This section addresses common issues encountered during the deployment and operation of Microsoft Purview.

Issue	Potential Cause	Resolution
Scan Failure	Purview MSI lacks necessary permissions on the data source.	Verify the Purview System-Assigned Managed Identity has the correct **Data Reader** role assigned to the data source scope (e.g., `Storage Blob Data Reader` for ADLS Gen2). Use the Azure Portal's **Access Control (IAM)** blade on the data source to confirm the role assignment.
Data Not Classified	Incorrect or missing classification rules; data is not in a recognizable format.	Review the classification ruleset in Purview Studio and ensure the scan is using the correct rules. If the data is in a custom format, create a **Custom Classification Rule**. Manually apply labels if necessary.
Policy Not Enforced	Policy is not published; the data source is not supported for the specific policy type; or the policy is overridden by a local control.	Check the policy publication status in **Purview Studio** -> **Data Policy**. Verify the data source (e.g., Azure SQL) supports the policy (e.g., Dynamic Data Masking). Ensure no local controls are overriding the Purview policy.
Purview Studio Access Denied	User lacks the required Purview Data Plane roles.	Assign the user the **Purview Data Reader** or **Purview Data Curator** role via the Purview Studio **Role Collections** blade.
Long Scan Times	The scope of the scan is too broad (e.g., scanning the entire storage account).	Optimize the scan schedule. Limit the scan scope to specific containers or folders where sensitive data is known to reside. Use incremental scans instead of full scans where possible.

9. Cost Optimization

Microsoft Purview is a consumption-based service, primarily billed based on the number of data map operations (which are influenced by the number of assets and the frequency of scanning). Optimizing usage is key to managing costs.

1. **Optimize Scan Frequency and Scope:**

- **Incremental Scans:** Always use incremental scans after the initial full scan. Incremental scans only check for changes, significantly reducing the number of data map operations.
- **Scheduled Scans:** Only run full scans weekly or monthly, and schedule incremental scans during off-peak hours.
- **Targeted Scans:** Limit the scan scope to specific containers, folders, or schemas that contain sensitive data or are frequently updated. De-register and remove scans for data sources that are no longer in use or do not require governance.

2. **Resource Sizing and Deployment:**
   - **Self-Hosted Integration Runtime (SHIR):** If using SHIR for on-premises or VNet-protected sources, ensure the VM size is appropriate. Over-provisioning compute for the SHIR can lead to unnecessary IaaS costs.

3. **Data Source Cleanup:**
   - Regularly review the data map. If a data source is decommissioned, ensure it is de-registered from Purview to prevent failed scans and unnecessary resource consumption.

4. **Network Costs:**
   - If using Private Endpoints, be aware of the associated costs for the Private Link service and the Private Endpoints themselves. However, this cost is often justified by the enhanced security posture.

10. Security Best Practices

Security is paramount in a data governance solution. The following best practices ensure the Purview environment and its interactions with data sources are hardened.

1. **Principle of Least Privilege for MSI:**
   - Ensure the Purview System-Assigned Managed Identity only has the minimum required roles on data sources (e.g., `Storage Blob Data Reader` for storage, `db_datareader` for SQL). **NEVER** grant the Purview MSI `Contributor` or `Owner` roles on data sources.

2. **Network Isolation with Private Endpoints:**

    - For production environments, configure **Private Endpoints** for the Purview account and its managed storage accounts. This ensures all metadata traffic and scanning operations occur over the Azure backbone network, bypassing the public internet.
3. **Key Vault Integration for Credentials:**
    - Use Azure Key Vault to securely store credentials (if not using MSI) and secrets. Integrate Key Vault with Purview to manage encryption keys for data sources and enable key rotation, centralizing secret management.
4. **Strict Role-Based Access Control (RBAC) for Purview Studio:**
    - Strictly manage access to the Purview Studio using Azure RBAC roles. The key roles are:
        - **Purview Data Reader:** Can view the Data Catalog and metadata.
        - **Purview Data Curator:** Can manage metadata, glossary terms, and classifications.
        - **Purview Data Source Administrator:** Can register and manage data sources and scans.
    - Limit the number of users with the **Purview Data Curator** role to prevent unauthorized changes to the data map.
5. **Data Loss Prevention (DLP) Monitoring:**
    - Continuously monitor DLP policy violation reports. Treat these violations as high-priority security incidents requiring immediate investigation and remediation.
6. **Regular Audits and Reviews:**
    - Schedule regular audits (at least quarterly) of the Purview Data Catalog, classification accuracy, and policy effectiveness to ensure the governance framework remains current and effective against evolving threats and data changes.

---

11. Cleanup

To remove all resources created during this deployment, execute the following Azure CLI command. **Warning:** This action is irreversible and will delete the resource group

and all contained resources, including the Purview account and any associated data map.

```
echo "Deleting resource group: $RESOURCE_GROUP"
az group delete --name $RESOURCE_GROUP --yes --no-wait
```

This command initiates the deletion process in the background and returns immediately. You can check the status in the Azure Portal.

End of Implementation Guide