

**Investment Domain Modernization: Data Ingress, Egress, Storage, and Metrics in AWS**

**Document Versioning**

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Date | Author | Description |
| 1.0 | 2025-07-08 | [Your Name] | Initial Draft |

**Table of Contents**

1. Overview
2. Data Storage Patterns
3. Data Ingress Patterns
4. Data Egress Patterns
5. Data Consumption Patterns
6. Metrics Logging Approaches
7. Pros, Cons, Limitations, Security Considerations
8. Appendix

**1. Overview**

This document outlines the possible data ingress and egress patterns, storage options, and metrics logging methods for the Investment Domain Modernization project on AWS. It includes a comparative analysis of each approach, focusing on pros, cons, limitations, and security considerations.

**2. Data Storage Patterns**

|  |  |  |
| --- | --- | --- |
| Storage Option | Description | Use Cases |
| Amazon S3 (Iceberg) | Object storage with Iceberg table format | Data lake, scalable storage |
| Amazon Redshift | Managed data warehouse for analytics | Complex analytics, reporting |

**3. Data Ingress Patterns**

**Pull Methods**

* **API Calls:** Scheduled or on-demand retrieval from external/internal APIs.
* **Scheduled DB Calls:** Periodic extraction from databases.
* **Scheduled SCP/SFTP Calls:** Automated file transfers from remote servers.
* **SharePoint/Confluence:** Manual or automated document ingestion.
* **JDBC/ODBC Calls:** Direct database connectivity for data extraction.

**Push Methods**

* **SFTP via AWS Transfer Family:** Secure file transfer into S3.
* **API Calls (Push):** External systems send data via REST APIs.
* **AWS CLI:** Manual uploads (not recommended for production).
* **NAS via AWS DataSync:** Automated file transfer from network storage.

**4. Data Egress Patterns**

**Push Methods**

* **SSH/SCP:** Secure file transfer to external servers.
* **SFTP via AWS Transfer Family:** Data delivery to external SFTP endpoints.
* **NAS via AWS DataSync:** Sync data to on-premises or cloud NAS.
* **SharePoint:** Exporting data to document management platforms.

**Pull Methods**

* **JDBC/ODBC:** Applications connect to Athena/Redshift for querying.
* **API Calls:** Data consumers retrieve data via APIs.
* **Non-AWS Services:** S3 Shortcuts on MS Fabric, IAM user credentials for access.

**5. Data Consumption Patterns**

* **Amazon Athena:** Serverless querying of S3 data.
* **Amazon Redshift:** Data warehouse analytics.
* **Snowflake:** Cloud data platform querying S3.
* **SharePoint/Confluence:** Data sharing and collaboration.
* **JDBC/ODBC:** BI tools and applications.
* **MS Fabric S3 Shortcuts:** Non-AWS analytics platforms.
* **IAM User Credentials:** Direct access for authorized users.

**6. Metrics Logging Approaches**

|  |  |  |
| --- | --- | --- |
| Method | Description | Typical Metrics |
| AWS CloudWatch | Native AWS service for logs and metrics | Ingestion rates, errors |
| AWS CloudTrail | API call and access logging | Access patterns, audit trail |
| S3 Access Logs | Detailed logs of S3 object access | File access, user activity |
| Redshift/Athena Logs | Query and performance logs | Query times, failures |
| Custom Application Logs | Application-specific metrics | Business KPIs, custom events |

**7. Pros, Cons, Limitations, Security Considerations**

**Data Ingress**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Pros | Cons | Limitations | Security Considerations |
| API Calls | Real-time, flexible | Requires API management | Rate limits, error handling | Secure endpoints, IAM roles |
| Scheduled DB Calls | Reliable, automated | Latency, scheduling complexity | Source DB load | Encrypted connections, audit logs |
| SCP/SFTP | Secure, widely supported | Requires credentials management | Not real-time | Key/cert management, encryption |
| SharePoint/Confluence | User-friendly, integrates with MS tools | Manual steps, less automation | Versioning, file formats | Access control, audit trails |
| JDBC/ODBC | Direct DB access, standard protocols | Network/firewall dependencies | Source DB compatibility | Encrypted channels, least privilege |
| SFTP (Transfer Family) | Managed, scalable, secure | Setup required, cost | Throughput limits | IAM integration, logging |
| AWS CLI | Flexible, scriptable | Not recommended for prod, error-prone | Manual, not scalable | Key rotation, audit logs |
| NAS (DataSync) | High throughput, automated | Setup complexity, costs | Network dependencies | Encrypted transfer, IAM roles |

**Data Egress**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Pros | Cons | Limitations | Security Considerations |
| SSH/SCP | Secure, standard | Manual, not scalable | Network/firewall restrictions | Key management, audit logs |
| SFTP (Transfer Family) | Managed, scalable, secure | Setup required, cost | Throughput limits | IAM integration, logging |
| NAS (DataSync) | Automated, high throughput | Setup complexity, ongoing costs | Network dependencies | Encrypted transfer, IAM roles |
| SharePoint | User-friendly, integrates with MS tools | Manual, less automation | File size/format limits | Access control, audit trails |
| JDBC/ODBC | Standardized, supports BI tools | Network/firewall dependencies | Query performance, cost | Encrypted channels, least privilege |
| API Calls | Flexible, programmatic access | Requires API management | Rate limits, error handling | Secure endpoints, IAM roles |
| S3 Shortcuts (MS Fabric) | Easy integration for non-AWS users | Limited features, compatibility | Access management | IAM user policies, audit logs |
| IAM User Credentials | Direct access, flexible | Risk of credential leakage | Key rotation, access control | Strong password, MFA, logging |

**Data Storage**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Storage Option | Pros | Cons | Limitations | Security Considerations |
| S3 (Iceberg) | Scalable, cost-effective, open format | Requires schema management | Query performance (vs. warehouse) | Encryption, bucket policies |
| Redshift | High-performance analytics | Cost, scaling complexity | Storage/compute quotas | VPC, IAM, encryption |

**8. Appendix**

**A. Versioning Strategy**

* Use document versioning table at the start of the document.
* Store all major revisions in a central repository (e.g., SharePoint, Confluence, or version-controlled S3 bucket).

**B. Security Best Practices**

* Enforce least-privilege IAM policies.
* Enable encryption at rest and in transit.
* Monitor all access with AWS CloudTrail and S3 access logs.
* Regularly rotate credentials and enforce MFA.

**C. References**

* AWS Documentation (S3, Redshift, Athena, DataSync, Transfer Family)
* Security best practices for AWS services

**Note:** This structure is intended for a high-level technical document. You can expand each section with more details or diagrams as needed for your audience.