**Diagrama de Arquitectura - Spaceflight News Data Pipeline**

**Componentes AWS y Justificación**

**1. AWS Lambda**

**Función**:

* Extracción diaria de datos de la API externa con paginación y reintentos.  
  **Justificación**:
* Serverless: Costo cero cuando no se ejecuta.
* Escalado automático para futuros incrementos de volumen.

**2. Amazon S3**

**Función**:

* Almacenamiento duradero de JSON en estructura particionada:

s3://spaceflight-raw-data/[endpoint]/year=\*/month=\*/day=\*/\*.json

**Justificación**:

* Costo mínimo ($0.023/GB-mes) + 99.999999999% durabilidad.

**3. AWS Glue Crawler**

**Función**:

* Actualización incremental del Data Catalog con esquemas JSON.  
  **Justificación**:
* Automatiza la detección de esquemas sin código.

**4. Amazon Athena**

**Función**:

* Consultas SQL sobre datos en S3 usando el Data Catalog.  
  **Justificación**:
* Pago por consulta ($5/TB escaneado), ideal para análisis ad-hoc.

**5. Apache Airflow (MWAA)**

**Función**:

* Orquestación del pipeline y manejo de reintentos.  
  **Justificación**:
* Integración nativa con AWS (Lambda, Glue).

**Flujo de Datos**

[Insertar diagrama tipo: API Externa → Lambda → S3 → Glue → Athena]

**Pasos detallados**:

1. **Extracción**:
   * Lambda consulta /articles, /blogs, y /reports con paginación (100 registros/llamada).
   * *Ejemplo de flujo*:

API → Lambda (procesa página 1) → S3 → Lambda (procesa página 2) → ...

1. **Almacenamiento**:
   * JSONs se guardan en S3 con particionado temporal:

s3://spaceflight-raw-data/articles/year=2024/month=09/day=15/data.json

1. **Catalogación**:
   * Glue Crawler escanea S3 y actualiza metadatos en el Data Catalog.
2. **Análisis**:
   * Usuarios ejecutan consultas SQL en Athena:

SELECT COUNT(\*) FROM articles WHERE year = '2024';

**Sistema de Backup y Recuperación**

**Backup Automático**

[Insertar diagrama de S3 Versioning + Lifecycle Policies]

* **Estrategias**:
  + **Versionado en S3**:
    - Histórico de cambios para recuperar versiones corruptas.
  + **Lifecycle Policies**:
    - Transición a S3 Glacier después de 1 año (costo: $0.004/GB-mes).

**Recuperación de Desastres**

[Insertar diagrama de recuperación desde Glacier]

* **Escenarios**:

|  |  |
| --- | --- |
| Caso | Acción |
| Eliminación accidental | Restauración desde versiones anteriores. |
| Corrupción masiva | Reprocesamiento desde la API usando Lambda |