



# Presentación Pipeline utilizando la API de Spaceflight News para el Análisis de Tendencias en la Industria Espacial

Título: "ETL y Análisis de Noticias en  
Google Cloud Platform"



Juan Carlos Cortes Mendez

Colombia-Bogotá-Febrero 2025



## Objetivo principal:

"Automatizar la ingesta, transformación, depuración y análisis de noticias usando Google Cloud Platform"

### ✓ Beneficios esperados:

Mayor eficiencia en el procesamiento de datos

"Este proyecto busca extraer datos de la API de Spark News API, procesarlos con Spark en Dataproc y analizarlos en Big Data, proporcionando insights sobre tendencias en noticias."



# Tabla de contenido

# Arquitectura del Pipeline

-  
  -  
  -  
  -  
  -  
  -  
  -  
  - 
  -  
  - 

# Flujo del Pipeline



# Modelo de Datos en BigQuery

- [REDACTED]
  - [REDACTED] 
  - [REDACTED]
  - [REDACTED]
  - [REDACTED]
  - [REDACTED]
  - [REDACTED]
  -  [REDACTED]
  -  [REDACTED]
  -  [REDACTED]
  -  [REDACTED]

# Tabla de contenido

## Análisis de Datos y Resultados

- ✅ 
- ✅ 
- ✅ 
- ✅ 
- ✅ 

## Integración con Machine Learning

- ✅ 
- 
- 
- 
- 
- 
- 
- 
- 
- 
- 
- 
- 

## Pruebas y Validaciones

- ✅ 
- ✅ 
- 
- ✅ 
- 

1

# Arquitectura del Pipeline

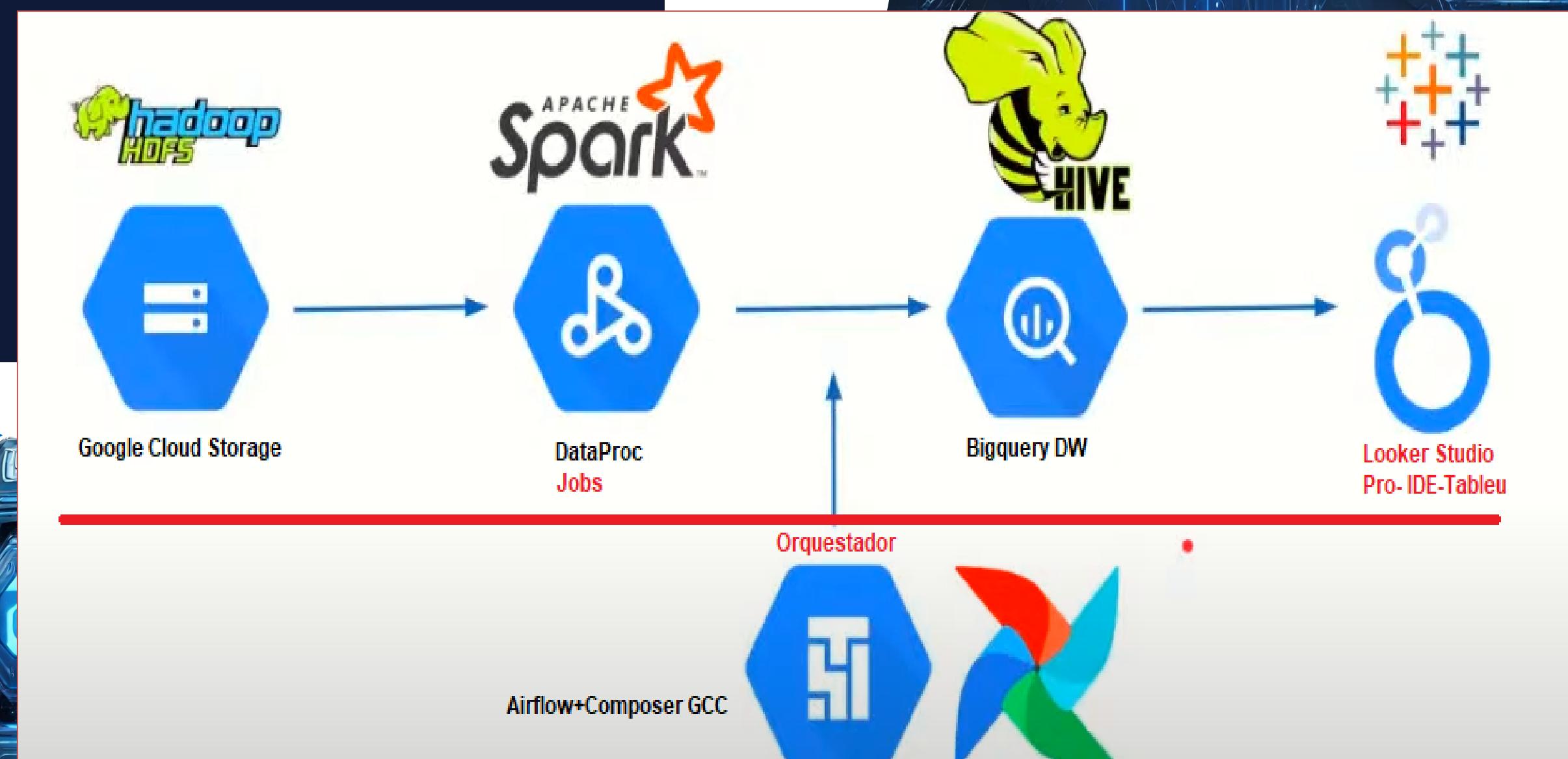
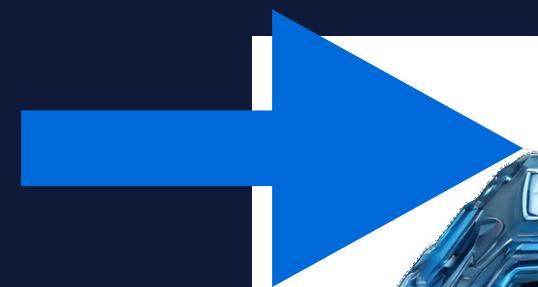
BATCH D

LINE



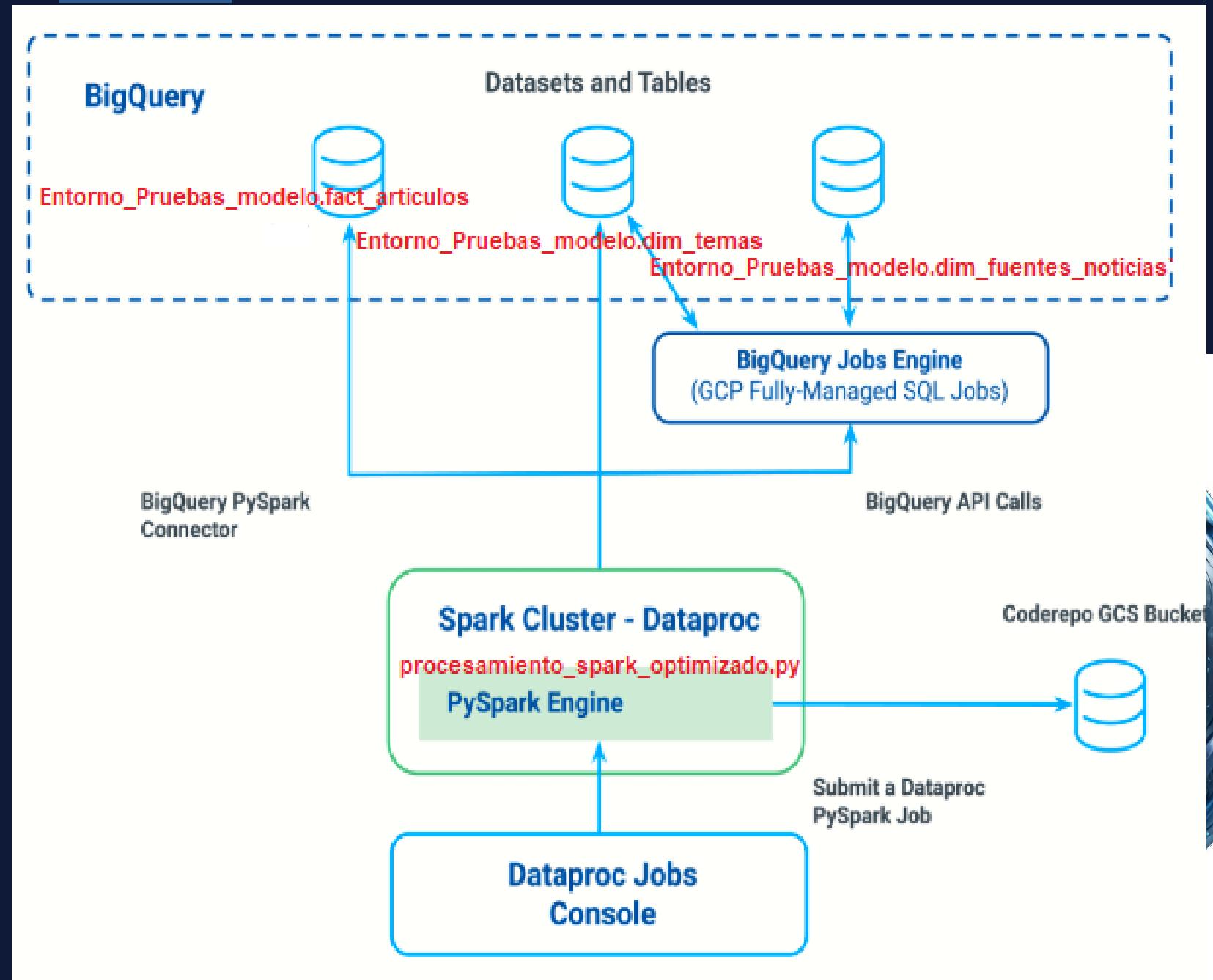
**API SpaceFlight  
News**

**INGESTA**



# 2

## TRANSFORMACIÓN DE DATOS



### Explicación del Flujo:

#### 1. Extracción de datos (`extract_articles`, `extract_blogs`, `extract_reports`):

- \* Utiliza PythonOperator para obtener datos de la API y guardarlos en archivos temporales.

#### 2. Limpieza y deduplicación con Spark (`limpia_y_deduplica.py`):

- \* Ejecuta un script de Spark almacenado en Google Cloud Storage para limpiar y deduplicar los datos.

- \* SparkSubmitOperator se conecta al cluster de Dataproc para ejecutar el trabajo.

#### Análisis y clasificación de temas (`proceso_analisis.py` y `identifica_topics.py`):

Estos trabajos de Spark realizan el análisis avanzado, extracción de palabras clave y clasificación de artículos.

#### Carga en BigQuery (`carga_data_procesada.py`):

Usa PythonOperator para cargar los datos procesados en BigQuery.

#### Generación de insights y actualización de dashboards (`generacion_diaria_insights`, `actualiza_dashboards`):

Genera reportes diarios y actualiza los dashboards.

# 3

# Modelo de Datos en BigQuery

## TRANSFORMACIÓN DE DATOS

Google Cloud analitica-contact-center-dev

Explorador + AGREGAR

Buscar recursos de BigQuery

Mostrar solo los destacados

- Cargas de trabajo
- Conexiones externas
- API CONSULTAS DINÁMICAS
- Consultas Tabla
- Entorno\_Pruebas\_modelo
- dim\_fuentes\_noticias
- dim\_temas
- fact\_articulos
- noticias\_procesadas

Consulta sin título

EJECUTAR GUARDAR DESCARGAR COMPARIR

```
81 MERGE INTO `analitica-contact-center-dev.Entorno_Pruebas_modelo.fact_articulos` AS destino
82 USING (
83     SELECT DISTINCT
84         ROW_NUMBER() OVER() AS article_id,
85         f.source_id,
86         t.topic_id,
87         TIMESTAMP(b.published_at) AS published_at,
88         b.title AS titulo,
89         b.summary AS resumen,
90         b.url,
91         CAST(FLOOR(RAND()*1000) AS INT64) AS visitas,
92         CAST(FLOOR(RAND()*500) AS INT64) AS compartidos
93     FROM `analitica-contact-center-dev.Entorno_Pruebas_modelo.noticias_procesadas` b
94     LEFT JOIN `analitica-contact-center-dev.Entorno_Pruebas_modelo.dim_fuentes_noticias` f
95     ON b.news_site = f.nombre
96     LEFT JOIN `analitica-contact-center-dev.Entorno_Pruebas_modelo.dim_temas` t
97     ON b.title = t.nombre
98 ) AS fuente
99 ON destino.article_id = fuente.article_id
100 WHEN NOT MATCHED THEN
101     INSERT (article_id, source_id, topic_id, published_at, titulo, resumen, url, visitas, compartidos)
```

Buscar (/) recursos, documentos, productos y más

\*Consulta... ulo Entorno... elo noticias... das dim\_fuen...ias

noticias\_procesadas

CONSULTA COMPARIR COPIAR INSTAN

ESQUEMA DETALLES VISTA PREVIA EXPLORADOR DE TABLAS VISTA PREVIA

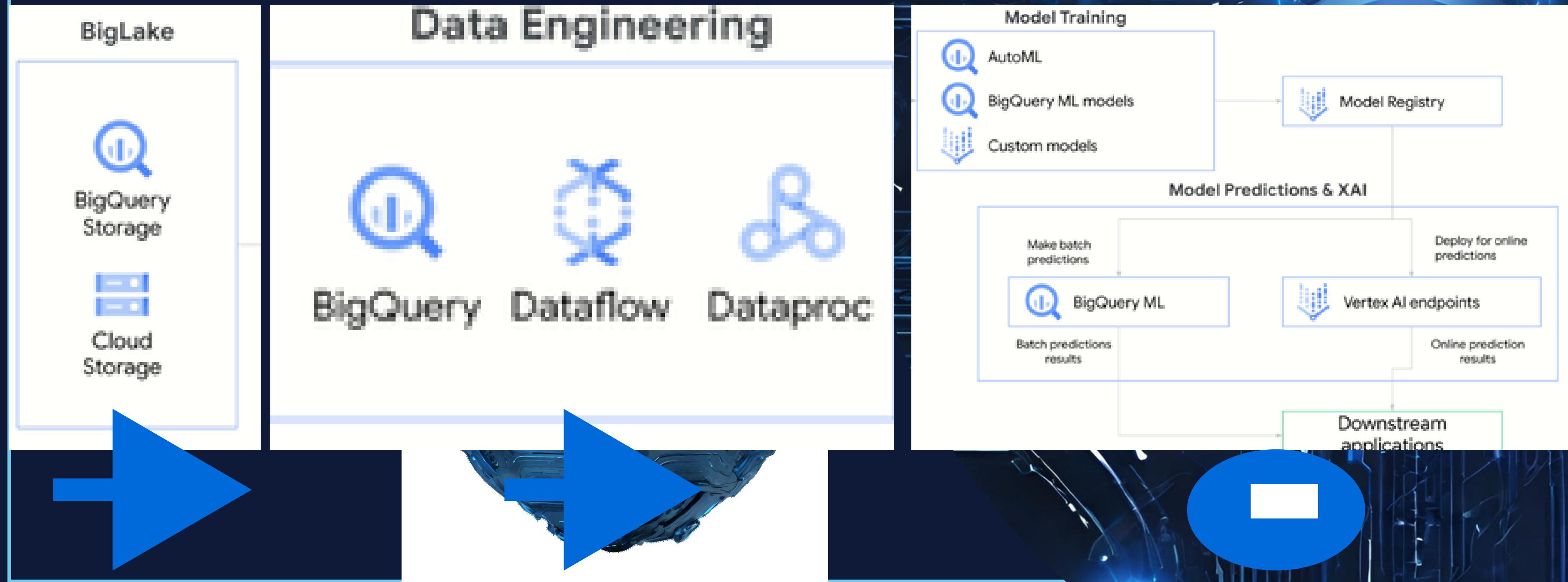
Filtro Ingresar el nombre o el valor de la propiedad

<input type="checkbox"/> Nombre del campo	Tipo	Modo	Clave	Intercalación	Valor predeterminado
<input type="checkbox"/> id_articulo	INTEGER	NULLABLE	-	-	-
<input type="checkbox"/> id_fuente	INTEGER	NULLABLE	-	-	-
<input type="checkbox"/> id_tema	INTEGER	NULLABLE	-	-	-
<input type="checkbox"/> fecha_publicacion	TIMESTAMP	NULLABLE	-	-	-
<input type="checkbox"/> titulo	STRING	NULLABLE	-	-	-
<input type="checkbox"/> resumen	STRING	NULLABLE	-	-	-
<input type="checkbox"/> url	STRING	NULLABLE	-	-	-
<input type="checkbox"/> visitas	INTEGER	NULLABLE	-	-	-
<input type="checkbox"/> compartidos	INTEGER	NULLABLE	-	-	-

4

# Integración con Machine Learning

## ARQUITECTURA

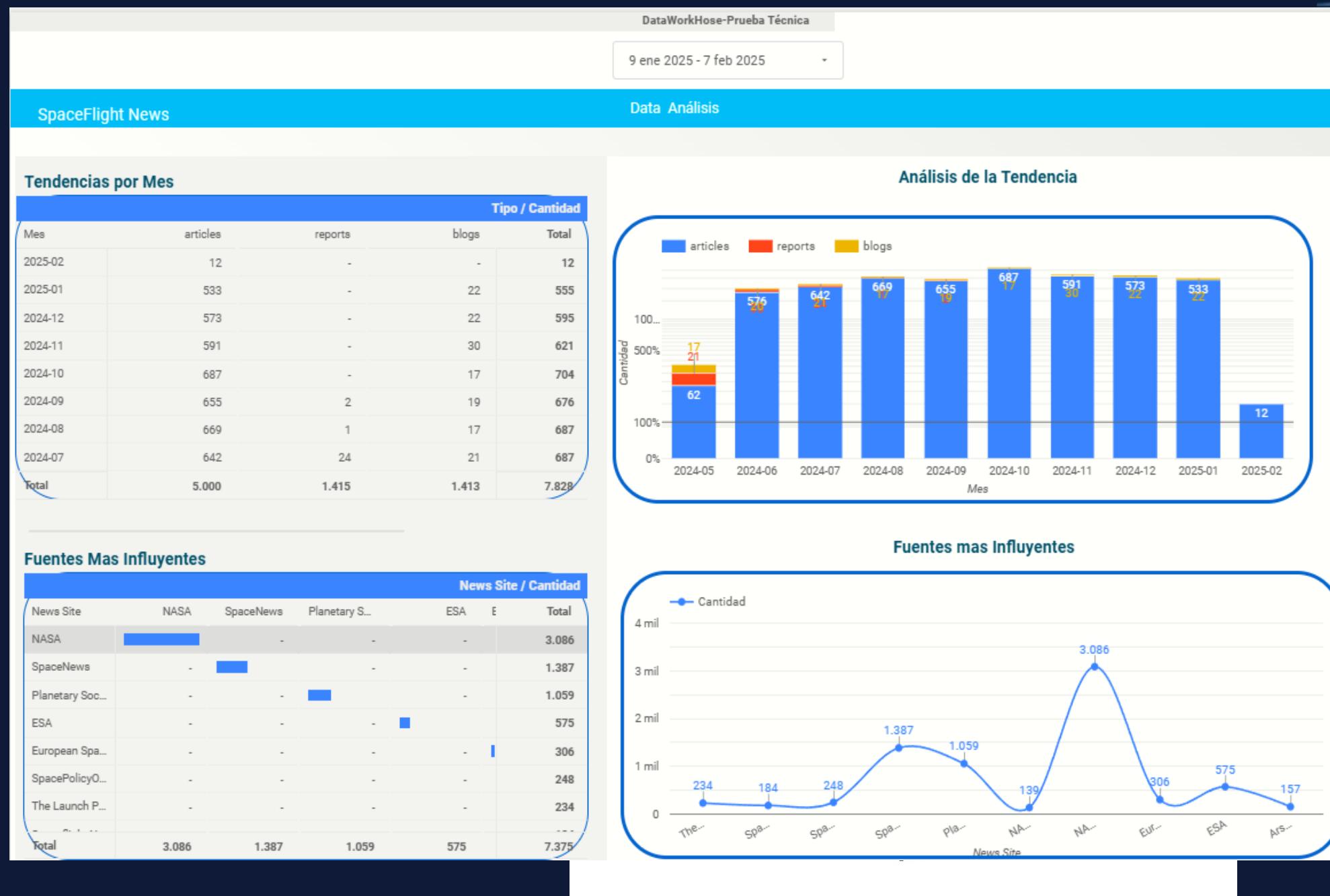


# 5

# Análisis de Datos y Resultados

## FUENTE DE NOTICIAS MAS INFLUYENTES

## TENDENCIAS POR MES

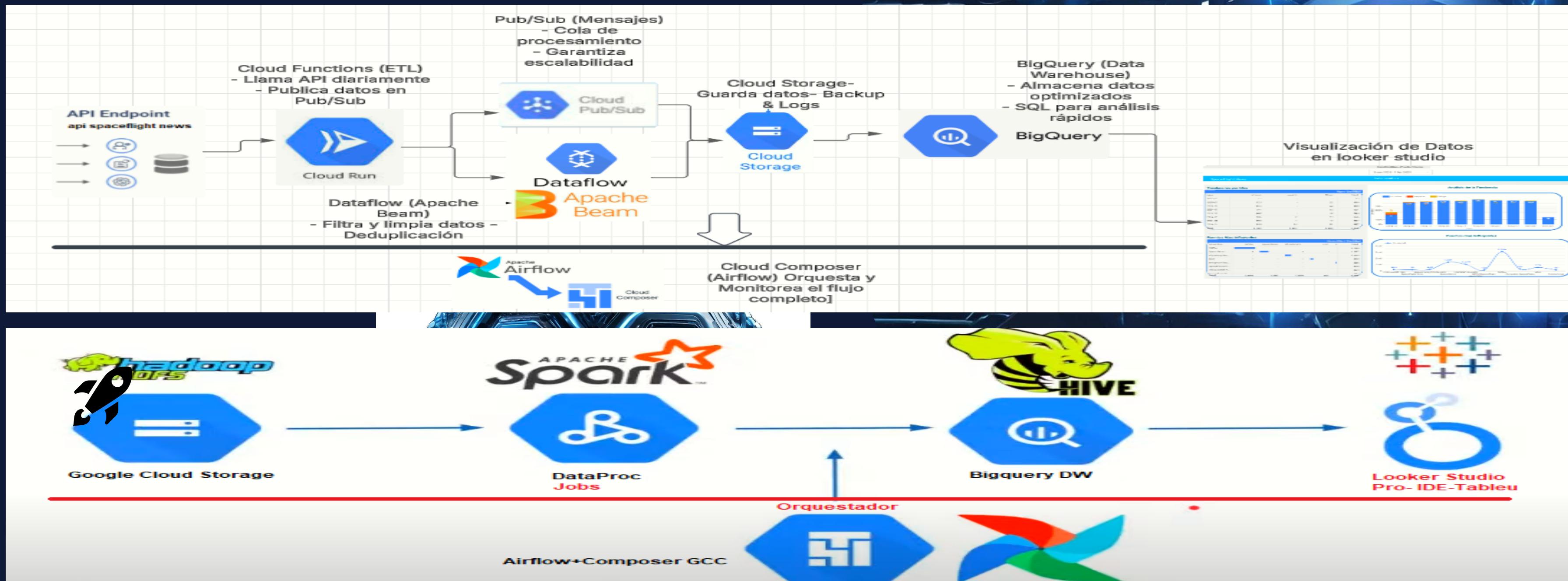


El tablero muestra el análisis de publicaciones sobre SpaceFlight entre 2024 y 2025. Octubre 2024 fue el mes con más publicaciones (704), mientras que febrero 2025 tuvo la menor actividad (12). Los artículos lideran el contenido con 5,000 publicaciones, seguido por reportes y blogs. La NASA es la fuente más influyente con 3,086 publicaciones, destacando claramente sobre SpaceNews y Planetary Society. Esta tendencia refleja una caída notable en 2025.

# 1.1 mejoras

## Opcion 2-Arquitectura del Pipeline

STREAMING Y 100% SERVELESS



## 1.2 Cloud

# Ventajas del Modelo de Pipeline en Google

### STREAMING Y 100% SERVERLESS

#### 1 Totalmente Automatizado y Orquestado

**Cloud Composer (Airflow)** la orquestación, lo que asegura que todas las tareas del pipeline se ejecuten en el orden correcto y sin intervención manual.

**Reintentos automáticos** en caso de fallos y **monitoreo constante** para asegurar la continuidad del proceso.

#### 2 Escalabilidad y Flexibilidad

• **Google Cloud Dataflow** permite procesar grandes volúmenes de datos en **tiempo real** o en **batch**, sin necesidad de administrar servidores.

• La **arquitectura serverless** garantiza escalabilidad automática según la carga de trabajo, lo que permite manejar tanto pequeñas como grandes cantidades de datos sin modificar la infraestructura.

#### 3 Optimización de Consultas y Costos

• **Particionamiento de datos históricos** en BigQuery mejora el rendimiento y reduce costos al ejecutar consultas solo en las particiones necesarias.



... necesidad de recalcular consultas, mejorando la eficiencia del sistema.

#### 4 Análisis y Visualización Avanzada

• BigQuery permite ejecutar consultas rápidas y complejas para identificar tendencias, analizar patrones y generar reportes automáticos.

• **Poker Studio** facilita la creación de dashboards dinámicos, proporcionando **visualización en tiempo real** de las principales métricas y tendencias.

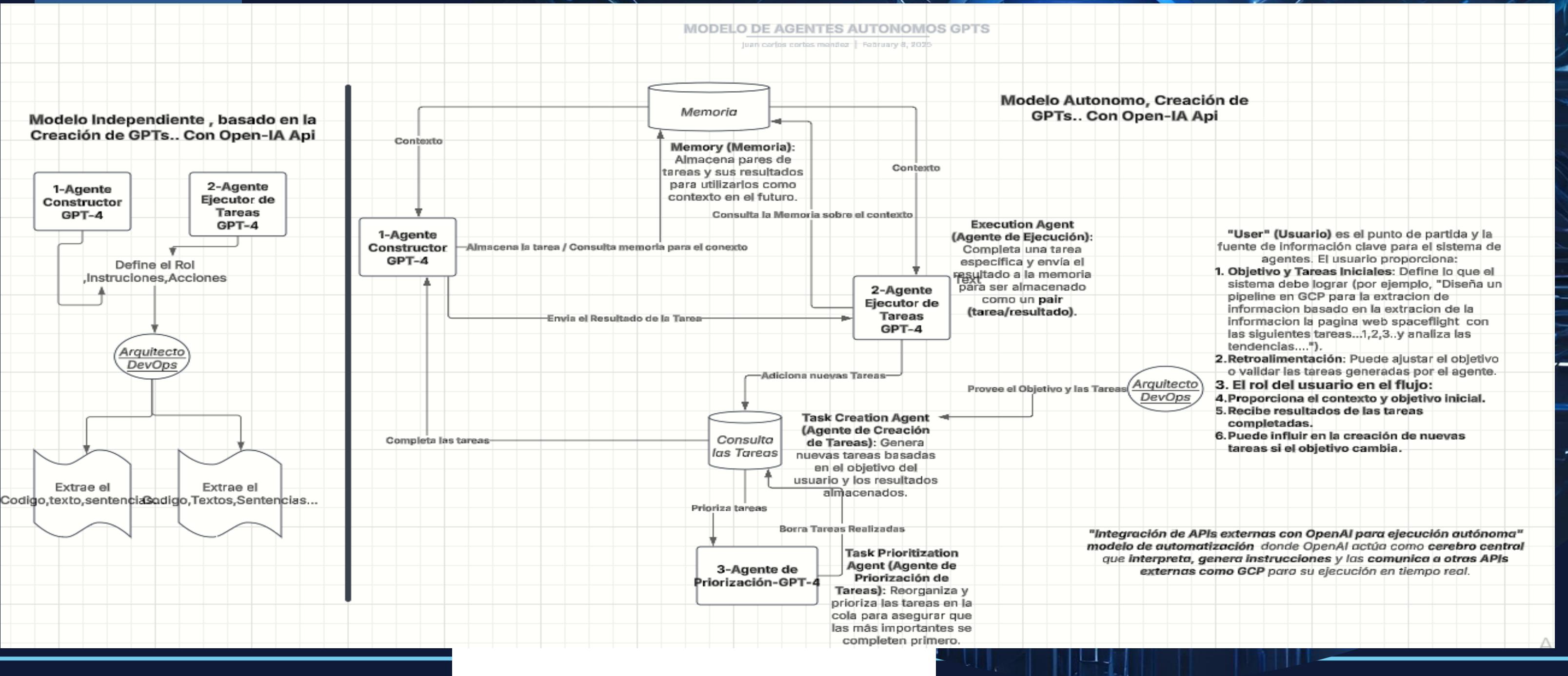
#### 5 Mantenimiento Simplificado

• **Arquitectura modular**: Cada componente (ingesta, procesamiento, almacenamiento) es independiente y fácil de actualizar, lo que simplifica el mantenimiento del sistema.

• **Cloud Functions y Dataflow** se integran directamente con Pub/Sub, reduciendo la complejidad del flujo de datos y mejorando la eficiencia.

# 1.3

# MODELO GENERADO CON IA-OPS BASADO EN AGENTES-AUTOGPTs.



# 6

# Integración con Machine Learning ☒

## MODELOS ML

```
-- ♦ Entrenar un Modelo de Predicción de Popularidad
CREATE OR REPLACE MODEL
`analitica-contact-center-dev.Entorno_Pruebas_modelo.modelo_prediccion_popularidad`
OPTIONS(
    model_type='LINEAR_REG',
    input_label_cols=['impacto_total']
) AS
SELECT
    f.nombre AS fuente,
    a.titulo,
    a.visitas,
    a.compartidos,
    (a.visitas + a.compartidos) AS impacto_total
FROM `analitica-contact-center-dev.Entorno_Pruebas_modelo.fact_articulos` a
JOIN `analitica-contact-center-dev.Entorno_Pruebas_modelo.dim_fuentes_noticias` f
ON a.source_id = f.source_id;
```

📌 Explicación:

- Entrena un modelo de Regresión Lineal con BigQuery ML
  - Usa datos históricos de visitas y compartidos como etiquetas
  - Predice el impacto de nuevos artículos
- o una opción de este modelo
- predicción al usar normalización y ajustar la métrica de impacto.
  - Más precisión con datos adicionales como categoria y duracion\_portada.
  - Modelo más robusto que no se ve afectado por escalas diferentes en las variables.

### Opción 2: Clasificación de Artículos con Vertex AI

Otra opción es usar Vertex AI para clasificar automáticamente los artículos en temas relevantes.

♦ Entrenar un Modelo de Clasificación en Vertex AI

- 1 Sube los datos a GCS  
bq extract --destination\_format CSV \
`analitica-contact-center-dev.Entorno\_Pruebas\_modelo.fact\_articulos` \
gs://us-central1-flujotransaccion-9cfbfa36-bucket/ml\_data/articulos.csv

2 Crea un Dataset en Vertex AI y entrena un modelo AutoML

```
gcloud ai datasets create --display-name="Dataset Noticias" --metadata-schema-uri=gs://google-cloud-
```

3 Desplegar el modelo y hacer inferencias

```
gcloud ai endpoints create --display-name="Clasificador Noticias"
```

```
gcloud ai models deploy --model=projects/analitica-contact-center-dev/models/clasificador_noticias
```

♦ Realizar una Predicción con el Modelo

```
gcloud ai endpoints predict \
--endpoint=projects/analitica-contact-center-dev/endpoints/clasificador_noticias \
--json-request=prediccion.json
```

📌 Explicación:

- Usa AutoML en Vertex AI para clasificar automáticamente los artículos
- Se puede conectar con BigQuery para análisis más avanzado

7

Preguntas  
Gracias

