

L U M B R E



Calidad y observabilidad de datos

Grado: 6

Materia: Ingeniería de Datos · Pipeline de Datos

11/10/2026

Antes de empezar

Porque un pipeline en verde no garantiza un dato confiable: dimensiones de calidad, contratos, pruebas y las cinco señales de observabilidad para atrapar errores silenciosos.

QUÉ APRENDERÁS

- Aplicar las dimensiones de calidad de datos, pruebas en el pipeline, contratos y observabilidad para detectar y contener datos malos antes de que lleguen al consumidor.

Calidad y observabilidad de datos

VIDEO

Calidad y observabilidad de datos

Video explicativo (inglés): los cinco pilares de la observabilidad de datos y su diferencia con la calidad de datos tradicional.



Escanea para verlo en Lumbre

<https://www.youtube.com/watch?v=JlvwiLRu3QE>

Correr no es lo mismo que ser confiable

Que el pipeline termine en verde no significa que el dato esté bien. Un valor nulo mal colocado o una fuente que dejó de alimentar puede dejar un KPI en cero sin que ningún trabajo falle. La observabilidad trata el dato como un servicio que se vigila.

Los datos viajan por cadenas de transformaciones donde un error silencioso se propaga aguas abajo sin aviso. La **calidad de datos** pone reglas que detectan esos errores; la **observabilidad** vigila la salud del pipeline en producción para enterarse antes que el consumidor.

Las dimensiones de la calidad

- Completitud: no faltan filas ni valores obligatorios.
- Unicidad: una entidad real no aparece duplicada.
- Validez: los valores respetan tipo, formato y rango esperados.
- Consistencia: no se contradicen entre sistemas que deberían coincidir.
- Oportunidad (frescura): el dato llegó a tiempo para ser útil.

Contratos de datos

Un **contrato de datos** es un acuerdo explícito entre quien produce un dataset y quien lo consume: esquema, semántica, formato y nivel de servicio. Cuando el productor cambia una columna sin avisar (schema drift), el contrato lo detecta en lugar de romper al consumidor en silencio.

Pruebas dentro del pipeline

Las mismas ideas que se aplican al código se aplican al dato: antes y después de transformar, el pipeline ejecuta **pruebas** que aprueban o bloquean. Herramientas como dbt o Great Expectations declaran esas reglas junto al modelo.

- No nulos y únicos en llaves primarias.
- Rango y accepted values en columnas sensibles.
- Relacional: toda clave foránea existe en su tabla padre.
- Volumen esperado: el conteo de hoy no cae abruptamente.



La calidad se define por dimensiones, se vigila con técnicas y se responde con acciones acotadas.

Observabilidad: las cinco señales

La observabilidad moderna de datos monitorea automáticamente cinco señales que suelen anticipar un problema: **frescura** (¿llegó a tiempo?), **volumen** (¿cuántas filas?), **esquema** (¿cambió la estructura?), **linaje** (¿qué depende de qué?) y **distribución** (¿se movieron los valores?).

Cada dimensión se traduce en una regla concreta que el pipeline puede ejecutar.

Dimensión	Pregunta que responde	Ejemplo de regla
Completitud	¿Faltan datos?	id_pedido nunca nulo
Unicidad	¿Hay duplicados?	email es único
Validez	¿El valor tiene sentido?	edad entre 0 y 120
Frescura	¿Está al día?	tabla actualizada en las últimas 6 h
Volumen	¿Es el esperado?	entre 80k y 120k filas/día

Linaje: de dónde viene y a dónde va

El **linaje** registra el camino de cada dato desde su origen hasta cada tablero. Sin linaje, un cambio en una fuente es una apuesta a ciegas; con linaje, se sabe exactamente qué reportes y equipos se verán afectados antes de publicar.

Cuando algo falla

- Detectar: la prueba o la señal de observabilidad dispara.
- Aislar: no propagar el dato malo aguas abajo.
- Alertar con contexto: qué regla, qué tabla, desde cuándo.
- Corregir y reprocesar (backfill) una vez arreglada la causa.

Errores frecuentes

- Validar solo al final de la cadena, cuando el daño ya se propagó.
- No poner umbral de volumen, así que una fuente que deja de mandar pasa inadvertida.
- Tener un tablero de calidad que nadie mira y sin dueño asignado.

Ejemplo resuelto: un KPI amanece en cero

El ingreso diario del directorio muestra cero esta mañana aunque ayer estaba bien.

Secuencia de diagnóstico:

1. Revisar frescura: ¿llegó el lote de hoy de la fuente de pagos?
2. Consultar volumen: el conteo de la tabla de hechos viene desplomado.
3. Seguir el linaje hacia arriba hasta el modelo que alimenta el KPI.
4. Confirmar la causa: cambió el esquema de la fuente y el parser dejó de leer una columna.
5. Aislar el modelo, arreglar el contrato y lanzar el backfill del día.

¿Para qué sirve en la realidad?

La confianza es el activo de un equipo de datos: un solo KPI errado publicado destruye la credibilidad de todo el almacén. La calidad y la observabilidad convierten errores silenciosos en alertas manejables.

1. ¿Qué distingue a la observabilidad de datos de simplemente "que el pipeline no falle"?

- ☐ Que el pipeline termine sin errores basta para confiar en el dato
- ☐ La observabilidad vigila frescura, volumen y distribución para detectar datos malos aunque el job termine bien
- ☐ Solo se necesita revisar el código SQL
- ☐ Reemplaza la necesidad de pruebas

2. Un contrato de datos ayuda a detectar el schema drift antes de que rompa a los consumidores.

☐ Verdadero

☐ Falso

3. La dimensión de calidad que pregunta si el dato llegó a tiempo se llama o frescura.

4. Clasifica cada control según a qué dimensión de calidad pertenece.

Elementos:

- edad entre 0 y 120
- email sin duplicados
- tabla actualizada en las últimas 6 h
- formato de correo correcto
- id de pedido único

Validez	. . .
Unicidad	. . .
Frescura	. . .

5. Une cada señal de observabilidad con lo que vigila.

5.	Frescura	-----
5.	Volumen	-----
5.	Esquema	-----
5.	Linaje	-----

Opciones: si el dato llegó a tiempo · cuántas filas entran · cambios en la estructura · dependencias aguas arriba y abajo

ACTIVIDAD INTERACTIVA DISPONIBLE EN
LUMBRE

6. Calidad y observabilidad



Calidad y observabilidad de datos

Calidad y observabilidad

Dimensiones

completitud

unicidad y validez

frescura y volumen

Técnicas

pruebas en el pipeline

contratos de datos

linaje

Respuesta

detectar y aislar

alertar con contexto

backfill

El mapa mental interactivo es interactivo en la versión online.



Pruebas de datos en dbt (documentación oficial) <https://docs.getdbt.com/docs/build/data-tests> Cómo se declaran pruebas de unicidad, no nulos y relaciones junto al modelo.

7. Amanece un tablero ejecutivo con una métrica a la mitad del valor habitual, pero todos los jobs terminaron en verde. ¿Qué señal de observabilidad debería haberlo atrapado y qué prueba o umbral implementarías para que no vuelva a pasar desapercibido?

Referencia curricular

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

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



<https://univ.lumbre.lat/object/cdia-ingd-005-calidad-y-observabilidad-de-datos>

Material educativo generado desde Lumbre · lumbre.lat