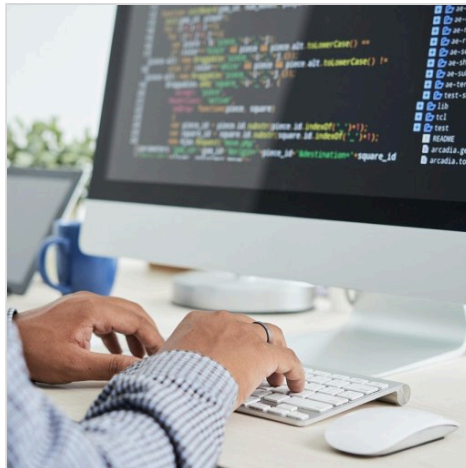


L U M B R E



Ingesta de fuentes heterogéneas

Grado: 6

Materia: Ingeniería de Datos · Ingeniería de Datos

11/10/2026

Antes de empezar

Los datos llegan siempre peores de lo que prometen: una API que cambia de versión, un CSV con comodines sueltos, una base con horarios imposibles. Ingerir con criterio es aislar ese caos en conectores robustos que no revientan al primer imprevisto.

QUÉ APRENDERÁS

- Conectar APIs, archivos y bases de datos dispares, normalizando formatos y manejo de errores para volcar todo en un destino único y fiable.

Ingesta de fuentes heterogéneas

Un marco, mil fuentes

Cada origen habla su idioma: REST con JSON paginado, archivos con delimitadores traicioneros, bases de datos con conexiones que caducan. La ingesta madura no es un script por fuente, sino un marco que absorbe las diferencias y entrega datos homogéneos.

VIDEO

Ingestión de datos para ingenieros de datos

Vídeo en español sobre cómo conectar orígenes dispares con conectores robustos.



Escanea para verlo en Lumbre

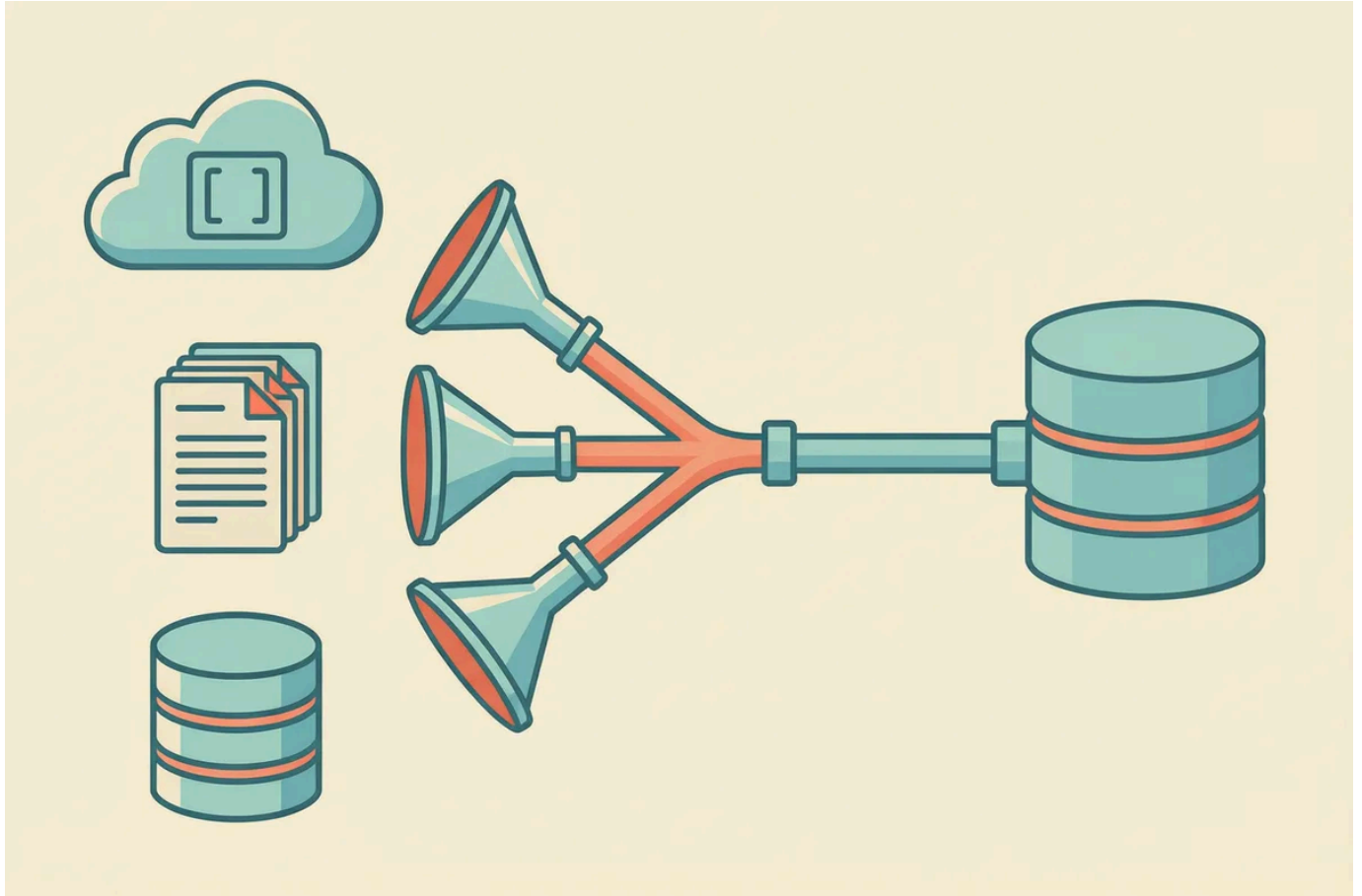
<https://www.youtube.com/watch?v=UXX-h7B4bf8>

Tipos de origen

Las APIs exponen datos paginados y con autenticación y límites de peticiones. Los archivos llegan por lotes con formatos y codificaciones variables. Las bases de datos permiten lecturas incrementales por marca temporal o registro de cambios.

El reto de la paginación

Una API rara vez devuelve todo de golpe: hay que recorrer páginas respetando cursores y cabeceras, sin perder ni repetir registros si el orden cambia entre llamadas. Mal hecho, duplica o deja huecos.



El marco de ingesta absorbe las diferencias de cada fuente y entrega un flujo homogéneo al destino.

Formatos y esquema

- Normaliza nombres y tipos a un esquema común antes de persistir.
- Decodifica y limpia delimitadores y caracteres en archivos planos.
- Registra el linaje: de qué origen y en qué lote llegó cada fila.

Manejo de errores

Los fallos son la norma: timeouts, cotas superadas, campos ausentes. Conviene reintentos con espera creciente, circuit breakers para no machacar un origen caído y cuarentena para filas malas en vez de abortar toda la carga.

Idempotencia de la lectura

Relanzar una ingesta no debe duplicar. Marcar cada lote con un identificador y escribir de forma que repetirlo no dañe deja la carga reproducible, cualidad que se paga diseñarla desde el primer conector.

Frecuencia y volumen

No todo se refresca igual: unas fuentes piden streaming casi continuo, otras lotes nocturnos. Fijar la cadencia según frescura requerida y coste evita sobrecargar orígenes y almacén sin necesidad.

Contratos de esquema

Cuando un origen cambia sin avisar, el pipeline debe detectarlo y no propagar el desastre. Validar contra un contrato y alertar ante desviaciones convierte una rotura silenciosa en un aviso controlado.

Errores frecuentes

- Programar un script distinto por fuente en vez de un marco que absorba las diferencias.
- Recorrer la paginación confiando en el orden y acabar con filas repetidas o perdidas.
- Abortar toda la carga por una fila mala en lugar de cuarentenarla y seguir con el resto.

Ejemplo resuelto: ingerir una API paginada de sensores

Una plataforma lee miles de lecturas repartidas en páginas de una API que además limita peticiones. El diseño:

1. Respetar el límite de peticiones con reintentos y espera creciente para no machacar el origen.
2. Avanzar con cursores estables, no con números de página, aunque el orden cambie entre llamadas.
3. Normalizar nombres y tipos al esquema común antes de persistir.
4. Escribir de forma idempotente marcando cada lote, para que relanzar no duplique.
5. Enviar a cuarentena las filas malformadas y registrar el linaje de cada lote.

¿Para qué sirve en la realidad?

Toda integración de datos, desde un panel climático hasta un catálogo de proveedores, depende de conectores que no revientan al primer imprevisto. La robustez aquí decide si el análisis tiene datos frescos o se queda ciego.

1. Una API REST devuelve resultados repartidos en páginas y el orden de los registros cambia entre llamadas. El riesgo principal de una ingesta mal diseñada es:

- ☐ que el almacén se quede sin espacio
- ☐ perder o duplicar registros al recorrer las páginas
- ☐ que el JSON llegue demasiado rápido
- ☐ que la autenticación sobre-escriba datos

2. Relanzar una ingesta idempotente sobre el mismo lote debe duplicar los registros ya cargados.

- ☐ Verdadero
- ☐ Falso

ACTIVIDAD INTERACTIVA DISPONIBLE EN LUMBRE

3. Ingesta



4. Ordena las fases de un conector de ingesta robusto:

Elementos:

- autenticarse y abrir canal con el origen
- recorrer páginas o archivos sin perder filas
- homogeneizar tipos y nombres al esquema común
- escribir de forma idempotente en el destino

Conectar	. . .
Leer	. . .
Normalizar	. . .
Persistir	. . .

Integración de datos https://en.wikipedia.org/wiki/Extract,_transform,_load Patrones de conexión a orígenes dispares y consolidación en un destino único.

5. Une cada fallo de la ingesta con su remedio.

5.	el origen limita peticiones por segundo	-----
5.	el orden cambia entre páginas	-----
5.	una fila llega malformada	-----
5.	el origen cambia de esquema sin avisar	-----

Opciones: backoff · cursores estables · cuarentena · contrato de esquema

Un marco, mil fuentes

Ingesta heterogénea

Origen

API paginada

archivos por lotes

base incremental

Robustez

reintentos con backoff

circuit breaker

cuarentena

Normalizar

esquema común

linaje por lote

contrato de esquema

Cadencia

streaming o lote

según frescura y coste

El mapa mental interactivo es interactivo en la versión online.



Qué es un pipeline de datos <https://www.ibm.com/mx-es/think/topics/data-pipeline> Visión general del movimiento y la preparación de datos desde orígenes dispares.

6. Un CSV semanal aparece a veces con una columna nueva y otra renombrada. ¿Tu ingesta debería fallar, ignorar o cuarentenar? Decide con el contrato de esquema y justifica.

Referencia curricular

Grado:	6
Materia:	Ingeniería de Datos · Ingeniería de Datos

Versión interactiva completa (con actividades, videos y autoevaluación):



<https://univ.lumbre.lat/object/cdia-ingd-002-ingesta-de-fuentes-heterogeneas>

Material educativo generado desde Lumbre · lumbre.lat