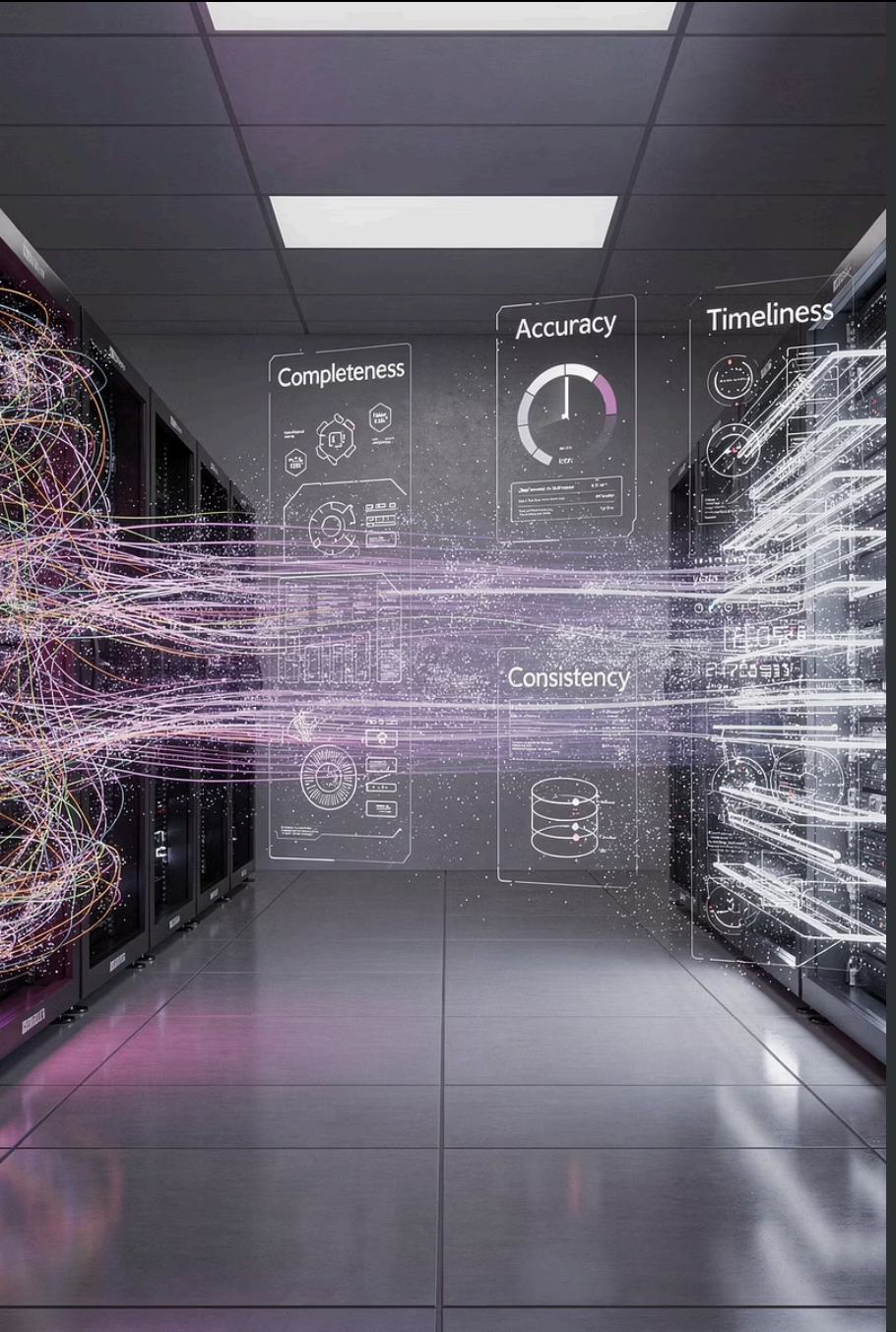

Calidad de Datos en el ETL: Las 6 Dimensiones

Criterios de aceptación

¿Qué evaluamos antes del ETL? Las 6 dimensiones de calidad de datos

Antes de definir criterios de aceptación o descarte, es necesario entender qué dimensiones de calidad están en juego. El framework DAMA establece 6 dimensiones medibles y replicables.

Exactitud

Los datos reflejan la realidad sin errores.

Coherencia

Los datos son consistentes entre tablas y fuentes.

Unicidad

No existen registros duplicados.

Integridad

No hay campos requeridos vacíos o nulos.

Puntualidad

Los datos están disponibles cuando se necesitan.

Validez

Los datos cumplen formatos y reglas definidas.

Fuente: DAMA International — Data Management Body of Knowledge (DMBOK)

Métricas cuantificables para cada dimensión

Cada dimensión se puede calcular con fórmulas simples sobre el dataset antes y después del proceso ETL.

Dimensión	Fórmula	Umbral recomendado
Integridad	(Valores no nulos / Total requeridos) × 100	> 95% campos críticos
Exactitud	(Registros correctos / Total muestreados) × 100	> 90%
Unicidad	(Registros únicos / Total registros) × 100	> 98%
Validez	(Registros con formato correcto / Total) × 100	> 92%
Coherencia	(Registros consistentes entre tablas / Total) × 100	> 90%
Puntualidad	(Registros a tiempo / Total requeridos) × 100	> 90% / < 2 días

Estas métricas deben calcularse tanto en el dataset fuente como en el resultado final del ETL para identificar degradaciones o mejoras en la calidad.

De la medición a la decisión: las dimensiones guían el ETL

Los resultados de medir las 6 dimensiones determinan directamente qué criterio aplicar en cada etapa del proceso ETL.

Medir

Pre-ETL

- Perfilar el dataset aplicando las 6 dimensiones
- Identificar campos con baja integridad, duplicados o formatos inválidos

Decidir

Criterios

- Aceptar si la métrica supera el umbral
- Descartar si está por debajo y el campo es crítico
- Imputar si el porcentaje de faltantes lo permite

Documentar

Post-ETL

- Registrar hallazgos por dimensión
- Consignar las transformaciones aplicadas
- Evidenciar los umbrales usados como referencia

Criterios de aceptación

Cumplen con las reglas de negocio

Los datos deben alinearse con los requisitos y la lógica definidos para el proyecto.

Tienen un formato y tipo de dato correctos

Se debe validar que los datos se ajustan al tipo de dato esperado, por ejemplo, números en columnas numéricas.

Son consistentes y únicos

Los datos deben ser consistentes entre sí y no deben incluir registros duplicados, a menos que sea apropiado para el análisis.

Son completos

Los datos esenciales no deben estar nulos o incompletos.

Criterios de descarte

Datos irrelevantes

Se eliminan los datos que no contribuyen a los objetivos del análisis, así como las columnas innecesarias.

Datos inconsistentes o erróneos

Se descartan los datos que contienen valores incorrectos o que no cumplen con los estándares de calidad.

Datos incompletos

Se pueden desechar registros que carecen de información esencial, especialmente si la ausencia de esta información puede invalidar el análisis posterior.

Datos duplicados

Si se detectan registros duplicados que no son válidos para el análisis, se descartan para mantener la unicidad de los datos.

Implementación práctica

Establecer reglas de validación claras

Definir las reglas de calidad y validación de los datos antes de comenzar el proceso ETL.

Implementar validaciones en cada etapa

Realizar controles de calidad en cada paso del proceso (extracción, transformación y carga) para detectar y corregir errores tempranamente.

Usar una estrategia de limpieza de datos

Realizar pasos de limpieza y transformación definidos para gestionar errores, valores nulos y datos innecesarios.

Implementación práctica

Documentar los procesos:

Mantener una documentación clara de los criterios y las transformaciones aplicadas para garantizar la transparencia y el mantenimiento del proceso.

Realizar pruebas:

Utilizar pruebas para validar que el proceso ETL está funcionando correctamente y que los datos finales cumplen con los estándares de calidad.

Tipo de columna	% de valores faltantes	Acción recomendada	Justificación
ID o clave primaria	> 0 %	❌ Descartar o revisar fuente	No se pueden imputar identificadores. Un solo nulo rompe la unicidad.
Campos críticos (fecha, monto, categoría obligatoria, etc.)	≤ 5 %	✅ Imputar (si posible) o dejar nulo	Pequeños porcentajes pueden corregirse sin sesgo.
	5 – 20 %	⚠️ Evaluar impacto; si el campo es necesario → imputar; si no → eliminar columna o registros.	Riesgo moderado de distorsión.
	> 20 %	❌ Descartar columna o registros según relevancia.	El campo deja de ser confiable.
Campos no críticos (comentarios, opcionales, etc.)	≤ 20 %	👍 Dejar nulos o imputar.	No afecta métricas clave.
	20 – 50 %	⚠️ Imputar si el campo es útil para modelos; sino eliminar.	Balance entre cobertura y ruido.
	> 50 %	❌ Eliminar columna.	No tiene suficiente información útil.

Tipo	% Valores faltantes	Acción recomendada	Nota / Impacto
Columnas numéricas	≤ 10 %	Imputar con media, mediana o KNN.	Bajo sesgo.
	> 10 – 30 %	Imputar con mediana o modelo predictivo si el campo es importante; si no, eliminar.	Evaluar impacto en correlaciones.
	> 30 %	Eliminar o marcar como "poco confiable".	Riesgo alto de distorsión.
Columnas categóricas	≤ 10 %	Imputar con moda o categoría "Desconocido".	Conserva estructura de clases.
	10 – 40 %	Crear categoría "Missing" explícita.	No sesga distribución.
	> 40 %	Eliminar columna o fusionarla con otras.	Exceso de información faltante.
Duplicados (filas idénticas)	≤ 1 %	Eliminar.	Normal en fuentes múltiples.
	> 1 – 5 %	Revisar origen; eliminar si no son eventos válidos.	Posible error de integración.
	> 5 %	Auditar fuente o redefinir claves de unicidad.	Alta probabilidad de error.