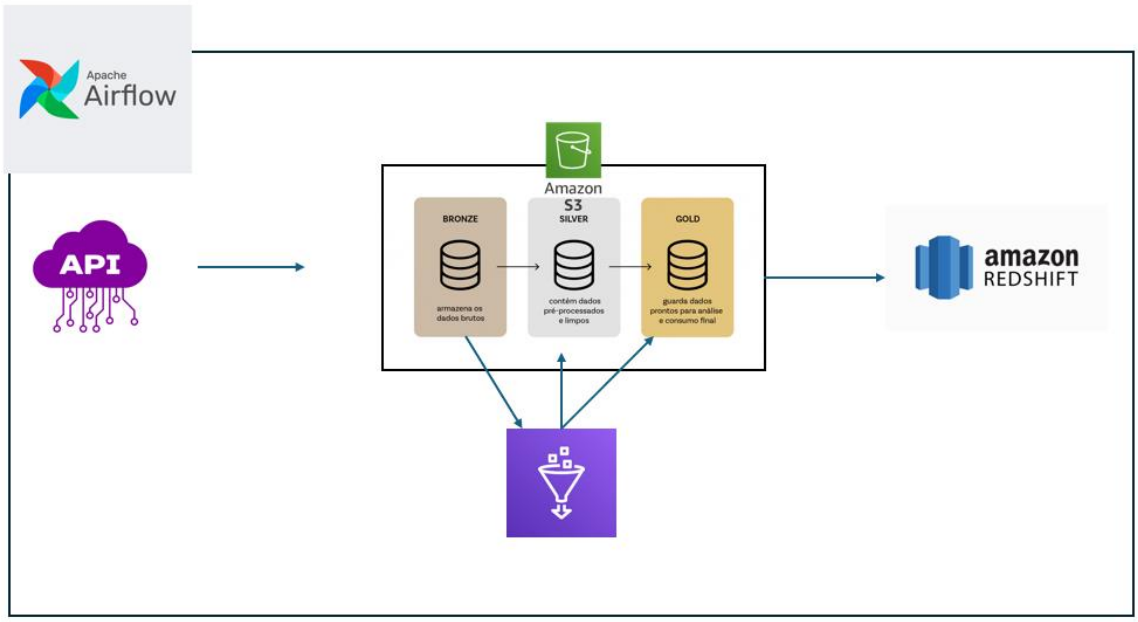
Diagrama y Arquitectura



Pipeline de Datos en AWS con Arquitectura de Almacenamiento en Capas (Medallón)

Razones para elegir los servicios

1. **Apache Airflow (Orquestador del pipeline):**
    - **Ventajas:**
        - Flexible y escalable para programar tareas complejas con dependencias claras.
        - Interfaz intuitiva para monitoreo.
        - Compatible con múltiples proveedores de nube, permitiendo integraciones nativas con AWS.
    - **Comparación:**
        - Frente a AWS Step Functions: Airflow es más adecuado para pipelines con dependencias complejas y lógica condicional.
        - Frente a herramientas como Prefect: Aunque Prefect tiene una configuración más sencilla, Airflow es más maduro y ampliamente adoptado.

2. **Amazon S3 (Almacenamiento):**
    - **Ventajas:**
        - Costo-eficiencia para almacenar grandes volúmenes de datos.
        - Integración nativa con Glue, Redshift, y otros servicios de AWS.
        - Soporte para versionado de datos y control de acceso granular.
    - **Comparación:**
        - Frente a Google Cloud Storage: S3 tiene mayor adopción y opciones avanzadas de ciclo de vida.
        - Frente a Azure Blob Storage: S3 ofrece precios más competitivos para almacenamiento masivo.
3. **AWS Glue (ETL con Spark):**
    - **Ventajas:**
        - Basado en Apache Spark, lo que asegura un procesamiento distribuido rápido y escalable.
        - Sin servidor, por lo que elimina la necesidad de manejar infraestructura.
        - Catálogo de datos integrado para descubrimiento y gestión.
    - **Comparación:**
        - Frente a Databricks: Glue es más económico y adecuado para tareas ETL estándar, mientras que Databricks sobresale en análisis avanzados y aprendizaje automático.
        - Frente a herramientas como Talend: Glue simplifica la integración con S3 y Redshift.
4. **Amazon Redshift (Data Warehouse):**
    - **Ventajas:**
        - Diseñado para análisis de datos a gran escala.
        - Compatibilidad con SQL estándar y herramientas de visualización como Tableau y Power BI.

        - Ofrece almacenamiento en columnas, optimizando consultas analíticas.
    - **Comparación:**
        - Frente a Snowflake: Redshift es más económico dentro del ecosistema AWS.
        - Frente a BigQuery: Redshift es más adecuado para usuarios con pipelines basados en AWS.

Arquitectura de Almacenamiento en Capa Medallón

La arquitectura en capa medallón divide los datos en tres niveles: **Raw, Processed, y Analytics**, asegurando un flujo claro desde la ingestión hasta el análisis.

1. **Capa Raw (Bronze):**
    - Datos sin procesar, directamente extraídos desde la API de Spaceflight News.
    - Almacenados en formato JSON o CSV.
    - Ejemplo: s3://spaceflight-news/raw/
2. **Capa Processed (Silver):**
    - Datos limpios y transformados para análisis intermedio.
    - Normalización de fechas, filtrado de campos irrelevantes y creación de métricas adicionales.
    - Almacenados en formato Parquet.
    - Ejemplo: s3://spaceflight-news/processed/
3. **Capa Analytics (Gold):**
    - Datos optimizados para reportes y análisis finales.
    - Almacenados directamente en Amazon Redshift.

Flujo de Datos del Pipeline

1. **Extracción y Almacenamiento:**

- **Airflow:** Orquesta una tarea para extraer datos de la API y almacenarlos en S3 (capa raw).
- **S3:** Garantiza la persistencia y versionado de los datos.

2. **Transformación y Limpieza:**
   - **AWS Glue:**
     - Ejecuta transformaciones utilizando Spark.
     - Convierte los datos a formato Parquet para almacenamiento eficiente.
     - Escribe los resultados en la capa processed.
3. **Carga y Análisis:**
   - **Airflow:** Gestiona la carga de datos transformados desde S3 a Redshift.
   - **Redshift:** Permite consultas SQL rápidas y escalables sobre los datos en la capa analytics.
4. **Backup y Recuperación:**
   - **S3:** Versionado y políticas de ciclo de vida configuradas para las capas raw y processed.
   - **Glue:** Scripts almacenados en repositorios como Git para mantenimiento.
   - **Redshift:** Snapshots automáticos y manuales.

Aspectos Técnicos Relevantes

1. **Formatos de Datos:**
   - JSON/CSV para datos raw (compatibilidad).
   - Parquet para datos procesados (rendimiento).
2. **Seguridad:**
   - Configuración de políticas IAM para acceso controlado.
   - Cifrado en reposo (S3, Redshift) y en tránsito (TLS).
3. **Monitoreo:**

- **CloudWatch:** Para logs y métricas.
- **Airflow UI:** Para monitoreo en tiempo real de los DAGs.
