

Presentación Pipeline de 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 y análisis de noticias usando GCP"

 Beneficios esperados: Mayor eficiencia en el procesamiento de datos

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



Juan Carlos Cortes Mendez

Colombia-Bogotá-Febrero 2024-



Tabla de contenido

Arquitectura del Pipeline

- Diagrama del flujo de datos:
- Herramientas utilizadas:
- Extracción → API Spaceflight News
- Orquestación → Cloud Composer (Airflow)
- Procesamiento → Dataproc (Apache Spark)
- Almacenamiento → Cloud Storage + BigQuery
- Visualización → Looker Studio
-

Flujo del Pipeline

- Explicación paso a paso:
- Se extraen noticias desde la API de Spaceflight News (con paginación y rate limits)
- Se almacenan en formato JSON en Google Cloud Storage
- Spark en Dataproc limpia y transforma los datos
- Se insertan en un modelo dimensional en BigQuery
- Se analizan tendencias con SQL y se visualizan en Looker Studio

Modelo de Datos en BigQuery

- Modelo de Datos en BigQuery
- Diagrama de las tablas en BigQuery
- dim_fuentes_noticias (Fuentes de noticias)
- dim_temas (Temas de artículos)
- fact_articulos (Datos principales con métricas de impacto)
- Optimización del Data Warehouse:
- Particionamiento por published_at
- Clustering por source_id y topic_id
- Estrategia de actualización con MERGE
-

Tabla de contenido

Análisis de Datos y Resultados

-  Consultas SQL clave:
-  Tendencias de temas por mes
-  Fuentes de noticias más influyentes
-  Predicción de impacto de artículos con ML

Integración con Machine Learning

-  Explicación paso a paso:
-  1 Se extraen noticias desde la API de Spaceflight News (con paginación y rate limits)
-  2 Se almacenan en formato JSON en Google Cloud Storage
-  3 Spark en Dataproc limpia y transforma los datos
-  4 Se insertan en un modelo dimensional en BigQuery
-  5 Se analizan tendencias con SQL y se visualizan en Looker Studio (Modelo IDE)

Pruebas y Validaciones

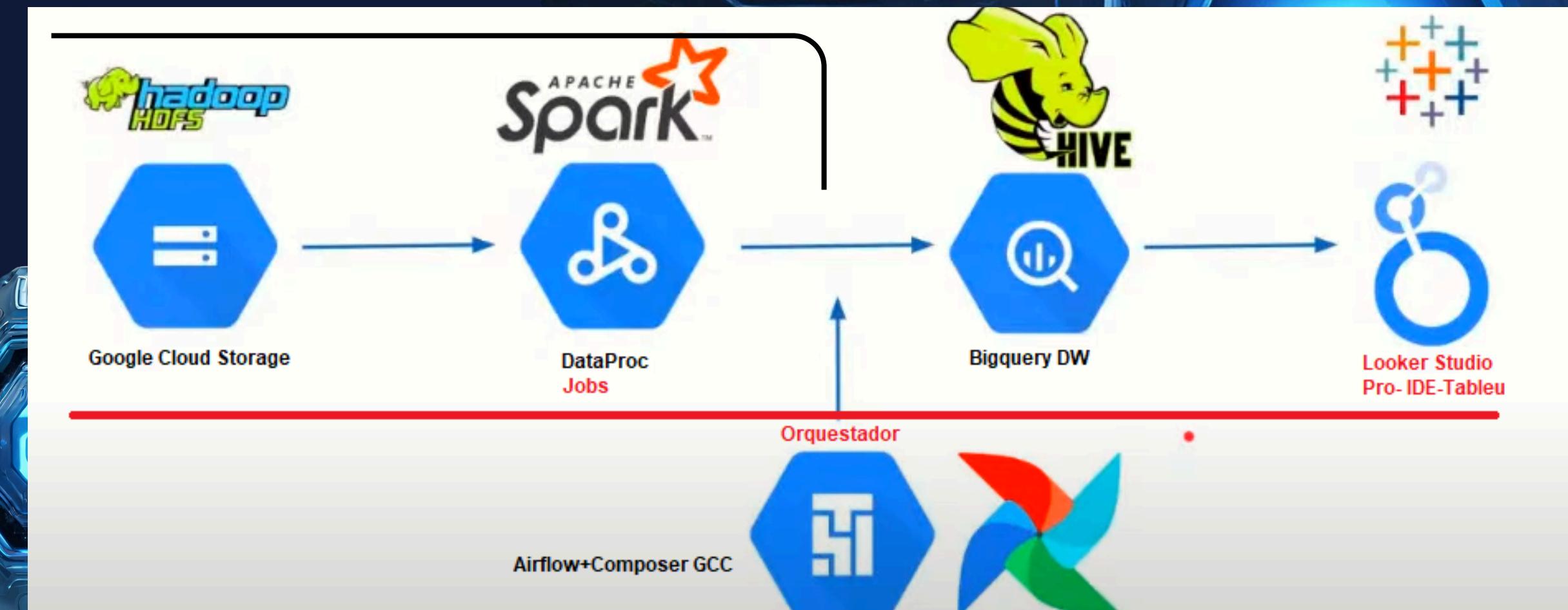
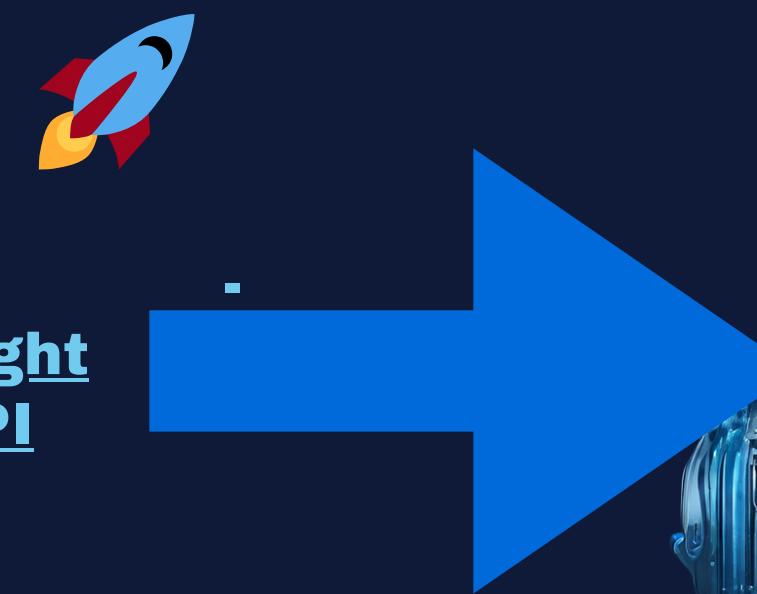
-  Tests unitarios:
-  Pruebas en Airflow → Verifica que el DAG funciona correctamente
-  Pruebas en BigQuery → Validación de datos antes de insertarlos

1

Arquitectura del Pipeline

INGESTA

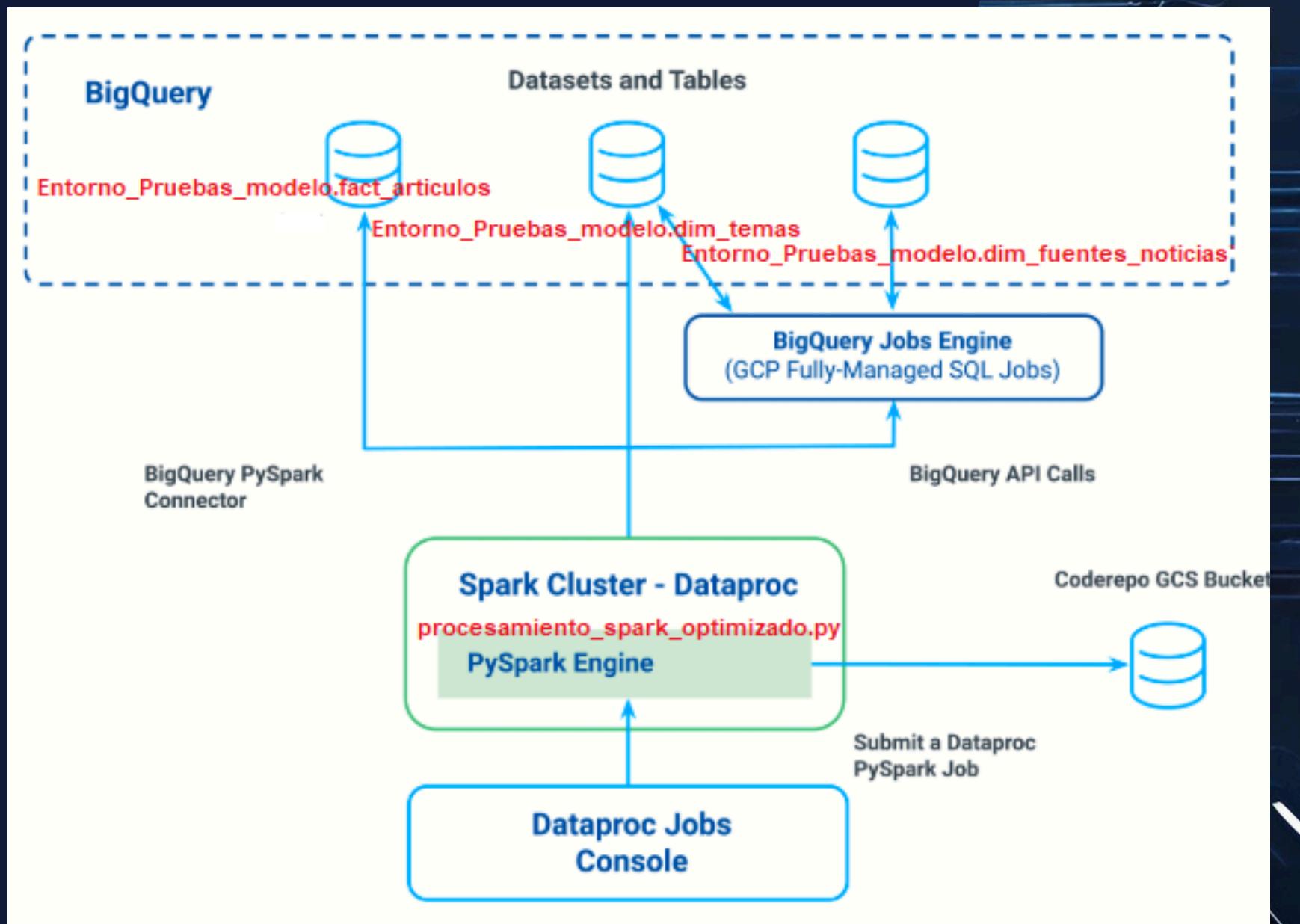
API
Spaceflight
News API



2

Flujo del PipeLine

TRANSFORMACIÓN DE DATOS



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 COMPARTE

```
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 COMPARTE COPIAR INSTAI

ESQUEMA DETALLES VISTA PREVIA EXPLORADOR DE TABLAS VISTA PREVIA

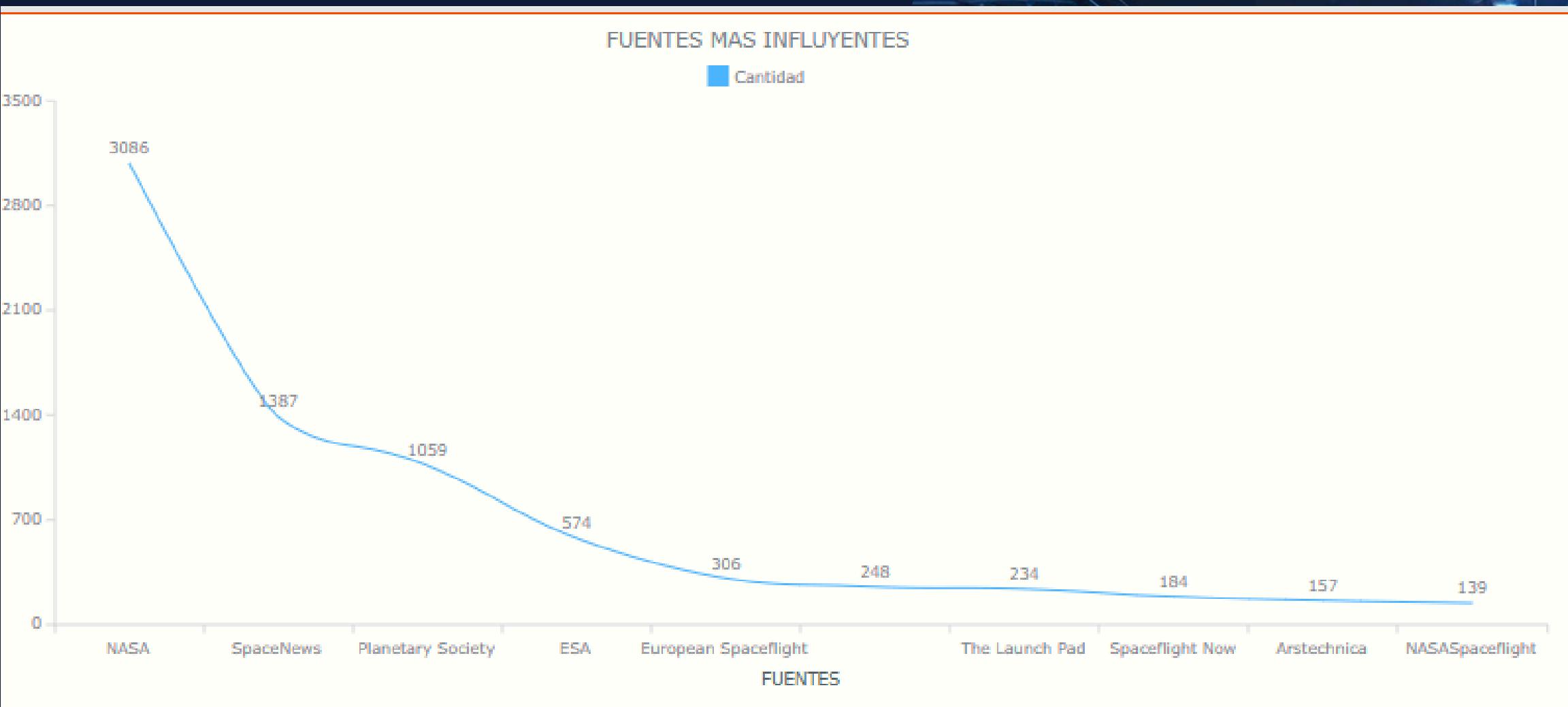
Filtro Ingresar el nombre o el valor de la propiedad

| <input type="checkbox"/> Nombre del campo | Tipo | Modo | Clave | Intercalación | Valor predete |
|--|-----------|----------|-------|---------------|---------------|
| <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

Análisis de Datos y Resultados

FUENTE DE NOTICIAS MAS INFLUYENTES

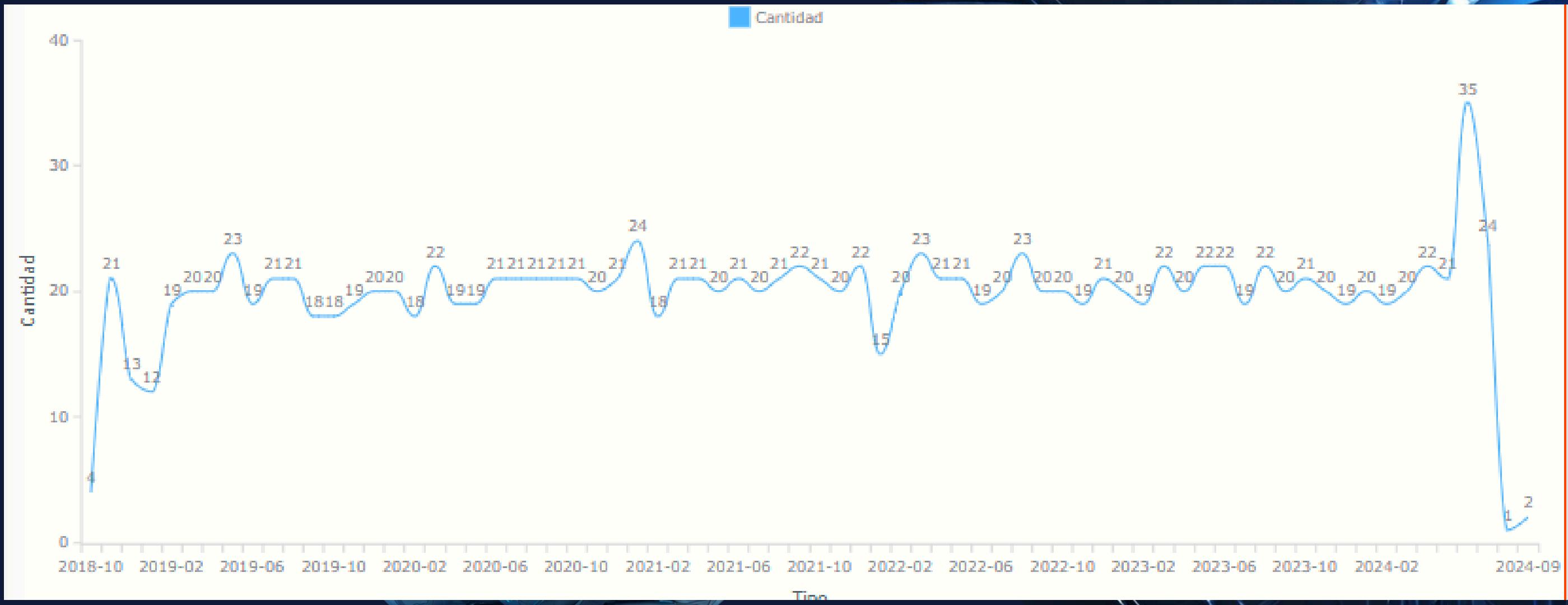


4

Análisis de Datos y Resultados



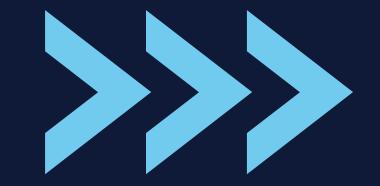
TENDENCIAS POR MES



4

Preguntas?





Gracias

