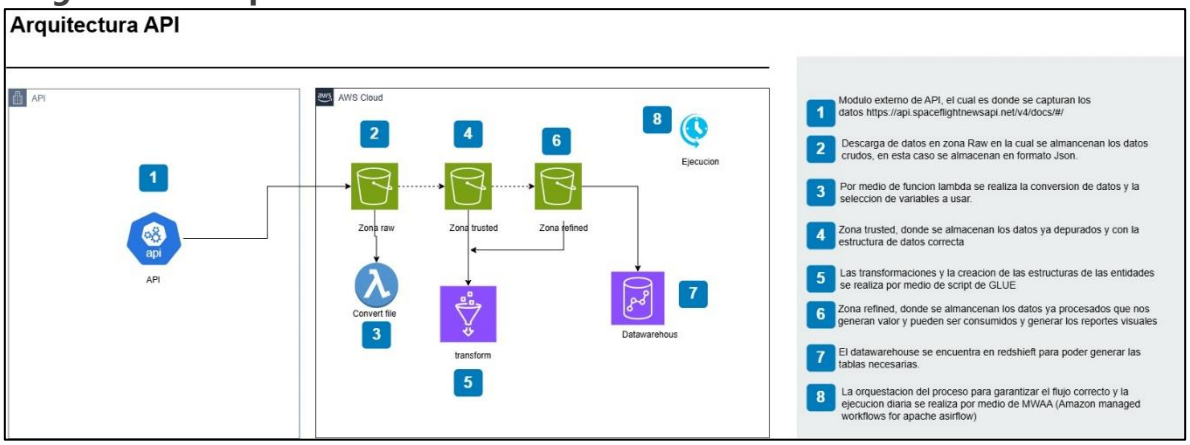
Documento Técnico: Flujo de Trabajo de Procesamiento de Datos

1. Introducción

- **Objetivo**: Describir el flujo de trabajo para la descarga, procesamiento y carga de datos desde la API https://api.spaceflightnewsapi.net/v4/docs/#/ hasta un data warehouse de Redshift.
- **Alcance**: Este documento cubre la arquitectura, estimación de volumen de datos, estrategia de almacenamiento, plan de contingencia y sistema de monitoreo.

2. Arquitectura del Flujo de Trabajo

Diagrama de Arquitectura



3. Estimación de Volumen de Datos

Fuente de Datos

- **API**: Datos de artículos, eventos y blogs relacionados con la industria espacial.
- **Frecuencia de Descarga**: Diaria.

Volumen Estimado Zona Raw

Fuente	Registros/Día	Tamaño / Día	Tamaño / Año
Articulos	~20	~20 kb	~7.2 Mb
Blogs	~2	~2 kb	~0.72 Mb
Reports	~1	~1 kb	~0.36 kb
Total	~**23**	~**23 kb**	~8.28 Mb

Volumen Estimado Zona Curada

Fuente	Registros/Día	Tamaño / Día	Tamaño / Año
dim_topic	~30	~2 kb	~0.72 Mb
Dim_source	~23	~2 kb	~0.72 Mb
Fact_article	~23	~2 kb	~0.72 kb
Total	~**23**	~6 **kb**	~2.16 Mb

Crecimiento Esperado

- Se espera un crecimiento del 20% anual en el volumen de datos debido al aumento de publicaciones y eventos, se estimaría 10 Mb por año, para la zona Raw y de 3 Mb por año para la zona Curada

4. Estrategia de Almacenamiento y Búsqueda

Almacenamiento en S3

- **Estructura de Carpetas**:
  - Datos crudos: s3://mi-bucket-s3/raw-data/date=<fecha>/.
  - Datos procesados: s3://mi-bucket-s3/processed-data/date=<fecha>/.
- **Formato de Archivos**:
  - Datos crudos: JSON.
  - Datos procesados: Parquet (optimizado para consultas y almacenamiento eficiente).
- **Retención de Datos**:
  - Datos crudos: 30 días (luego se eliminan o archivan en S3 Glacier).
  - Datos procesados: 1 año.

Búsqueda y Consultas

- **Amazon Athena**:

    - Para consultas SQL sobre datos en S3.
    - Ejemplo: Consultar artículos publicados en un rango de fechas.
- **Amazon Redshift**:
    - Para consultas analíticas sobre datos procesados.
    - Ejemplo: Agregaciones y análisis de tendencias.

5. Plan de Contingencia

Escenarios de Falla

1. **Falla en la Descarga de Datos**:
    - **Causa**: La API no está disponible o devuelve errores.
    - **Acción**:
        - Reintentar la descarga después de un tiempo de espera.
        - Notificar al equipo mediante alertas de CloudWatch.
2. **Falla en el Procesamiento de Datos**:
    - **Causa**: Errores en el script de PySpark o falta de recursos en Glue.
    - **Acción**:
        - Reintentar el job de Glue.
        - Escalar el tamaño del clúster de Glue si es necesario.
3. **Falla en la Carga a Redshift**:
    - **Causa**: Problemas de conexión o errores en el script de carga.
    - **Acción**:
        - Reintentar la carga.
        - Verificar la configuración de Redshift y los permisos de IAM.