**Documento Técnico: Pipeline de Datos para Spaceflight News API**

**Control de Versiones y Aprobaciones**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Versión | Fecha | Cambios | Autor | Aprobado por |
| 1.0 | **15/10/2024** | **Versión inicial del documento** | **[Tu Nombre]** | **Jefe de Ingeniería** |
| 1.1 | **20/10/2024** | **Adición de endpoint /reports** | **Equipo Datos** | **CTO** |
| 1.2 | **25/10/2024** | **Inclusión de políticas de retención específicas** | **Auditoría TI** | **DPO** |

**1. Estimación de Volumen de Datos**

**1.1 Volumen Diario**

* **Máximo histórico observado**: 34 registros/día por endpoint (/articles, /blogs y /reports).
* **Tamaño por registro**: ≤ 2 KB.
* **Cálculos**:

Volumen diario = 34 registros/endpoint × 3 endpoints × 2 KB = 204 KB/día

**1.2 Proyección Anual**

204 KB/día × 365 días = 74.46 MB/año

**1.3 Implicaciones**

* **Almacenamiento en S3**:
  + Costo estimado: 0.023/GB−mes→∗∗≈0.023/*GB*−*mes*→∗∗≈0.0012/año\*\*.
* **Procesamiento en Athena**:
  + 49.64 MB escaneados por consulta → $0.00025/consulta.

**2. Estrategia de Almacenamiento y Búsqueda**

**2.1 Almacenamiento en S3**

* **Estructura de carpetas**:

Texto

Descripción generada automáticamente

* **Ventajas**:
  + Particionamiento nativo para consultas eficientes en Athena.
  + Bajo costo de almacenamiento (S3 Standard).

**2.2 Búsqueda con Amazon Athena**

* **Configuración**:
  + Tablas creadas por AWS Glue Crawler con esquema inferido automáticamente.
  + Sintaxis de consulta estándar SQL.

SELECT \* FROM spaceflight\_news.articles WHERE year = '2025' AND month = '01';

* **Justificación de Athena**:
  + **Costos**: $5/TB escaneado (ideal para volúmenes pequeños y consultas ocasionales).
  + **Mantenimiento**: Sin servidores ni índices que gestionar.

**3. Plan de Contingencia**

**3.1 Escenarios de Fallo y Mitigación**

|  |  |
| --- | --- |
| Escenario | Acción |
| Fallo en la API externa | - Reintentos en Lambda con backoff exponencial (3 intentos). |
| Timeout de Lambda | - Retorno de estado PARTIAL con URL next para continuar descargas. |
| Corrupción de datos en S3 | - Versionamiento de S3 activado para recuperar versiones anteriores. |
| Error en Glue Crawler | - Notificaciones via CloudWatch + re-ejecución manual desde Airflow. |

**3.2 Recuperación de Datos**

* **Mecanismo**:
  + Ejecución manual de Lambda con parámetro next para reprocesar lotes fallidos.
  + Uso de Airflow para monitorizar reintentos automáticos.

**4. Sistema de Monitoreo**

**4.1 Métricas Clave (CloudWatch)**

|  |  |  |
| --- | --- | --- |
| Métrica | Umbral de Alerta | Acción |
| Ejecuciones fallidas de Lambda | ≥ 1 en 5 minutos | Notificación al equipo vía SNS. |
| Tiempo de ejecución de Lambda | ≥ 800 segundos | Optimizar código o dividir en pasos. |
| Registros no procesados en S3 | Incremento ≥ 10% vs histórico | Verificar Glue Crawler y Lambda. |

**4.2 Orquestación con Airflow**

* **Implementación propuesta**:
  + **DAG diario** para:
    1. Ejecutar Lambda.
    2. Lanzar Glue Crawler al finalizar.
    3. Verificar métricas de CloudWatch.
  + **Opción de despliegue**:
    1. **AWS Managed Workflows for Apache Airflow (MWAA)**: Integración nativa con servicios AWS (recomendado).
    2. **Airflow local**: Menor costo, pero requiere gestión de infraestructura.

**6. Arquitectura de Airflow**

**6.1 Opciones de Despliegue**

|  |  |  |
| --- | --- | --- |
| Criterio | Airflow Local | AWS MWAA |
| Costo | $0 (infra propia) | ≈ $300/mes (entorno básico) |
| Integración AWS | Manual (roles IAM) | Nativa (VPC, CloudWatch) |
| Mantenimiento | Alto (servidores, updates) | Bajo (AWS gestiona la infra) |

**Recomendación**: Usar AWS MWAA si el presupuesto lo permite, para simplificar operaciones.