Contexto del Proyecto

Desarrollarás un pipeline de datos completo utilizando la API de Spaceflight News

(https://api.spaceflightnewsapi.net/v4/docs/) para crear un sistema de análisis de

tendencias en la industria espacial. El sistema procesará artículos de noticias,

eventos y blogs para generar insights sobre la industria espacial.

Por: Juan de Dios Camacho Rocha

1. **La solución se implementa en AWS, incluyendo los siguientes componentes:**

Diagrama

Descripción generada automáticamente

AWS Lambda: Para la extracción de datos desde la API de Spaceflight News.

Amazon S3: Almacenamiento de datos en formato parquet para eficiencia.

AWS Glue: Transformación y limpieza de datos usando Spark.

Amazon Redshift: Data Warehouse para consultas y análisis.

Amazon QuickSight: Visualización de datos y reportes.

Apache Airflow: Orquestación del pipeline.

Flujo de los datos:

1. **Zona de datos crudos:**

Extracción: AWS Lambda consume los endpoints de la API y almacena los datos en S3.

* Se usa AWS Lambda para hacer requests a los endpoints:

articles, blogs, reports, info.

* Almacenamiento de datos en formato JSON en S3.
* Validaciones con glue

1. **Zona de Procesos:**

Procesamiento: AWS Glue limpia y transforma los datos con Apache Spark.

Procesamiento con Spark:

Limpieza de datos, normalización de fechas y campos nulos.

Extracción de palabras clave con NLP.

Clasificación de artículos por tema.

Caching de datos más consultados.

Particionamiento de datos por mes y fuente.

1. **Zona de Resultados:**

Carga: Los datos procesados se almacenan en Amazon Redshift.

1. **Análisis y Visualización:**

Queries SQL en Redshift y reportes en QuickSight.

1. Orquestador: Flujo del proceso en Airflow

El DAG en Airflow sigue estos pasos:

* Ejecuta una función Lambda para extraer datos desde la API de Spaceflight News y almacenarlos en Amazon S3.
* Ejecuta un job de AWS Glue para transformar y limpiar los datos.
* Carga los datos transformados en Amazon Redshift para análisis.
* Refresca los dashboards en QuickSight para visualizar los resultados.

from airflow import DAG

from airflow.providers.amazon.aws.operators.lambda\_function import AwsLambdaInvokeFunctionOperator

from airflow.providers.amazon.aws.operators.glue import GlueJobOperator

from airflow.providers.amazon.aws.operators.redshift\_sql import RedshiftSQLOperator

from airflow.providers.amazon.aws.operators.quicksight import QuickSightCreateIngestionOperator

from datetime import datetime

default\_args = {

'owner': 'diego',

'start\_date': datetime(2024, 1, 1),

'retries': 1

}

dag = DAG('space\_news\_pipeline', default\_args=default\_args, schedule\_interval='@daily')

# Ejecutar Lambda para extraer datos y almacenarlos en S3

extract\_data\_lambda = AwsLambdaInvokeFunctionOperator(

task\_id='invoke\_lambda',

function\_name='SpaceNewsDataExtractor',

aws\_conn\_id='aws\_default',

invocation\_type='Event',

dag=dag

)

# Ejecutar AWS Glue para procesar y limpiar los datos

glue\_job = GlueJobOperator(

task\_id='run\_glue\_job',

job\_name='space\_news\_processing',

aws\_conn\_id='aws\_default',

dag=dag

)

# Cargar los datos transformados en Redshift

load\_to\_redshift = RedshiftSQLOperator(

task\_id='load\_redshift',

redshift\_conn\_id='redshift\_default',

sql="""

COPY fact\_article

FROM 's3://datos-inetum/processed\_data/'

IAM\_ROLE 'arn:aws:iam::123456789012:role/MyRedshiftRole'

FORMAT AS PARQUET;

""",

dag=dag

)

# Refrescar datos en Amazon QuickSight

refresh\_quicksight = QuickSightCreateIngestionOperator(

task\_id="refresh\_quicksight",

aws\_conn\_id="aws\_default",

data\_set\_id="dataset-space-news",

ingestion\_id="refresh-ingestion",

dag=dag

)

Explicación del flujo

invoke\_lambda: Ejecuta la función Lambda (SpaceNewsDataExtractor) que extrae los datos desde la API y los guarda en Amazon S3.

run\_glue\_job: Inicia un AWS Glue Job (space\_news\_processing) que transforma y limpia los datos. Convierte los archivos JSON a formato Parquet para optimizar la consulta en Redshift.

load\_redshift: Ejecuta una consulta SQL en Amazon Redshift usando COPY para cargar los datos desde S3. Los datos transformados en Glue se almacenan en una tabla de Redshift

refresh\_quicksight: Refresca los dashboards en QuickSight para mostrar los datos actualizados.

1. **Modelado y Análisis**

Diseño del Data Warehouse:

**Tablas Dimensionales:**

CREATE TABLE dim\_news\_source (

source\_id INT PRIMARY KEY,

name VARCHAR(255),

url VARCHAR(255),

reliability\_score FLOAT

);

CREATE TABLE dim\_topic (

topic\_id INT PRIMARY KEY,

name VARCHAR(255),

category VARCHAR(255)

);

**Tabla de Hechos:**

CREATE TABLE fact\_article (

article\_id INT PRIMARY KEY,

source\_id INT REFERENCES dim\_news\_source(source\_id),

topic\_id INT REFERENCES dim\_topic(topic\_id),

published\_at TIMESTAMP

);

**Queries SQL para Análisis**

**Tendencias de temas por mes:**

SELECT topic\_id, DATE\_TRUNC('month', published\_at) as month, COUNT(\*) as total\_articles

FROM fact\_article

GROUP BY topic\_id, month

ORDER BY month DESC;

**Fuentes más influyentes:**

SELECT s.name, COUNT(a.article\_id) as total\_articles

FROM fact\_article a

JOIN dim\_news\_source s ON a.source\_id = s.source\_id

GROUP BY s.name

ORDER BY total\_articles DESC;