

Sesión 4: Calidad de los datos

En la era de los datos, tener información no es suficiente: lo que realmente agrega valor es tener datos confiables, relevantes y bien gestionados. La calidad de los datos no es solo una cuestión técnica; es un factor estratégico que afecta directamente la competitividad, la productividad y la capacidad de innovar de una organización.

 por Kibernet Capacitación S.A.

Objetivos y principios de la calidad de los datos

El objetivo principal es garantizar que los datos sean adecuados para su propósito. Esto implica que:

- Estén disponibles cuando se necesiten.
- Sean precisos, es decir, que representen correctamente la realidad.
- Sean relevantes para los usuarios que los consumen.
- Sean completos, sin omisiones críticas.
- Sean oportunos, actualizados y vigentes.

Una mala calidad de datos puede parecer un problema menor en el corto plazo, pero a mediano y largo plazo aumenta los costos, reduce la confianza del usuario y debilita las decisiones estratégicas.

Según un estudio de Gartner citado por Dataversity (2022), las organizaciones pierden un promedio de 12.9 millones de dólares al año debido a problemas de calidad de datos. Estas pérdidas están asociadas a ineficiencias operativas, errores en la toma de decisiones y deterioro de la experiencia del cliente.

Fuente: Dataversity – Understanding the Impact of Bad Data

Principios rectores de la calidad de datos

La calidad se puede construir si se definen principios claros y medibles. A continuación, se presentan los más relevantes:

Principio	Descripción práctica
Accuracy (Precisión)	Los datos deben reflejar fielmente la realidad. Un error tipográfico en el campo de salario o edad puede cambiar por completo un análisis.
Completeness (Compleitud)	Todos los campos críticos deben estar presentes. Por ejemplo, una dirección sin código postal no permite realizar envíos correctos.
Consistency (Consistencia)	El dato debe mantener el mismo valor en todas sus representaciones. Un cliente no puede tener dos nombres distintos en sistemas integrados.
Integrity (Integridad)	El dato debe mantener relaciones correctas dentro de la base. Ejemplo: no puede haber una venta sin cliente asignado.
Reasonableness (Razonabilidad)	Los datos deben tener sentido en su contexto. No es razonable que un cliente tenga 150 años o que un producto pese 0 kg.
Timeliness (Actualización)	El dato debe estar vigente y reflejar los cambios recientes. Una fecha de renovación de contrato desactualizada puede llevar a sanciones.
Uniqueness/Deduplication (Unicidad)	Cada entidad debe estar registrada una única vez. Duplicados afectan reportes y generan redundancia en campañas o costos.
Validity (Validez)	Los datos deben ajustarse a formatos y dominios predefinidos. Por ejemplo, el RUT chileno debe tener estructura y dígito verificador correctos.
Accessibility (Accesibilidad)	Los datos deben estar disponibles para quienes los necesitan, cuando los necesitan, según permisos definidos por gobernanza.

Reflexión: Si uno de estos principios se rompe, ¿en qué parte del proceso de negocio crees que se vería afectado primero?



Analogía de los datos como ingredientes

Piensa en los datos como los ingredientes de una receta. Aunque tengas una receta muy bien escrita (los procesos de negocio), si los ingredientes están en mal estado (los datos), el resultado final será un fracaso, sin importar la calidad de la receta.

Pirámide de la Jerarquía de datos

La pirámide de la jerarquía de datos muestra cómo los datos brutos se transforman en información, conocimiento y finalmente sabiduría. Cada nivel de la pirámide representa un mayor grado de refinamiento y valor para la organización.



Conceptos esenciales de la calidad de los datos

En la gestión moderna de datos, es fundamental comprender que la calidad no es una propiedad universal, sino contextual. Esto significa que un dato puede ser técnicamente correcto y, sin embargo, no aportar valor si no es pertinente para el análisis o la decisión que se requiere tomar.

Un dato es de calidad cuando satisface las expectativas del usuario dentro de un contexto de uso específico.

Por ello, el concepto de calidad de los datos se apoya en cuatro pilares esenciales: validez, completitud, confiabilidad y relevancia. Estos conceptos permiten detectar cuándo un dato, aunque aparentemente correcto, puede estar limitando la generación de valor en una organización.

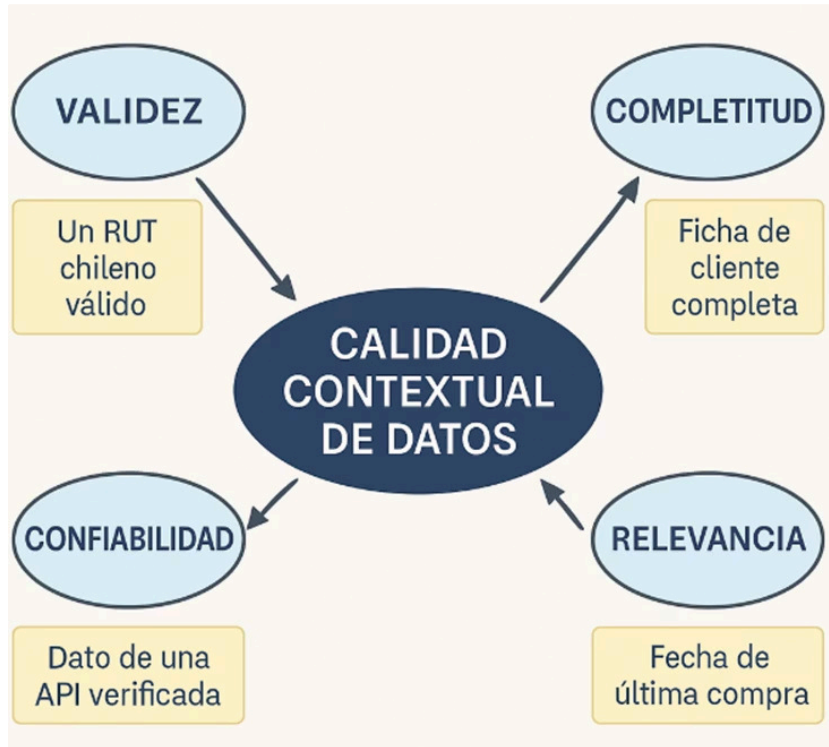
Idea para discusión: ¿Puede existir un dato correcto pero irrelevante?

Conceptos clave de calidad de datos

Concepto	Definición clara	Ejemplo práctico
Dato válido	Cumple con el formato, tipo de dato y estructura definida para ese campo.	Un RUT chileno que contiene el dígito verificador correcto y estructura esperada.
Dato completo	Todos los campos necesarios están presentes y no vacíos.	Una ficha de cliente con nombre, dirección, comuna y correo electrónico.
Dato confiable	Proviene de una fuente verificada, autorizada o con trazabilidad de origen.	Ingresado desde un sistema de autenticación o proveniente de una API validada.
Dato relevante	Tiene utilidad directa para los objetivos del análisis, operación o decisión.	El campo "fecha de última compra" es útil para segmentar promociones por inactividad.

Mapa conceptual de calidad de datos

Mapa conceptual centrado en la "Calidad contextual de los datos", destacando sus cuatro pilares fundamentales: validez, completitud, confiabilidad y relevancia.



Dimensiones de la calidad de los datos

En el contexto de la ingeniería de datos, hablar de calidad no es suficiente sin establecer formas concretas de evaluarla, diagnosticarla y mejorarla. Es aquí donde entran en juego las dimensiones de la calidad de datos.

¿Qué es una "dimensión" de calidad?

El término dimensión se refiere a un criterio o atributo específico que permite medir, analizar y comparar la calidad de los datos desde distintos ángulos. Según el marco DAMA-DMBOK (Data Management Body of Knowledge), una dimensión es una categoría que explica un aspecto evaluable del dato, y que puede ser expresado mediante indicadores o métricas.

Ejemplo: la "precisión" es una dimensión que permite verificar qué tan bien representa un dato la realidad que busca modelar.

Las dimensiones son especialmente útiles para establecer estándares organizacionales, construir dashboards de monitoreo, definir reglas de validación y aplicar procesos de gobernanza de datos.

Dimensiones fundamentales de la calidad de datos

Basado en fuentes como Talend, DAMA-DMBOK v2 y Data Management Association (DAMA), las siguientes son algunas de las dimensiones más utilizadas y reconocidas en la industria:

Dimensión	Definición operativa	Ejemplo práctico
Precisión (Accuracy)	Evalúa si el dato refleja con exactitud la realidad que representa.	Un RUT mal digitado altera la identidad del cliente.
Compleitud (Completeness)	Verifica si todos los datos obligatorios están presentes y no están vacíos.	Un registro sin comuna o dirección está incompleto.
Consistencia (Consistency)	Examina si los datos mantienen coherencia entre diferentes fuentes o sistemas.	El mismo cliente aparece con distinta comuna en el ERP y el CRM.
Actualización (Timeliness)	Mide si los datos están vigentes según su propósito y fecha de uso.	Un número telefónico desactualizado impide contactar al usuario.
Trazabilidad (Traceability)	Indica si es posible rastrear el origen, la modificación y el uso del dato.	Saber quién modificó un precio y cuándo se hizo.
Accesibilidad (Accessibility)	Revisa si los usuarios autorizados pueden acceder al dato necesario cuando lo requieren.	Un analista no puede acceder a las métricas clave por restricciones mal configuradas.

Estas dimensiones permiten al equipo de datos:

- Detectar errores sistemáticos o casos aislados.
- Priorizar acciones correctivas.
- Evaluar la eficacia de las herramientas de validación y limpieza.
- Alinear los datos con los objetivos del negocio y de los usuarios.

Además, cada dimensión puede ser convertida en una métrica cuantificable, lo que permite alimentar tableros de control, informes de calidad, y procesos de auditoría de datos.

Conexión entre principios rectores y dimensiones medibles

Aunque los términos "principios" y "dimensiones" puedan parecer similares, su función en el proceso de gestión de calidad de datos es distinta pero complementaria.

Principios rectores

Definen el "deber ser": representan los fundamentos éticos, estratégicos y conceptuales que guían la gestión del dato.

Ejemplo:

El principio de precisión nos dice que los datos deben reflejar fielmente la realidad.

→ La dimensión precisión se puede medir con la fórmula:

$$(\text{Registros sin errores} / \text{Total de registros}) \times 100.$$

Dimensiones

Son el "cómo medirlo": establecen criterios técnicos para evaluar cuantitativamente si esos principios se están cumpliendo en la práctica.

Introducción práctica: Explorando la calidad de datos con Pandas

Ahora vamos a dar un paso hacia la práctica, utilizando ejemplos simples con la librería Pandas. Si aún no la conoces, no te preocupes: más adelante la exploraremos en mayor profundidad, pero esta sección servirá como puente entre los conceptos de calidad de datos y su aplicación concreta con Python.

Para seguir este ejemplo necesitaras:

1. Descargar el archivo CSV

- Haz clic aquí para descargarlo:
 - [clientes_calidad_demo.csv](#)
- Guárdalo en una carpeta de trabajo, por ejemplo:
C:\Users\TuNombre\Documentos\clase_pandas o una carpeta en tu escritorio.

2. Crear un nuevo archivo Python

- Abre VSCode.
- Crea un nuevo archivo: analisis_demo.py.
- Guarda este archivo en la misma carpeta donde descargaste el CSV.

Configuración del entorno para análisis de datos

Primero, vamos a crear un entorno virtual. Esto nos permitirá trabajar de forma aislada con las librerías necesarias para el análisis.

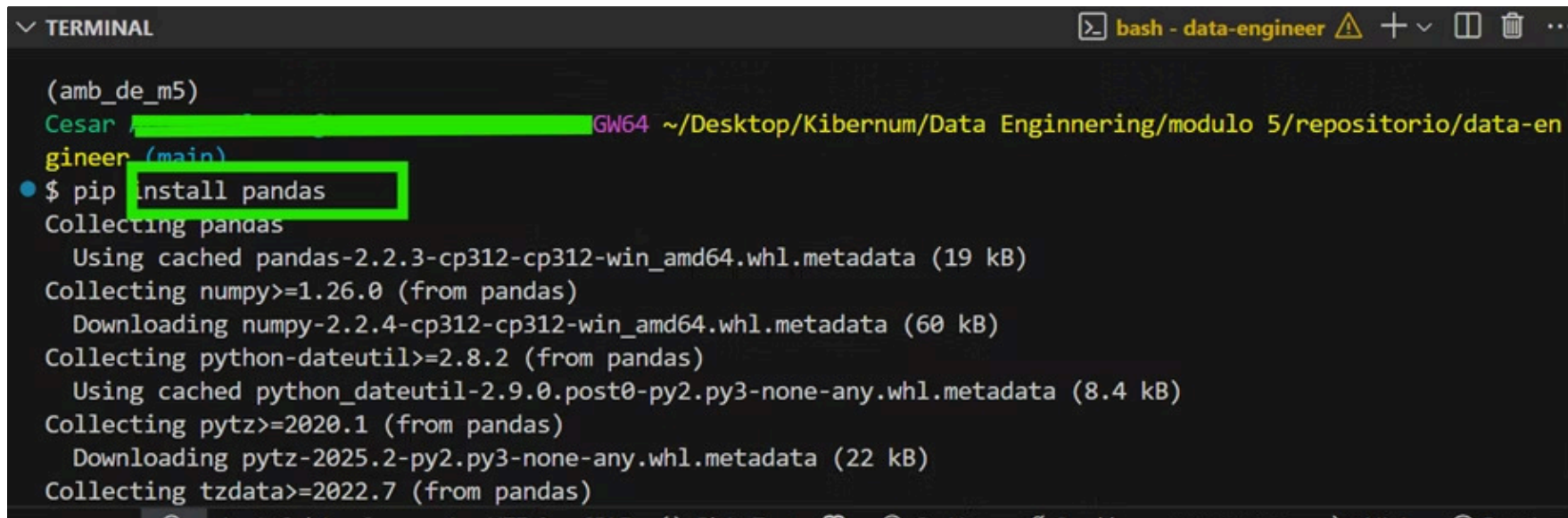
Los prints de pantallas son de Windows, Git Bash y virtualenvwrapper-win, puedes ejecutar:

```
mkvirtualenv amb_de_m5  
workon amb_de_m5
```

También puedes usar cualquier otro entorno que domines, como venv o Conda.

Instalamos Pandas con pip:

```
pip install pandas
```



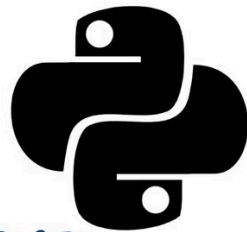
```
▼ TERMINAL bash - data-engineer [?] + - [ ] [ ] ...  
  
(amb_de_m5)  
Cesar [redacted] GW64 ~/Desktop/Kibernum/Data Enginnering/modulo 5/repositorio/data-en  
gineer (main)  
• $ pip install pandas  
Collecting pandas  
  Using cached pandas-2.2.3-cp312-cp312-win_amd64.whl.metadata (19 kB)  
Collecting numpy>=1.26.0 (from pandas)  
  Downloading numpy-2.2.4-cp312-cp312-win_amd64.whl.metadata (60 kB)  
Collecting python-dateutil>=2.8.2 (from pandas)  
  Using cached python_dateutil-2.9.0.post0-py2.py3-none-any.whl.metadata (8.4 kB)  
Collecting pytz>=2020.1 (from pandas)  
  Downloading pytz-2025.2-py2.py3-none-any.whl.metadata (22 kB)  
Collecting tzdata>=2022.7 (from pandas)
```

Play

with

csv files

using python



PANDAS



KODURI BHARGAV

Cargar un archivo CSV y comenzar el análisis

Una vez descargado el archivo `clientes_calidad_demo.csv`, crea un archivo `analisis_demo.py` en nuestro editor, en mi caso VSC y escribe lo siguiente:

```
analisis_demo.py M x clientes_calidad_demo.csv M
data-engineer > modulo-5 > analisis_demo.py > ...
1 import pandas as pd
2
3 df = pd.read_csv("clientes_calidad_demo.csv")
4 print(df.head(5))
5
```

Esto mostrará las primeras 5 filas del dataset y nos permitirá visualizar qué tipo de datos vamos a analizar.

```
✓ TERMINAL
(amb_de_m5)
MINGW64 ~/Desktop/Kibernum/Data Engi
$ py analisis_demo.py
      email      nombre  telefono  empresa  prioridad
0  sofia@empresa.cl    Sofia   912345678  EmpresaA         5.0
1   mario@correo     Mario  no definido  EmpresaB        11.0
2 catalina@empresa.com  Catalina  987654321  EmpresaC         7.0
3    ana@mail.com      NaN   965874123  EmpresaB        NaN
4 catalina@empresa.com  Catalina  987654321  EmpresaC         7.0
```

Ejemplos básicos de validación con Pandas

A continuación, exploramos algunos fragmentos de código que nos ayudan a diagnosticar problemas comunes en los datos:

Validar formato de correos electrónicos

La expresión regular permite verificar si los correos tienen un formato válido (usuario@dominio). El resultado se almacena en una nueva columna `email_valido`. Para evitar modificar el DataFrame original, se recomienda trabajar sobre una copia utilizando `.copy()`.

```
# Priemro creamos una copia patra no modificar el original
# Se recomienda crear una copia del DataFrame original para no modificarlo directamente
df_copia = df.copy()

|

# Verificar si el campo 'email' es un email válido
# Se considera un email válido si tiene al menos un carácter antes y después del '@' y al r
df_copia['email_valido'] = df_copia['email'].str.contains(r'^\S+@\S+\.\S+$', na=False)
```

Salida:

	email	nombre	telefono	empresa	prioridad	email_valido
0	sofia@empresa.cl	Sofia	912345678	EmpresaA	5.0	True
1	mario@correo	Mario	no definido	EmpresaB	11.0	False
2	catalina@empresa.com	Catalina	987654321	EmpresaC	7.0	True
3	ana@mail.com	NaN	965874123	EmpresaB	NaN	True
4	catalina@empresa.com	Catalina	987654321	EmpresaC	7.0	True

(amb_de_m5)

Medición de completitud y detección de duplicados

Medir la completitud por columna

```
completitud = df.notnull().mean() * 100  
print(completitud)
```

```
-----  
email          100.0  
nombre         80.0  
telefono       100.0  
empresa        100.0  
prioridad      80.0  
dtype: float64  
-----
```

Esta instrucción calcula el porcentaje de valores no nulos por columna, permitiendo identificar campos con datos faltantes o poco utilizados. Esta operación no modifica el DataFrame original, ya que crea un nuevo objeto llamado completitud.

¿Te resulta familiar? Esta lógica es comparable a una consulta en SQL que utiliza IS NOT NULL para evaluar la presencia de datos.

¿Qué otras similitudes encuentras entre Pandas y las consultas SQL?

A medida que avancemos en el curso, profundizaremos en estas y otras funciones de Pandas para automatizar tareas, validar datos y preparar conjuntos limpios para el análisis.

Detectar registros duplicados

```
duplicados = df.duplicated().sum()  
print(f"Duplicados detectados: {duplicados}")  
|
```

Cuenta cuántas filas del dataset están repetidas, lo cual puede afectar la unicidad y generar errores en reportes o métricas.

```
Duplicados detectados: 1  
(amb_de_m5)
```

Validar rango de prioridad

```
df_copia['prioridad_valida'] = df_copia['prioridad'].between(1, 10)  
print(df_copia)
```

Verifica si los valores de prioridad se encuentran dentro del rango esperado. Es una validación común para variables categóricas o niveles.

```
df_copia['prioridad_valida'] = df_copia['prioridad'].between(1, 10)
```


Causas comunes de problemas de calidad de datos

En la práctica, muchas organizaciones acumulan enormes volúmenes de datos que, a pesar de estar disponibles, no son confiables ni útiles. La calidad de los datos suele verse afectada por problemas que, en su mayoría, son evitables si se identifican a tiempo y se corrigen con buenas prácticas.

¿Por qué ocurren los problemas de calidad de datos?

A continuación, exploramos algunas de las causas más frecuentes, acompañadas de ejemplos reales.

Causa	Descripción	Ejemplo práctico
Ingreso manual incorrecto	Errores al digitar datos, muchas veces por falta de validación o capacitación insuficiente.	En un formulario, se ingresa 987654321@ como correo.
Duplicación de registros	Cuando un mismo dato se registra más de una vez, sin claves únicas ni controles adecuados.	Un cliente aparece dos veces con el mismo correo y teléfono, pero con nombres escritos de forma distinta.
Falta de estandarización	Datos que deberían tener un formato uniforme se registran de forma inconsistente.	Algunas fechas se ingresan como 10/01/2024, otras como 2024-01-10.
Datos obsoletos	Información que no ha sido actualizada y que ya no representa la situación actual.	El número de teléfono de un cliente ya no existe, pero sigue en la base de datos.
Integraciones defectuosas	Fallas en los procesos de integración entre sistemas generan datos incompletos o mal relacionados.	Al integrar dos sistemas distintos, las comunas no se cruzan correctamente por usar distintos nombres.

Puedes usar una herramienta como [draw.io](#) o [Canva](#) para crear un gráfico con ramas conectando causas técnicas y humanas con las consecuencias.

Conexión con la práctica

Una buena gestión de calidad de datos no solo requiere herramientas técnicas (como Pandas, SQL o ETL), sino también procesos, validaciones humanas y estándares bien definidos. Conocer estas causas te permite anticiparte a los errores antes de que impacten en los reportes, la toma de decisiones o incluso en la experiencia del cliente.



Principales características de la norma ISO 8000

La norma ISO 8000 es un estándar internacional desarrollado por la Organización Internacional de Normalización (ISO) para establecer buenas prácticas en la gestión de la calidad de los datos. Aporta un marco robusto para garantizar que los datos utilizados en procesos empresariales sean confiables, interoperables y reutilizables.

¿Qué busca ISO 8000?

La norma tiene como objetivo asegurar que los datos, especialmente los datos maestros (como clientes, productos, proveedores), cumplan criterios de precisión, completitud, consistencia y trazabilidad, de forma independiente de los sistemas tecnológicos donde se almacenen.

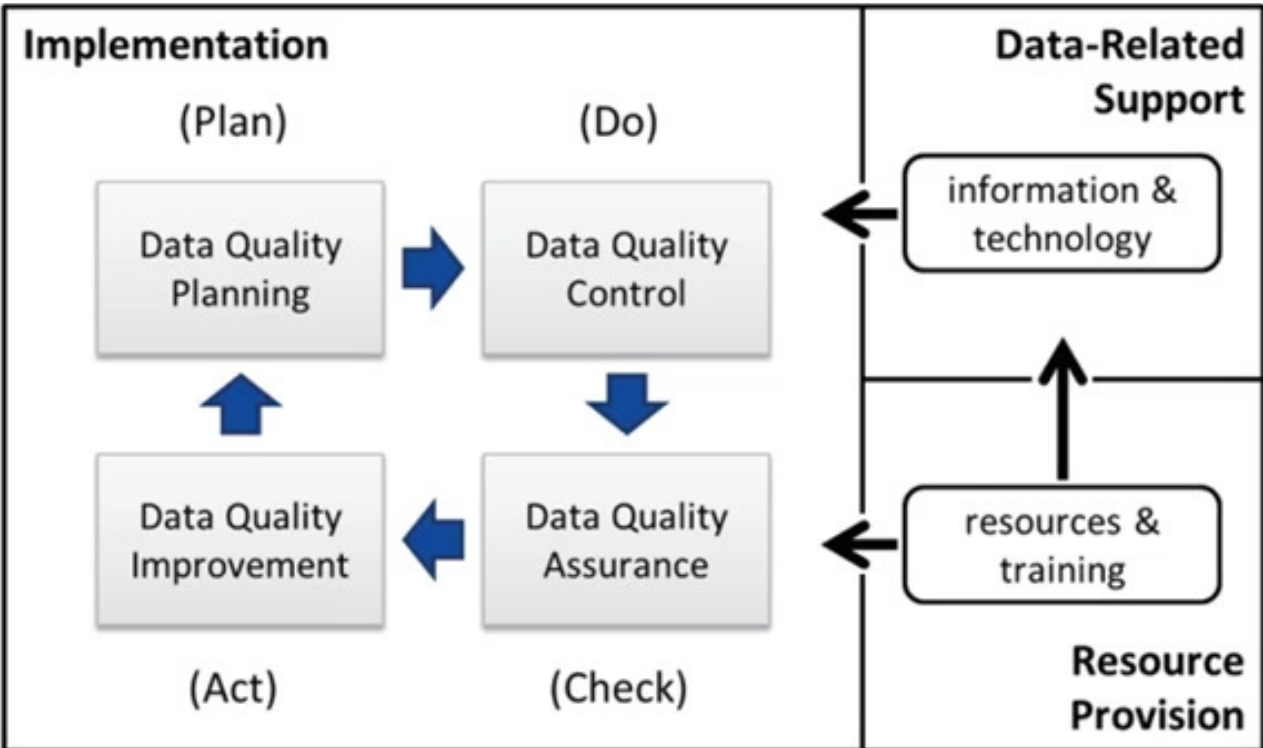
Característica	Descripción
Calidad como activo	Reconoce los datos como activos estratégicos, comparables con recursos financieros o humanos.
Interoperabilidad	Propone formatos y estructuras que permiten que los datos sean compartidos entre distintos sistemas.
Trazabilidad	Requiere poder rastrear el origen, los cambios y el uso de los datos a lo largo de su ciclo de vida.
Validación y consistencia	Exige procesos de verificación estructurados para mantener datos consistentes en toda la organización.
Estandarización de vocabularios	Promueve el uso de catálogos, taxonomías y reglas comunes para evitar ambigüedades.
Independencia tecnológica	La norma no está atada a herramientas específicas, lo que permite su adopción en diversos entornos.

¿Dónde aplica la ISO 8000?

- Gestión de inventarios.
- Información de productos (PIM).
- Directorios de clientes o proveedores.
- Datos de referencia en sistemas ERP o CRM.

Ejemplo: Si una empresa registra productos con nombres distintos en distintas áreas (ej. "Notebook HP i5" vs. "HP Intel Core i5"), incumple el principio de vocabulario estandarizado.

Estructura básica de la gestión de la calidad de los datos.



Evaluaciones de calidad de datos y métricas para monitoreo

Evaluaciones de calidad de datos (Data Quality Assessments)

Una evaluación de calidad de datos (o Data Quality Assessment) no es simplemente una revisión técnica o automatizada: es un proceso integral y estratégico, que busca diagnosticar el estado de los datos con base en criterios objetivos, conocimiento del negocio y buenas prácticas normativas, como la ISO 8000.

Una evaluación robusta debe considerar al menos tres pilares:

- Indicadores cuantitativos: Métricas concretas como porcentaje de completitud, duplicación, precisión o validez.
- Conocimiento del negocio: No basta con revisar formatos: es necesario entender qué datos son realmente críticos para los procesos operativos y decisionales.
- Cumplimiento normativo: Aplicación de estándares como ISO 8000, políticas internas de gobernanza de datos y buenas prácticas del sector.

Te compartimos el siguiente checklist de ejemplo:

Criterio a evaluar	Si Cumple	No Cumple	Comentario o hallazgo (Ejemplo)
¿Todos los campos críticos están completos?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Ej. 20% de direcciones faltan
¿Hay datos duplicados?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Se repiten 15 RUT de clientes
¿Los formatos cumplen lo esperado?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Correos inválidos detectados
¿Se puede rastrear el origen de los datos?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Falta historial de edición
¿Los datos están actualizados?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Teléfonos no vigentes

Checklist de ejemplo, puedes descargarlo [acá](#).

Medidas y métricas para monitoreo

Evaluar datos una sola vez no es suficiente: la calidad debe monitorearse de forma continua. Para ello, se utilizan métricas cuantificables, que permiten detectar desviaciones, aplicar mejoras y generar reportes de evolución.

Métrica	Fórmula sugerida	Interpretación
Tasa de completitud	(Campos no nulos / Total de campos) × 100	¿Qué porcentaje de campos obligatorios están llenos?
Tasa de duplicación	(Registros duplicados / Total registros) × 100	¿Qué tan frecuente se repite la misma entidad?
Tasa de precisión	(Registros correctos / Total registros) × 100	¿Qué tan bien representan los datos la realidad esperada?
Tasa de valores válidos	(Valores válidos / Total valores) × 100	¿Cuántos valores cumplen con su formato o dominio esperado?

Estas métricas pueden calcularse fácilmente con herramientas como Python (Pandas), SQL, Power BI o incluso con fórmulas en Excel, dependiendo del nivel técnico del equipo.

Técnicas y herramientas para medir calidad de datos

Cuando hablamos de asegurar la calidad de los datos, no basta con identificar errores: es fundamental actuar sobre ellos con herramientas adecuadas, según el contexto técnico y los recursos de la organización.

En esta sesión comprendimos que la calidad de los datos es un elemento clave para la toma de decisiones y la eficiencia organizacional. Exploramos los principios y dimensiones de calidad según la norma ISO 8000, identificamos causas comunes de errores y aprendimos a evaluarlos mediante métricas y herramientas prácticas, como Pandas, que utilizamos en un ejercicio aplicado.

Tabla comparativa: Técnicas y herramientas comunes

Técnica / Herramienta	Aplicación principal
Regex / Expresiones regulares	Validación de patrones y formatos en campos como RUT, correo, etc.
OpenRefine	Limpieza interactiva, estandarización de nombres, detección de outliers
Talend Data Quality (DQ)	Perfilado automático de datos, detección de duplicados y calidad por dimensiones
Pandas (Python)	Análisis y validación automatizada mediante código, ideal para integrarlo con otras tareas de ingeniería de datos
SQL	Verificación de reglas de integridad, búsquedas de valores nulos o inconsistentes, deduplicación

¿Por qué Pandas?

Durante la evaluación formativa de esta sesión, usaremos Pandas como la herramienta central para realizar validaciones reales sobre un conjunto de datos.

Pandas es una librería en Python muy popular entre analistas e ingenieros de datos, ya que permite:

- Cargar archivos estructurados (como CSV) fácilmente.
- Detectar errores como duplicados, nulos o valores fuera de rango.
- Generar métricas y visualizar resultados rápidamente.
- Aplicar transformaciones y dejar el dataset en condiciones óptimas para su uso.

Importante: Aunque trabajaremos con Pandas en este módulo, es fundamental comprender que existen múltiples herramientas y enfoques que permiten lograr los mismos objetivos: garantizar que nuestros datos sean confiables, útiles y de alta calidad.

En esta sesión comprendimos que la calidad de los datos es un elemento clave para la toma de decisiones y la eficiencia organizacional. Exploramos los principios y dimensiones de calidad según la norma ISO 8000, identificamos causas comunes de errores y aprendimos a evaluarlos mediante métricas y herramientas prácticas, como Pandas, que utilizamos en un ejercicio aplicado.