

Calidad de datos

Stratio

Alfonso Fernández Revenga

Curso ____ - Edición ____

Fecha 09/01/2021

SOBRE MI

Alfonso Fernández Revenga

<https://www.linkedin.com/in/alfonsofernandezrevenga/>

alfonsofernandez@stratio.com



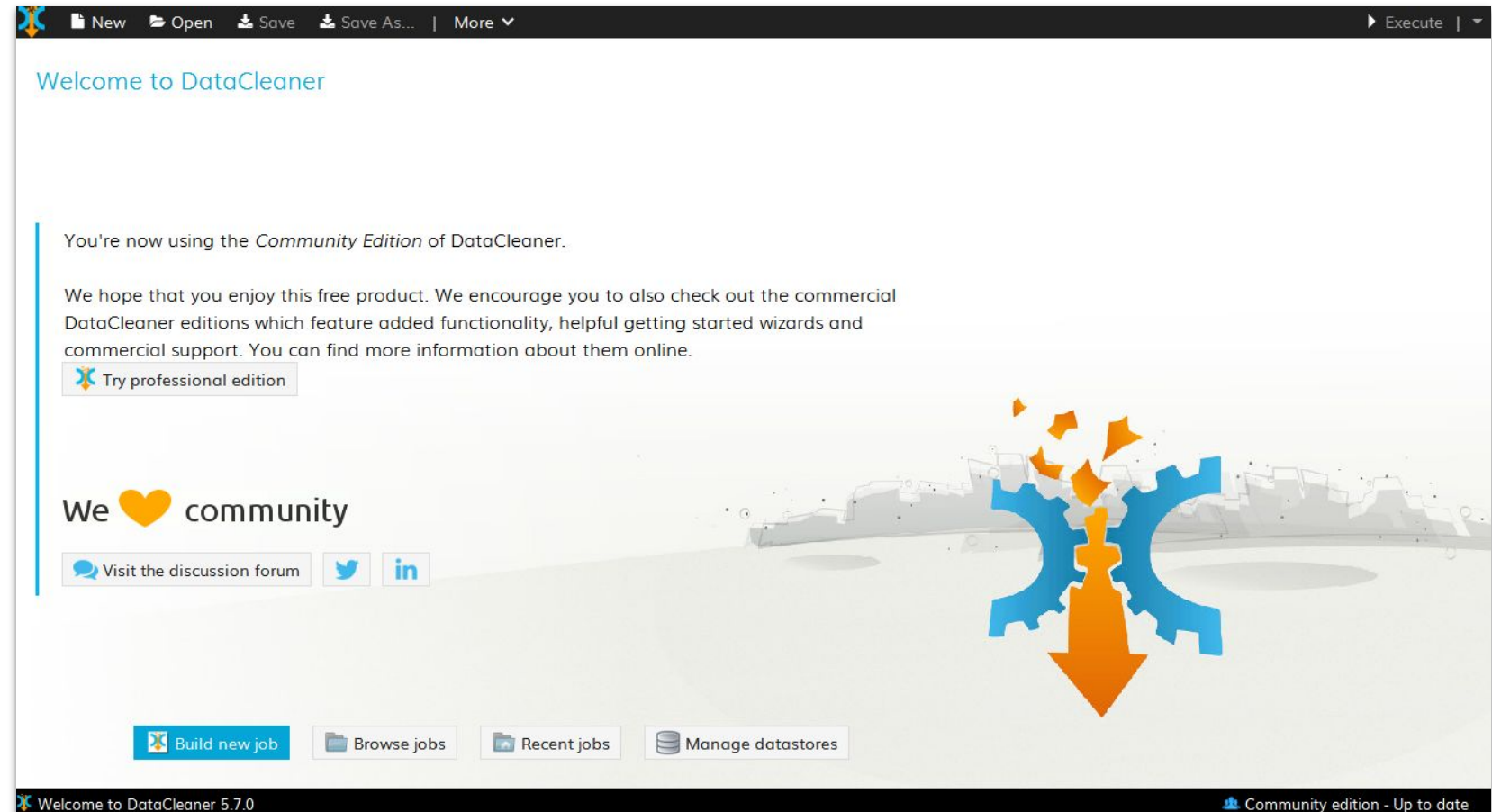
EDEM
Escuela de Empresarios



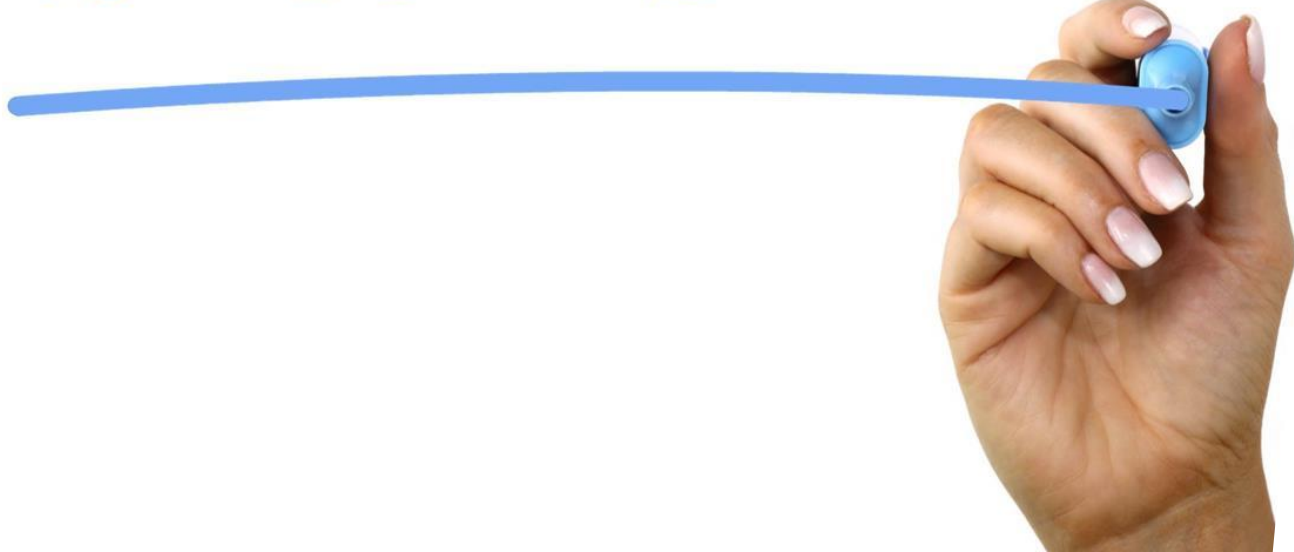
Caso práctico

Data Cleaner

<https://datacleaner.github.io/downloads>



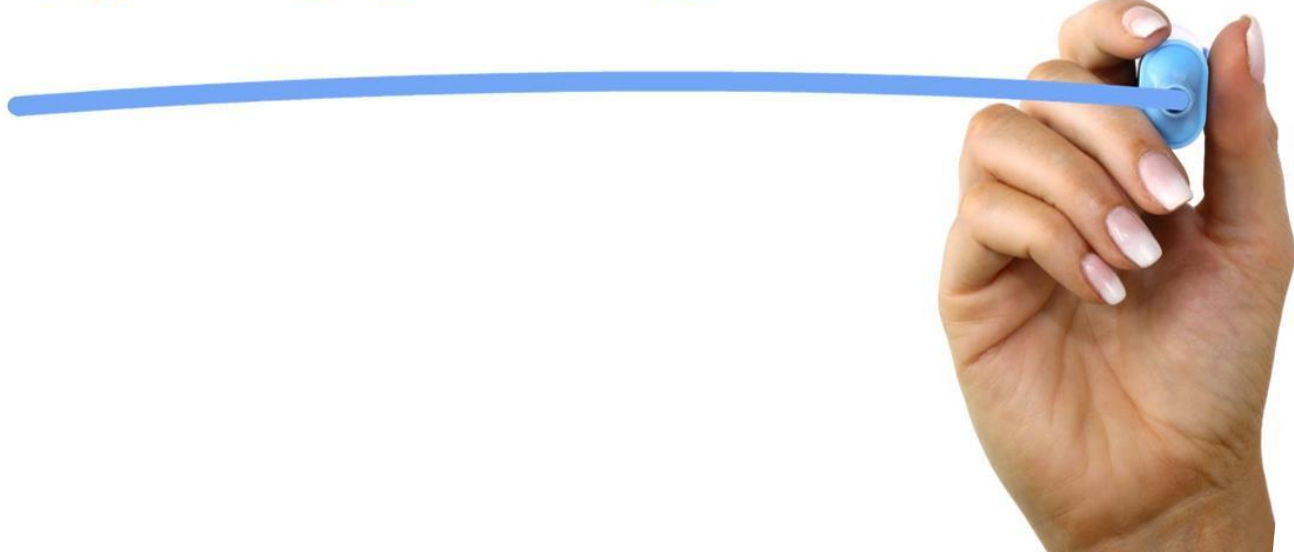
INDEX



Índice

1. Necesidades de calidad de datos
2. ¿Qué es calidad de datos?
3. Dimensiones de calidad
4. Ciclo de vida
5. Herramientas
6. Caso práctico
7. Conclusiones

INDEX



Índice

1. **Necesidades de calidad de datos**
2. ¿Qué es calidad de datos?
3. Dimensiones de calidad
4. Ciclo de vida
5. Herramientas
6. Caso práctico
7. Conclusiones

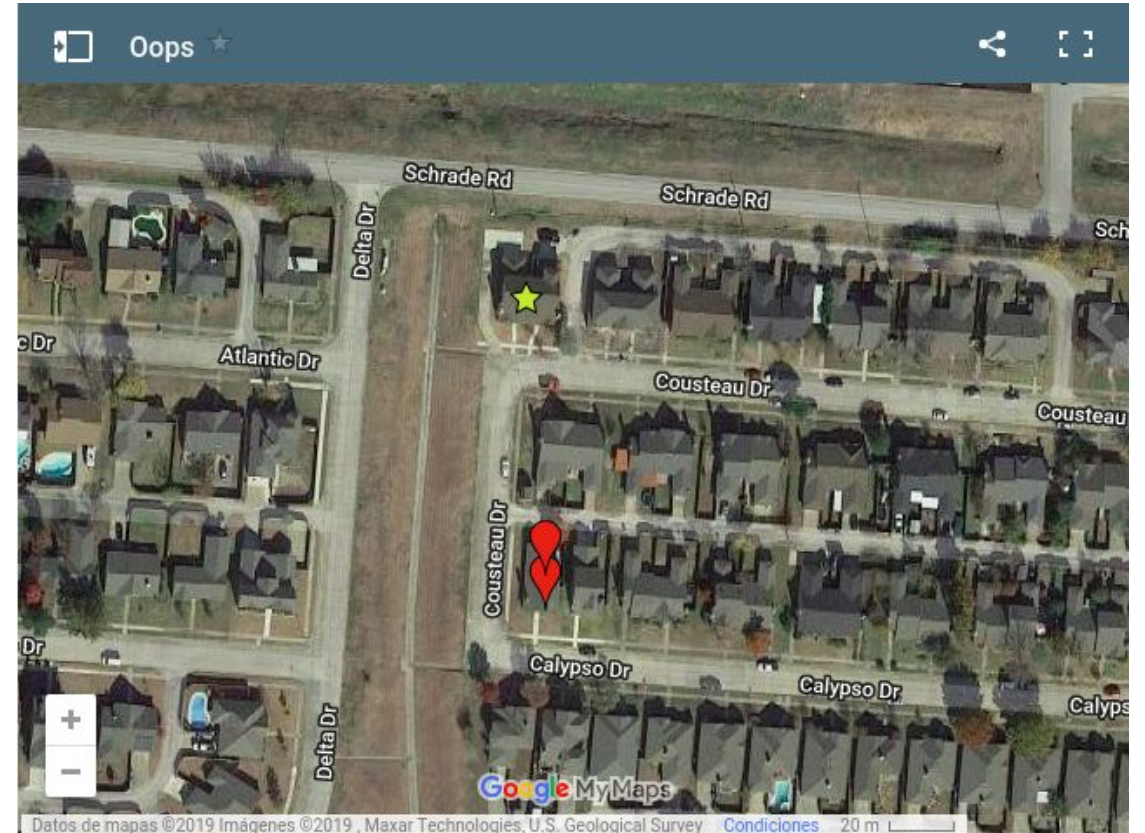
Necesidades de calidad de datos

Historias para no dormir...

La casa equivocada...

A principios de 2016, la casa de un hombre en Texas fue demolida debido a un error de GPS de Google Maps. Los trabajadores de demolición se habían basado en Google Maps para llevarlos a la dirección correcta, sin embargo Google tenía el lugar equivocado y la casa real era en realidad una cuadra de distancia.

¡Y lo que realmente da miedo es que esta no es la primera vez que esto sucede en Texas!



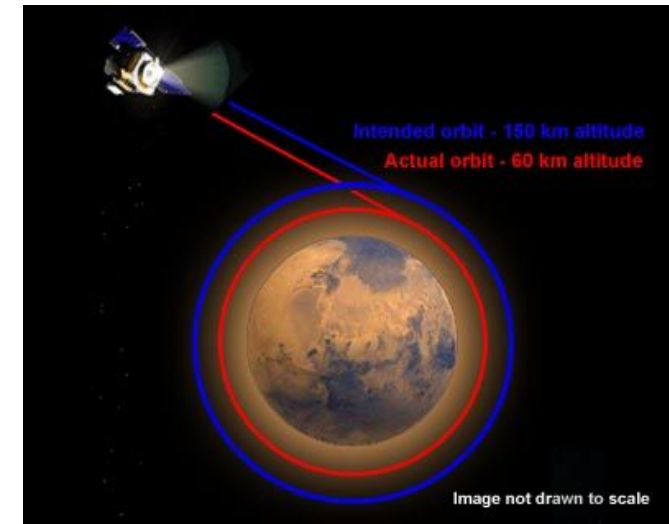
Necesidades de calidad de datos

Historias para no dormir...

Mars Climate

El Mars Climate Orbiter, una parte clave del programa de la NASA para explorar el planeta Marte, desapareció en septiembre de 1999 después de que se dispararon cohetes para llevarlo a la órbita del planeta. Más tarde, una junta de investigación descubrió que los ingenieros de la NASA no convirtieron las medidas inglesas de millas a kilómetros, y esa fue la raíz de la pérdida de la nave espacial. El orbitador se estrelló contra el planeta en lugar de alcanzar una órbita segura.

Esta discrepancia entre las dos medidas, que era relativamente pequeña, hizo que el orbitador se acercara a Marte a una altitud demasiado baja. El resultado fue la pérdida de una nave espacial de \$ 125 millones y un revés significativo en la capacidad de la NASA para explorar Marte.



Necesidades de calidad de datos

Historias para no dormir...

Fraude de correo CDs

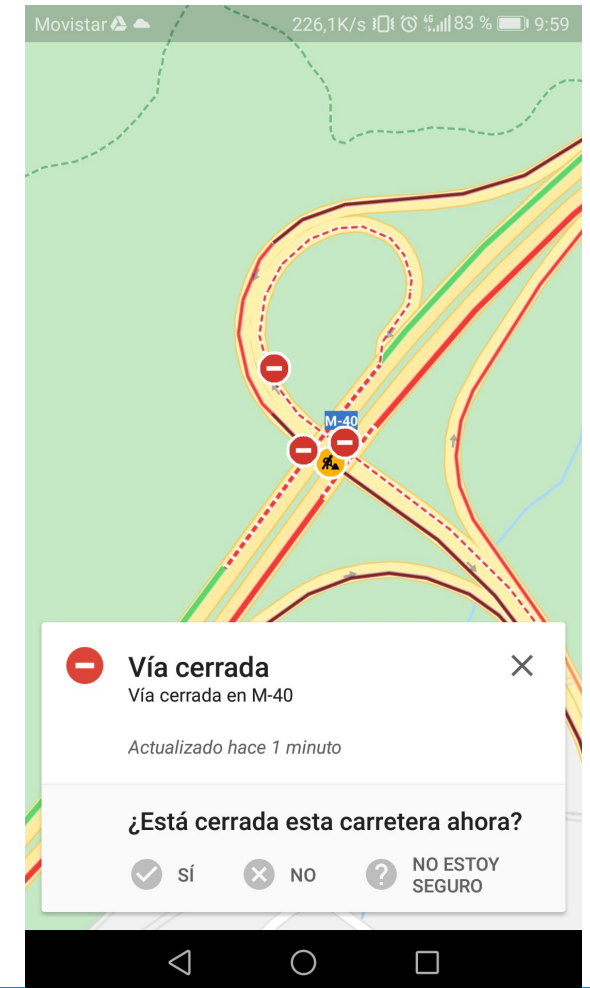
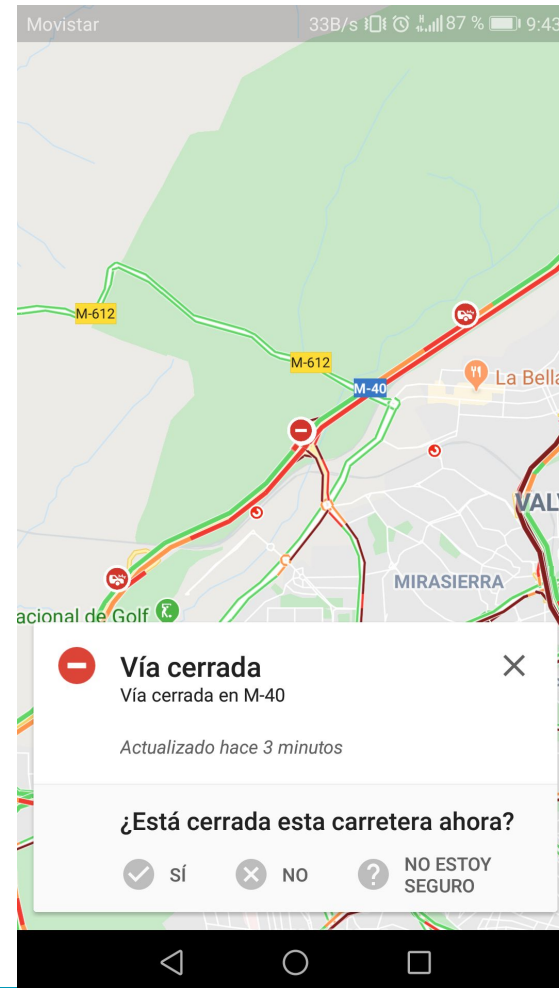
David Russo, de 33 años, de Sayreville, Nueva Jersey, admitió que recibió 22.260 CDs al hacer que cada dirección, incluso si figuraba en el mismo apartado postal, fuera lo suficientemente diferente como para evadir los programas informáticos de detección de fraude. Entre sus métodos: agregar números de apartamentos ficticios, abreviaturas de dirección innecesarias y signos de puntuación adicionales. Se cree que la estafa es la más grande de su tipo en la nación, dijo el fiscal federal adjunto Scott S. Christie, quien procesó el caso. La oferta proporcionaba nueve CDs gratuitos con la compra de un CD al precio normal, más gastos de envío. Russo pagó alrededor de \$ 56,000 en CD, dijo Paul B. Brickfield, su abogado, o un promedio de \$ 2.50 cada uno. Luego vendió los CD en los mercados por alrededor de \$ 10 cada uno (ganó unos 170.000\$).

Necesidades de calidad de datos

Historias para no dormir...

Google maps. Vivido en mis propias carnes.

Madrid, 17 de Octubre de 2019. Al ir a trabajar, Google me dice que la M40 está cortada por obras...



Necesidades de calidad de datos

Historias para no dormir...

Al no creermelo, recurro a la info del Ayuntamiento de Madrid. Lo que realmente estaba cortado era el ramal de acceso, pero NO la M40. Mucha gente llegó tarde simplemente por no confirmar la información.

Y yo tardé menos que otros días!



Necesidades de calidad de datos

Debate

¿De quién creéis que es la responsabilidad de estos problemas?

¿Cómo lo solventaríais?

Necesidades de calidad de datos

Debate

¿Y los datos de contagios de COVID?

¿Hay problema de calidad de datos?

24 de enero a las 15:30

Fecha actualización

1.418.726

Casos notificados últimos 14 días

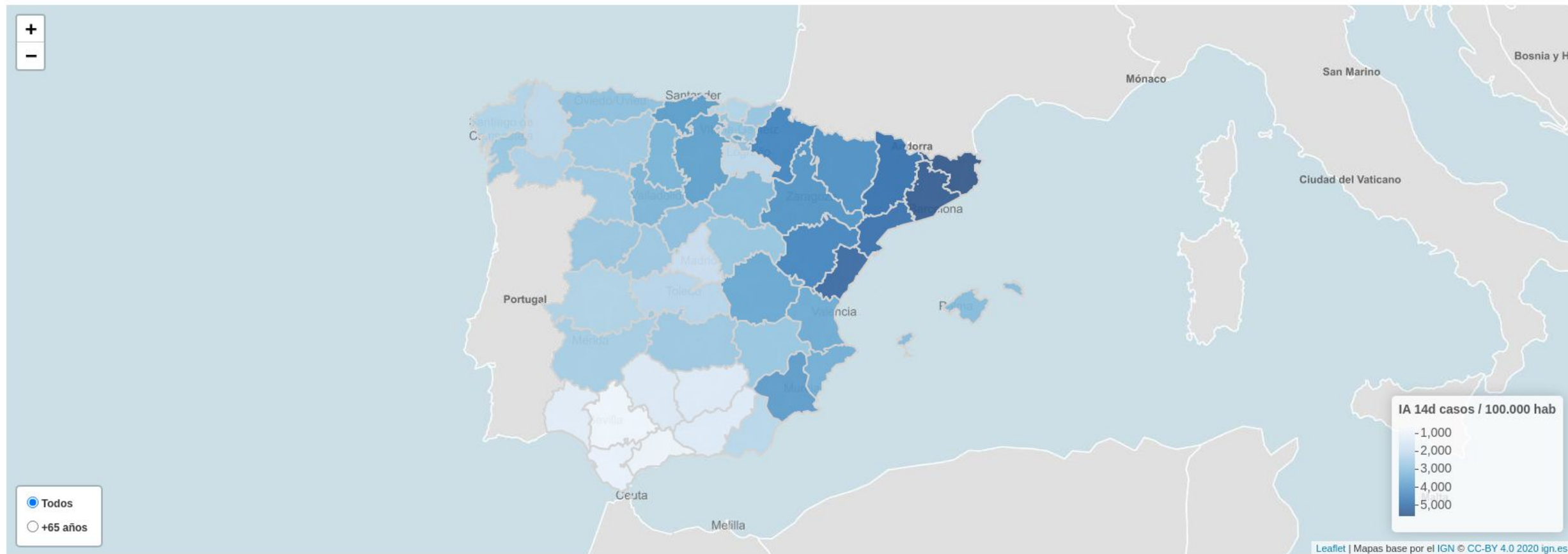
3017

Incidencia acumulada a 14 días / 100.000 hab

Incidencia acumulada a 14 días

Incidencia acumulada a 7 días

Razón de tasas



Mapa de incidencias acumuladas por provincia en los últimos 14 días (10 de enero a 23 de enero), para toda la población y para población de 65 y más años, calculadas a partir de los datos individualizados notificados a la RENAVE. Es importante resaltar que todos los resultados son provisionales y deben interpretarse con precaución porque se ofrece la información disponible en el momento de la extracción de datos. El número de casos y las IA cambian en cada actualización diaria del panel.

Necesidades de calidad de datos

Debate

Nota informativa en relación a la notificación de la información de vigilancia a la RENAVE

Los datos publicados en el Panel COVID-19 proceden de la **declaración individualizada** de casos COVID-19 a la Red Nacional de Vigilancia Epidemiológica (RENAVE) a través de la aplicación informática SiViEs.

La COVID-19 es una enfermedad de declaración obligatoria y, como tal, la responsabilidad de la notificación de acuerdo a los criterios establecidos en cada momento corresponde a los facultativos y servicios de Salud Pública de las comunidades autónomas que realizan esta notificación.

En SiViEs se contabilizan **todos los casos notificados, siguiendo la estrategia de vigilancia vigente en cada momento** (Estrategia de detección precoz, vigilancia y control de COVID-19 disponible en https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/COVID19_Estrategia_vigilancia_y_control_e_indicadores.pdf).

La información obtenida por la vigilancia epidemiológica es distinta a cualquier información que se obtiene con fines estadísticos. Se recoge y usa, en tiempo real, para la toma de decisiones en el ámbito de la salud pública. Por este motivo puede ser incompleta, contener errores y sufrir retrasos en distinta medida. En definitiva, se precisa de un tiempo para su depuración y consolidación. Para elaborar los informes publicados por el ISCIII se realizan procesos de depuración, imputación y relación con otras bases de datos secundarias para paliar los defectos de la notificación. Los análisis de una base extraída un día puedan variar de los extraídos al día siguiente si se ha actualizado la información. Por este motivo, se considera que la base de datos ofrecida semanalmente no está consolidada y los resultados extraídos de ella deben ser tratados con cautela.

Es importante resaltar que todos los resultados son provisionales y deben interpretarse con precaución porque se ofrece la información disponible en el momento de la extracción de datos. En los casos en los que no se haya notificado la provincia de residencia, ésta aparecerá en blanco. Es por tanto posible que no coincida la suma de los casos notificados por provincias con el dato de la Comunidad Autónoma al que pertenecen.

El número de casos de enero de 2020 oscila diariamente debido a errores en la inserción manual de datos de vigilancia en la plataforma SiViEs. Se revisan y actualizan conforme se confirman por las CCAA.

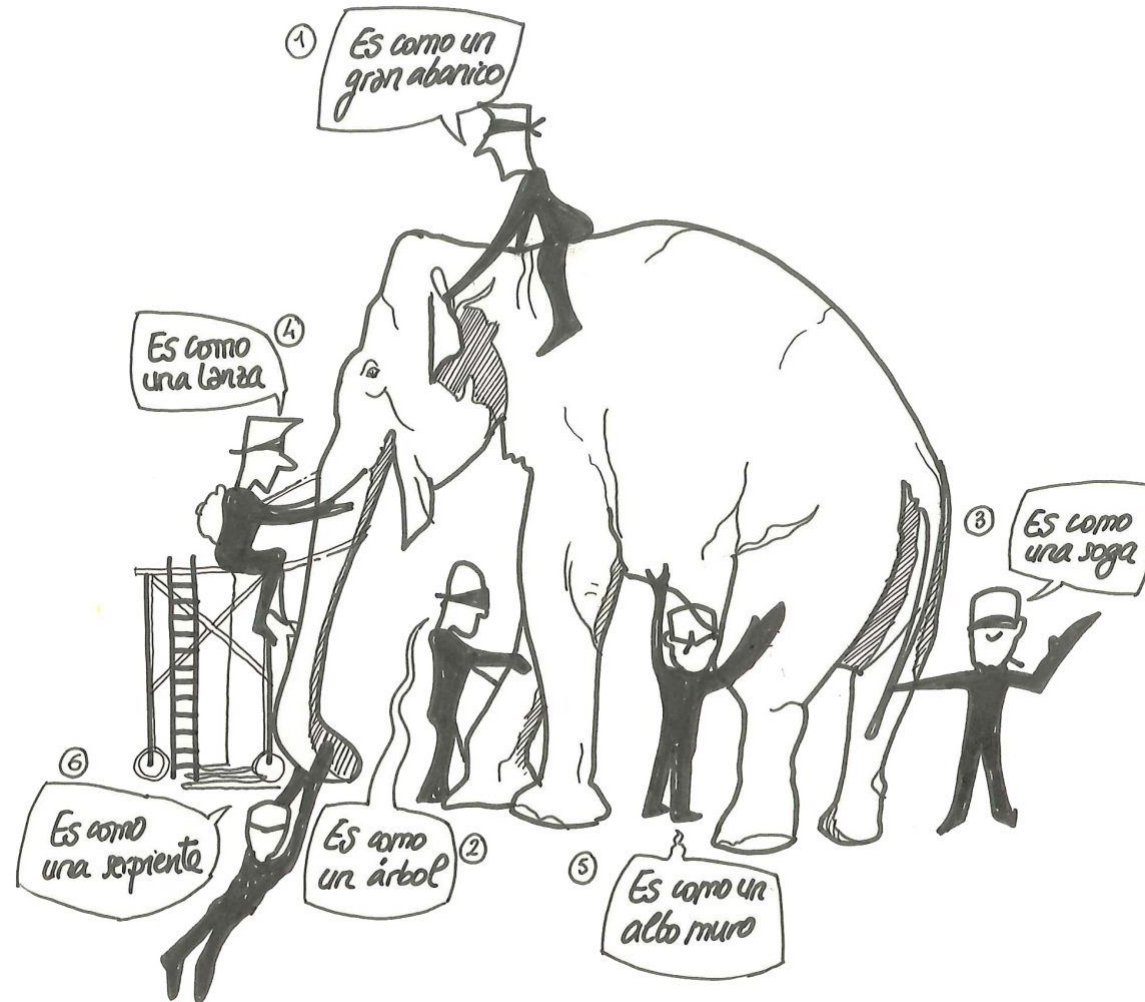
Todo ello debe tenerse en cuenta a la hora de interpretar los datos globales que se ofrecen en este Panel COVID-19.

Fuente: <https://cnecovid.isciii.es/covid19/>

→ Idoneidad.... luego lo veremos.

Necesidades de calidad de datos

Los 6 sabios ciegos y el elefante



Necesidades de calidad de datos

No existe una concepción universal de la calidad de los datos, sino que compiten muchas perspectivas diferentes

Problema:

- La mayoría de las organizaciones abordan los problemas de calidad de datos de la misma manera que los ciegos se acercaron al elefante: las **personas tienden a ver sólo los datos que están frente a ellos**.
- Poca **cooperación** a través de las fronteras, al igual que los ciegos no pudieron transmitir sus impresiones sobre el elefante para **reconocer a toda la entidad**.
- Esto conduce a la **confusión, disputas y opiniones** enfrentadas.

Solución:

- La gestión de calidad de datos puede ayudar a lograr una **imagen más completa** y facilitar las comunicaciones transfronterizas

Necesidades de calidad de datos

Verdadero o falso

- ? Puedes **arreglar** los datos
- ? La calidad de los datos es un **problema de IT**.
- ? El problema está en las **fuentes** de datos o en la **entrada** de datos.
- ? Los data warehouse proporcionarán una **versión única** de la verdad.
- ? El nuevo sistema proporcionará una **versión única** de la verdad.
- ? La **estandarización eliminará el problema** de las diferentes "verdades" representadas en los informes o en los análisis de datos

Necesidades de calidad de datos

Primeros malentendidos...

- X** Puedes **arreglar** los datos
- X** La calidad de los datos es un **problema de IT**.
- X** El problema está en las **fuentes** de datos o en la **entrada** de datos.
- X** Los data warehouse proporcionarán una **versión única** de la verdad.
- X** El nuevo sistema proporcionará una **versión única** de la verdad.
- X** La **estandarización eliminará el problema** de las diferentes "verdades" representadas en los informes o en los análisis de datos

Necesidades de calidad de datos

Y si es tan importante...

Pregunta: ¿Por qué las organizaciones no han adoptado un enfoque más proactivo en la calidad de los datos?

Necesidades de calidad de datos

Y si es tan importante...

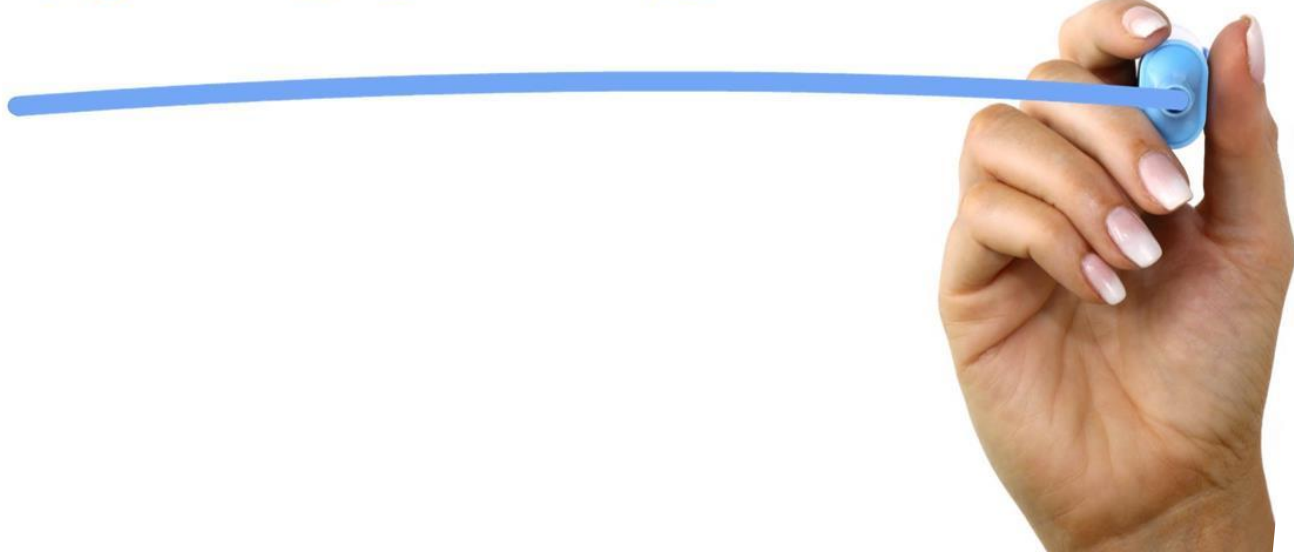
Pregunta: ¿Por qué las organizaciones no han adoptado un enfoque más proactivo en la calidad de los datos?

Respuesta:

- Arreglar problemas de calidad de datos no es fácil
- Es peligroso (vendrán por ti!!!)
- Es probable que sus esfuerzos sean mal entendidos
- Podrías empeorar las cosas (o eso piensan)
- Un solo problema de calidad de datos puede convertirse en una inversión significativa e inesperada
- Es complicado “vender” a negocio el beneficio de invertir en mejorar la calidad de datos

El grupo de trabajo de **calidad de datos** de **DAMA España** está trabajando en una línea para poder convencer a las direcciones de las empresas para invertir en calidad de datos. Primer artículo [aquí](#).

INDEX

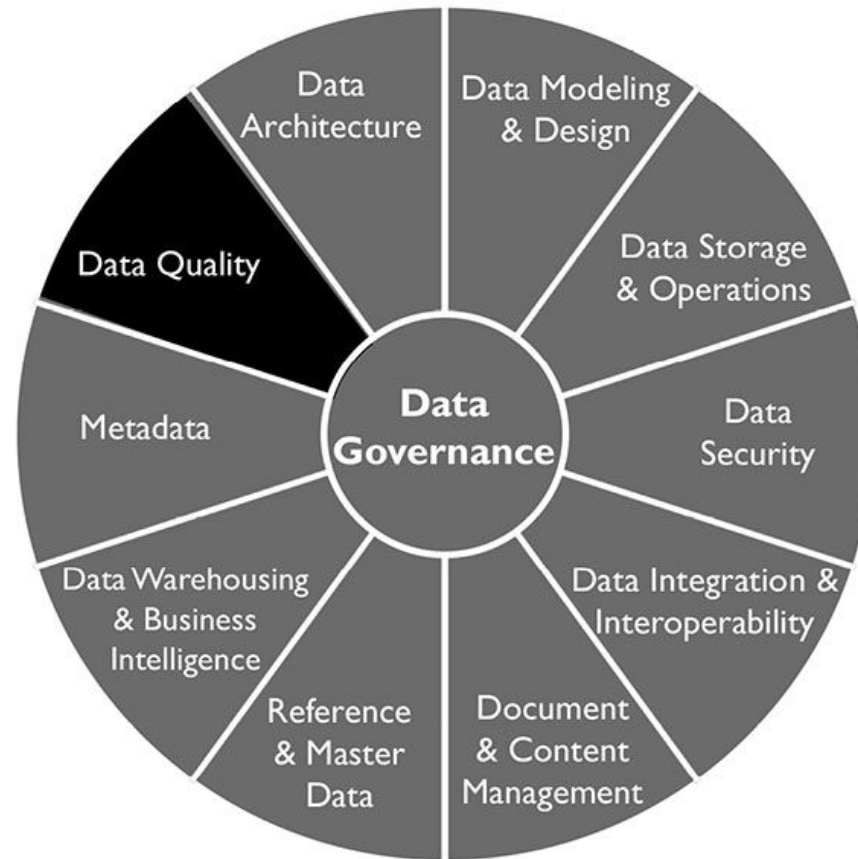


Índice

1. Necesidades de calidad de datos
2. **¿Qué es calidad de datos?**
3. Dimensiones de calidad
4. Ciclo de vida
5. Herramientas
6. Caso práctico
7. Conclusiones

¿Qué es calidad de datos?

DAMA - Data Management Body Of Knowledge



DAMA-DMBOK2 Data Management Framework

Copyright © 2017 by DAMA International

¿Qué es calidad de datos?

¿Qué es Data Quality Management?

La **mala gestión** de la calidad de los datos **no equivale** a una **mala calidad** de los datos... es peor!!!

Cuando no tienes una buena gestión de calidad de datos ...

- El nivel actual de calidad de datos será desconocido
- Mantener un nivel suficiente de calidad de datos será el resultado de la improvisación y propenso a errores
- El riesgo para el negocio aumentará con el tiempo

Es infinitamente más sensato garantizar una buena calidad de datos mediante una buena gestión a través de un conjunto coherente de políticas, estándares, procesos y tecnología de soporte.

¿Qué es calidad de datos?

¿Qué es Data Quality Management?

“Los errores en los datos pueden costar a una empresa millones. Y alienar a los clientes, proveedores y socios comerciales, para implementar nuevas estrategias puede que sea difícil o incluso imposible.

La existencia misma de la organización puede verse amenazada por datos deficientes ”

Joe Peppard – European School of Management and Technology

“La mala calidad de los datos es como suciedad en el parabrisas. Es posible que pueda conducir durante mucho tiempo con una visión que se degrada lentamente, pero en algún momento debe detenerse y limpiar el parabrisas o arriesgarlo todo”

Ken Orr, The Cutter Consortium



Gestión de Calidad de Datos

Definición: La planificación, implementación, y actividades de control que aplican técnicas de gestión de calidad a los datos, en orden de asegurar que sean aptos para su consumo y satisfagan las necesidades de consumidores de datos.

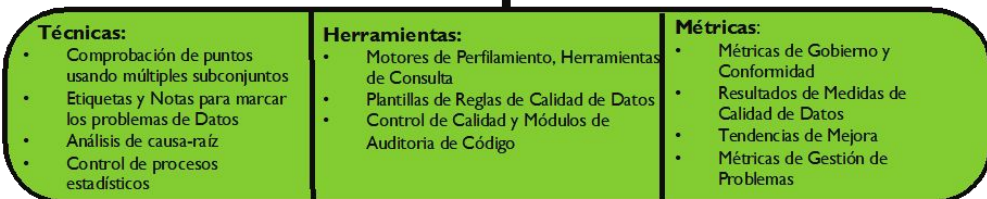
Metas:

1. Desarrollar un enfoque gobernado para que los datos cumplan con su propósito basado en los requerimientos del consumidor de datos.
2. Definir estándares, requerimientos, y especificaciones para el control de calidad de datos como parte del ciclo de vida de los datos.
3. Definir e implementar procesos para medir, monitorear y reportar los niveles de calidad de datos.
4. Identificar y abogar por oportunidades para mejorar la calidad de datos, a través de mejoras de procesos y sistemas.

Motivadores de Negocio



Motivadores Técnicos



(P) Planificación, (C) Control, (D) Desarrollo, (O) Operaciones

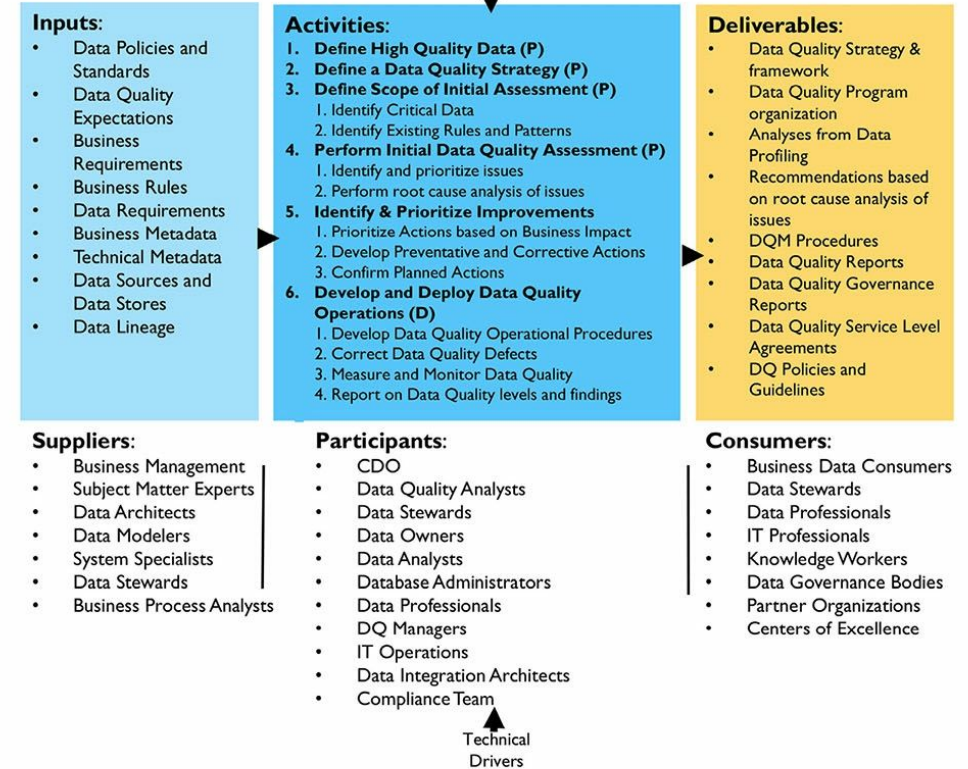
Data Quality Management

Definition: The planning, implementation, and control of activities that apply quality management techniques to data, in order to assure it is fit for consumption and meets the needs of data consumers.

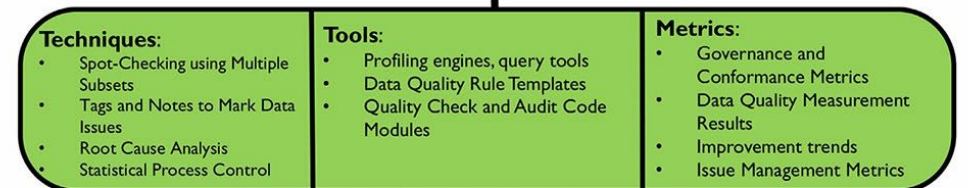
Goals:

1. Develop a governed approach to make data fit for purpose based on data consumers' requirements.
2. Define standards, requirements, and specifications for data quality controls as part of the data lifecycle.
3. Define and implement processes to measure, monitor, and report on data quality levels.
4. Identify and advocate for opportunities to improve the quality of data, through process and system improvements.

Business Drivers



Technical Drivers



(P) Planning, (C) Control, (D) Development, (O) Operations

Figure 91 Context Diagram: Data Quality

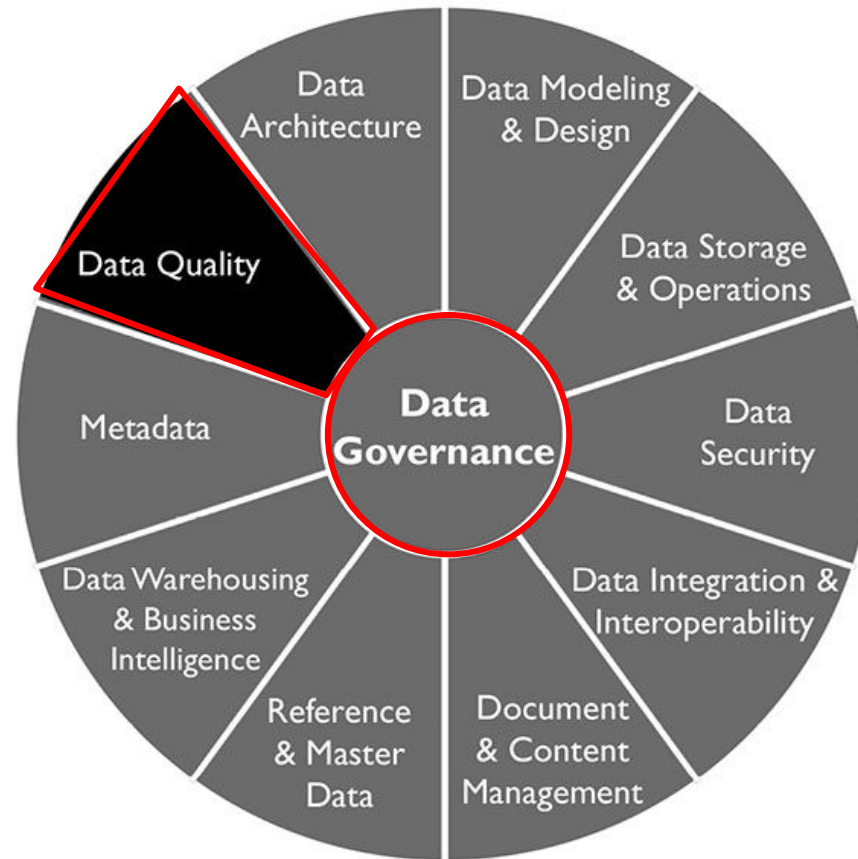
¿Qué es calidad de datos?

Desarrollar y promover el conocimiento de calidad de datos

1. **Promover y evangelizar** la importancia de la calidad de los datos lo antes posible, mejorará las posibilidades de éxito de cualquier programa de calidad de datos.
2. Esto debe suceder en **todos los niveles** dentro de la organización, desde la alta gerencia y las partes interesadas clave, hasta los usuarios y el personal operativo
3. Establecer una comunidad de interés de calidad de datos puede ayudar a crear un entendimiento común y proporcionar un foro para compartir conocimientos y mejores prácticas
4. La gestión de la calidad de los datos no puede sobrevivir sin la propiedad y la responsabilidad, por lo que debe estar alineado con la estrategia de **gobierno de datos**

¿Qué es calidad de datos?

El vínculo entre Gobierno y Calidad



DAMA-DMBOK2 Data Management Framework

Copyright © 2017 by DAMA International

¿Qué es calidad de datos?

El Gobierno de datos es clave

- **Involucrar** a socios comerciales que trabajarán con el equipo de calidad de datos y defenderán el programa **DQM** (Data Quality Management)
- Identificar las **funciones y responsabilidades** de ownership de datos, incluidos los miembros de la junta de gobierno de datos y los administradores (stewards)
- Identificar áreas clave de calidad de datos para abordar y **directivas** para la organización en torno a estas áreas clave
- Asignar responsabilidades por elementos de **datos críticos** y DQM
- Sincronice los elementos de datos utilizados en todas las líneas de negocio y proporciona **definiciones claras y sin ambigüedades**, uso de dominios de valor y reglas de calidad de datos.
- Informar continuamente sobre los **niveles** de calidad de datos (medir y monitorizar).
- Introducir los conceptos de análisis de **requisitos de datos** como parte del ciclo de vida general del desarrollo del sistema.

¿Qué es calidad de datos?

Beneficio e impacto

Beneficios de la buena calidad

- Adhesión a los actos **corporativos** y **regulatorios**.
- Mayor **confianza** en los datos.
- Reducción del "**busy work**" en arqueología de datos. (muerte a la excel de datos!!!!)
- **Satisfacción del cliente** enriquecida
- Mejor toma de **decisiones**.
- Marketing y publicidad efectivos
- Eficiencias de **costes**
- Operativa mejorada

Impacto de la mala calidad

- Publicidad y marketing ineficaces
- **Daño reputacional**
- Cumplimiento **regulatorio** disminuido (multas!)
- Disminución de la **satisfacción del cliente**.
- Procesos comerciales no económicos
- Salud, **seguridad** y protección comprometidas
- Inteligencia comercial errática
- Riesgo corporativo amplificado
- Agilidad empresarial deteriorada

¿Qué es calidad de datos?

Beneficio e impacto (debate)

Ahora que veis los beneficios y sobretudo el impacto, ¿habéis sufrido alguna vez alguna consecuencia de mala calidad?

¿Cómo creéis que se hubiese arreglado?

¿Qué es calidad de datos?

Definir requisitos de calidad de datos

Requisitos

- La calidad de los datos sólo puede considerarse dentro del contexto del uso previsto de los datos, es decir, **la calidad idónea para el propósito de uso**. → **Fit for purpose**
- El **nivel** requerido de **calidad** de datos para un componente en particular **depende**, por lo tanto, de la recopilación de **procesos** que interactúan con el componente
- Estos a su vez están impulsados por las políticas de negocio subyacentes, que en última instancia son la fuente de muchos requisitos de calidad de datos
- Determinar la **idoneidad para el propósito** requiere informar sobre métricas significativas asociadas con **dimensiones de calidad de datos bien definidas**.

Acciones

- ✓ Identificar los **componentes clave** de datos asociados con las políticas de negocio.
- ✓ Determinar cómo las “**afirmaciones**” de datos identificadas **afectan** el negocio
- ✓ Evaluar cómo se **clasifican los errores** de datos dentro de un conjunto de dimensiones de calidad de datos. Niveles de alerta.
- ✓ Proporcionar un medio para implementar procesos de **medición** que evalúen la **conformidad** con esas reglas de negocio

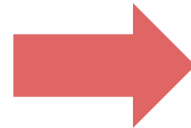
¿Qué es calidad de datos?

Ejemplos de idoneidad

En febrero de 2011, el gobierno del Reino Unido lanzó un sitio web de **mapeo de delitos** para Inglaterra y Gales (www.police.uk).

Desafortunadamente, por varias razones, el código postal asignado a un incidente policial específico **no siempre se correspondía** con la ubicación precisa del crimen.

El resultado fue que la poca precisión en el registro de información geográfica llevó a muchas calles residenciales tranquilas a ser **identificadas incorrectamente** como puntos críticos de delincuencia.



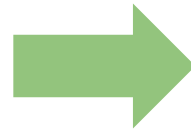
¿Qué es calidad de datos?

Ejemplos de idoneidad

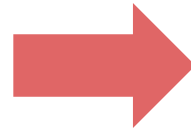
En febrero de 2011, el gobierno del Reino Unido lanzó un sitio web de **mapeo de delitos** para Inglaterra y Gales (www.police.uk).

Desafortunadamente, por varias razones, el código postal asignado a un incidente policial específico **no siempre se correspondía** con la ubicación precisa del crimen.

El resultado fue que la poca precisión en el registro de información geográfica llevó a muchas calles residenciales tranquilas a ser **identificadas incorrectamente** como puntos críticos de delincuencia.



En el contexto de la creación de **estadísticas agregadas** para evaluar las tasas de delincuencia relativa entre los condados, la calidad de los datos es perfectamente aceptable.



Sin embargo, si una compañía de seguros utiliza los mismos datos, hay un **problema** para los propietarios que reciben primas de seguro de vivienda infladas.

¿Qué es calidad de datos?

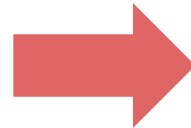
Ejemplos de idoneidad

Una compañía farmacéutica tenía cinco **centros principales de fabricación** en el Reino Unido, cada uno con su propio almacén de **repuestos** para las máquinas en las fábricas.



En teoría, los cinco sitios compartían un **sistema común**, por lo que las piezas de repuesto (65,000 artículos de inventario en total) se podían pedir desde otra ubicación.

Pero en realidad, el sistema era **difícil de usar**, por lo que cada uno de los sitios separados construyó su propio inventario de repuestos suficientes para sus necesidades. Más que suficiente, de hecho: después de una limpieza de datos, se descubrió que la compañía tenía suficientes repuestos para durar 90 años en algunos casos



¿Qué es calidad de datos?

Ejemplos de idoneidad

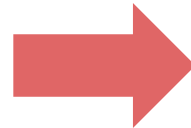
Una compañía farmacéutica tenía cinco **centros principales de fabricación** en el Reino Unido, cada uno con su propio almacén de **repuestos** para las máquinas en las fábricas.

En teoría, los cinco sitios compartían un **sistema común**, por lo que las piezas de repuesto (65,000 artículos de inventario en total) se podían pedir desde otra ubicación.

Pero en realidad, el sistema era **difícil de usar**, por lo que cada uno de los sitios separados construyó su propio inventario de repuestos suficientes para sus necesidades. Más que suficiente, de hecho: después de una limpieza de datos, se descubrió que la compañía tenía suficientes repuestos para durar 90 años en algunos casos

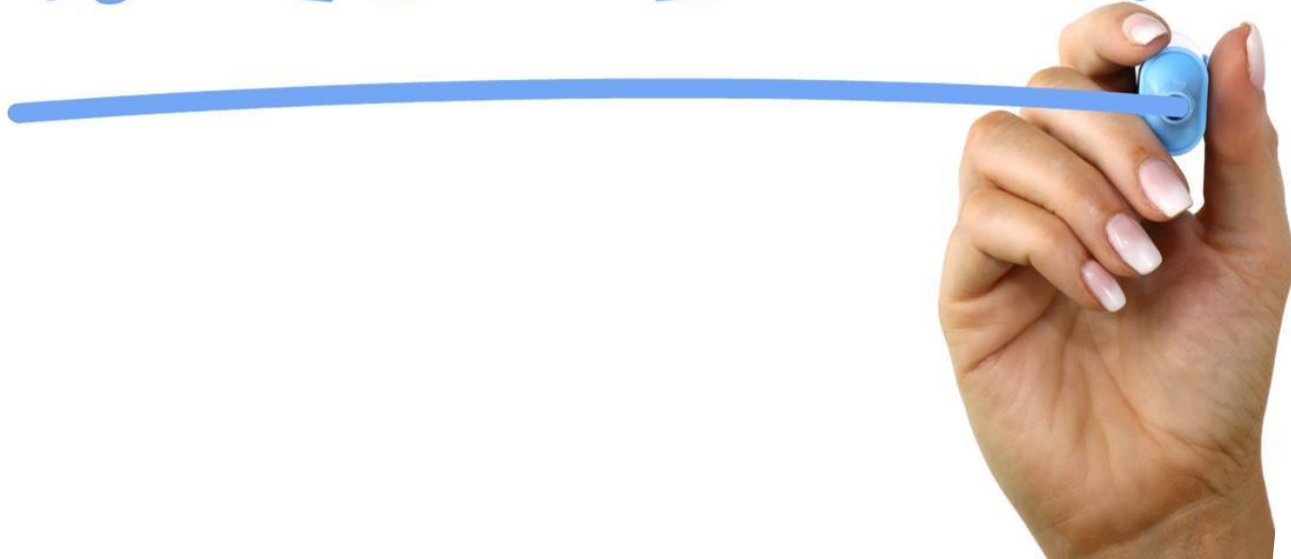


En el contexto de la gestión del riesgo de tiempo de inactividad de la máquina, esto es aceptable.



Sin embargo, con la visión holística del costo de las piezas de repuesto, esto es ridículo.

INDEX



Índice

1. Necesidades de calidad de datos
2. ¿Qué es calidad de datos?
- 3. Dimensiones de calidad**
4. Ciclo de vida
5. Herramientas
6. Caso práctico
7. Conclusiones

Dimensiones de calidad

¿Qué es una dimensión de calidad?

Una dimensión de calidad de datos es un término reconocido utilizado por los profesionales de gestión de datos para **describir una característica de los datos** que se puede **medir o evaluar** según estándares definidos para determinar la calidad de los datos.

Por ejemplo:

- Un set de datos tiene un 93% de cumplimiento de calidad
- El resultado de una evaluación de precisión para un set de datos fue 84%

Dimensiones de calidad

Cómo usar las dimensiones

Las organizaciones seleccionan las **dimensiones** de calidad de los datos y los **umbrales** de las dimensiones asociadas en **función de su contexto** de negocio, requisitos, niveles de riesgo, etc. Hay que tener en cuenta que es probable que cada dimensión tenga una **ponderación** diferente y, para obtener una medida precisa de la calidad de los datos, la organización deberá determinar cuánto contribuye cada dimensión a la calidad de los datos en su conjunto.

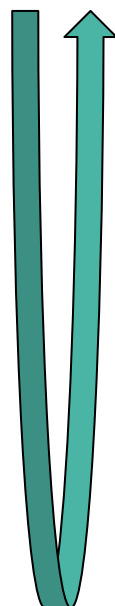
Por eso es necesario conocer el impacto de tener problemas en la calidad de datos. Algunos ejemplos:

- direcciones de correo electrónico incorrectas inexistentes tendrían un impacto significativo en cualquier campaña de marketing
- datos personales inexactos pueden conducir a oportunidades de ventas perdidas o un aumento en las quejas de los clientes
- los productos pueden enviarse a ubicaciones incorrectas
- las medidas incorrectas del producto pueden ocasionar problemas de transporte importantes, es decir, el producto no cabe en un camión, o por el contrario, se puede haber pedido demasiados camiones para el tamaño de la carga real (ineficiencia)

Dimensiones de calidad

Cómo usar las dimensiones

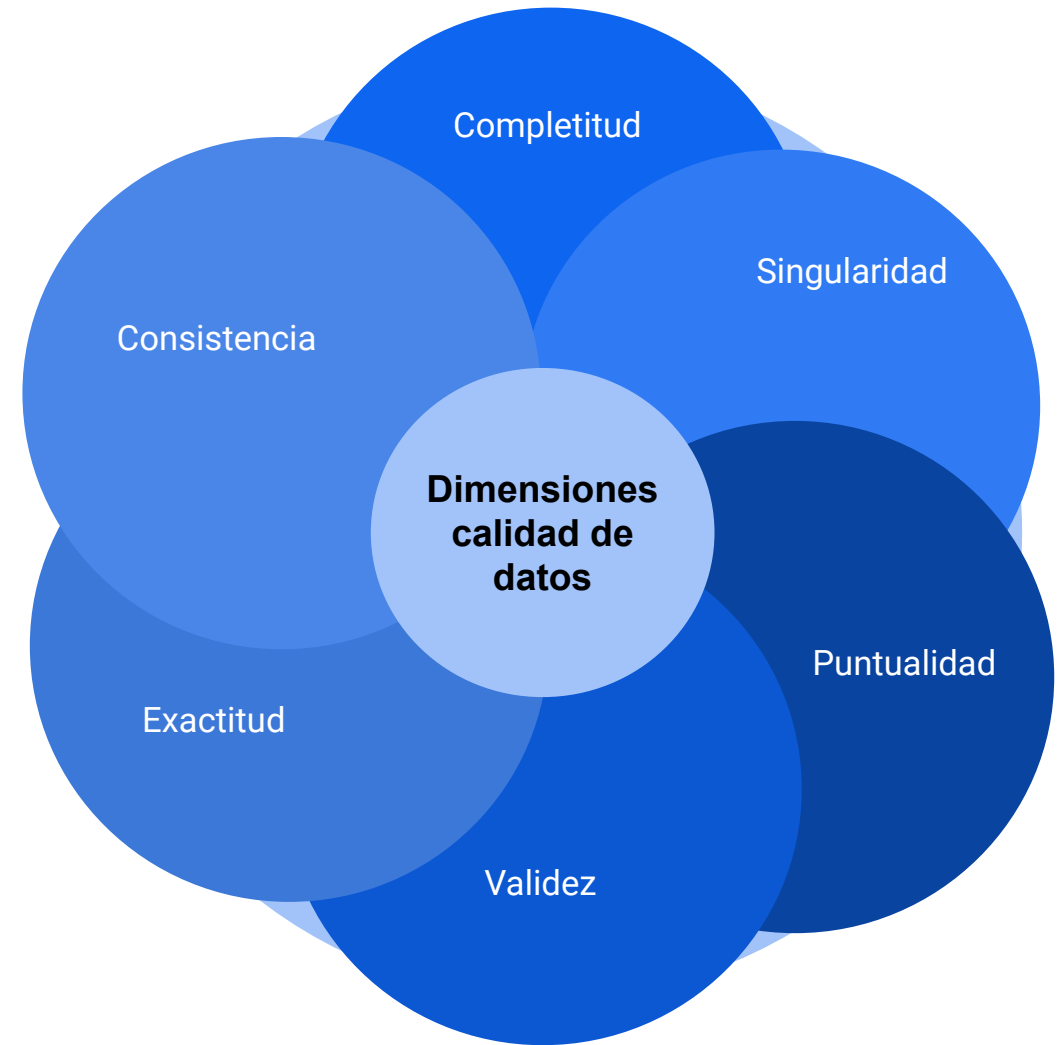
Un enfoque típico de evaluación de la calidad de los datos podría ser:

- 
1. **Identificar** qué **elementos** de datos deben evaluarse para determinar la calidad de los datos, normalmente serán elementos de datos considerados críticos para las operaciones comerciales y los informes de gestión asociados
 2. **Evaluar** qué **dimensiones** de calidad de datos usar y su ponderación asociada
 3. Para **cada dimensión** de calidad de datos, definir **valores o rangos** que representen datos de buena y mala calidad. Tener en cuenta que, como un conjunto de datos puede admitir múltiples requisitos, una cantidad de datos diferentes es posible que deba realizarse una serie de evaluaciones de calidad de datos diferentes
 4. Aplicar los **criterios de evaluación** a los elementos de datos.
 5. Revisar los resultados y determinar si la calidad de los datos es aceptable o no
 6. En su caso, tomar medidas correctivas, p.e. limpiar los datos y mejorar los procesos de manejo de datos para evitar futuras recurrencias (**plan de remediación**)
 7. Repetir lo anterior periódicamente para monitorear las tendencias en la calidad de los datos. (Evitar degradaciones)

Dimensiones de calidad

Hay varias dimensiones de calidad de datos. Depende del framework, se definen más o menos. En el caso de DAMA UK, definen 6 dimensiones principales:

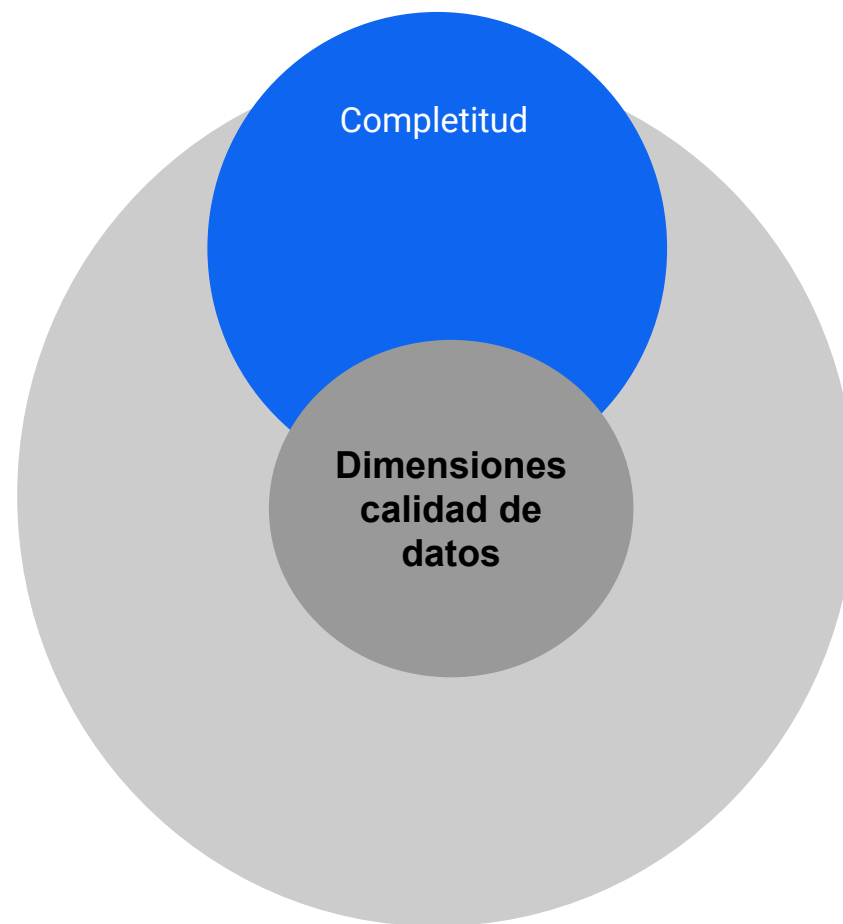
1. **Compleitud:** ¿Se registran para todos los sets de datos, cada uno de sus campos?
2. **Singularidad:** ¿Existe una vista única del conjunto de datos?
3. **Puntualidad:** ¿Los datos se guardan en tiempo aceptable?
4. **Validez:** ¿Los datos cumplen las reglas?
5. **Exactitud:** ¿Los datos reflejan la realidad?
6. **Consistencia:** ¿Los datos coinciden en todos los datastores?



Dimensiones de calidad

Completitud

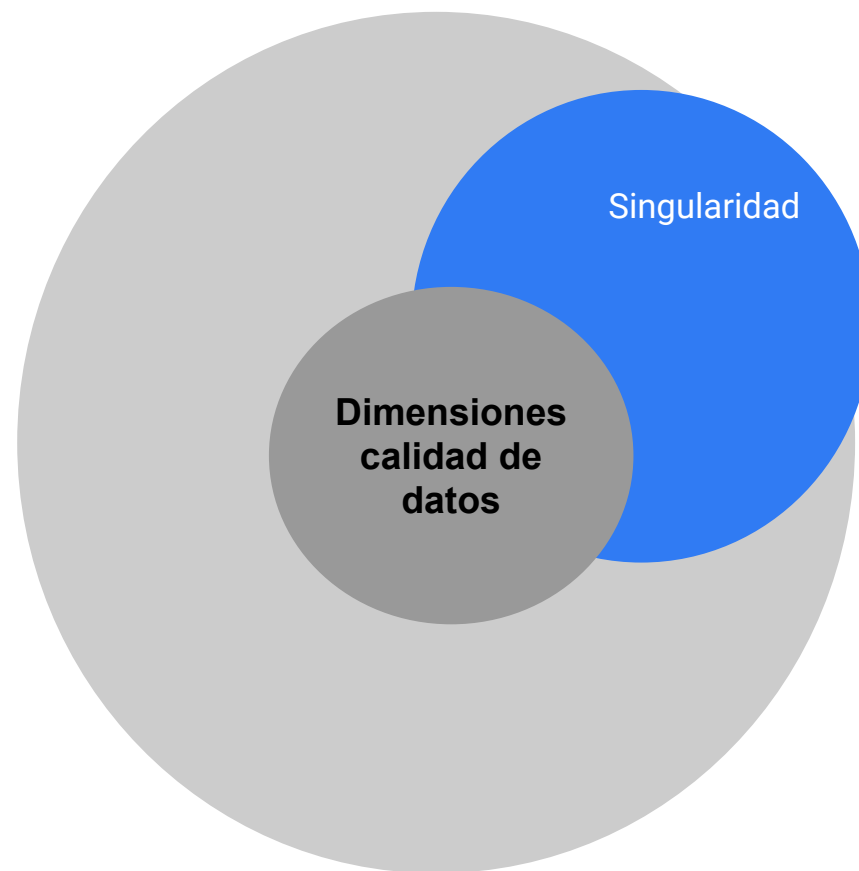
Definición	La proporción de datos almacenados frente al potencial de "100% completado"
Medida	Medir la ausencia de valores en blancos (cadena nula o vacía) o la presencia de valores no-nulos. Unidad de medida: %
Ejemplo	<p>Los padres de los nuevos estudiantes en el colegio deben completar una hoja de recopilación de datos que incluye condiciones médicas y detalles de contacto de emergencia, así como confirmar el nombre, la dirección y la fecha de nacimiento del estudiante.</p> <p><i>Situación:</i></p> <p>Al final de la primera semana del trimestre de otoño, se realizó un análisis de datos en el elemento de datos "Primer número de teléfono de contacto de emergencia" en la tabla de contactos. Hay 300 estudiantes en la escuela y 294 de un total de 300 registros se completaron, por lo tanto, $294/300 \times 100 = 98\%$ de completitud para este elemento de datos en la tabla de contacto</p>
Pseudo código	(Count 'Primer número de teléfono de emergencia' where not blank in the tabla Contacto) / (Count all estudiantes in the Contact table.)



Dimensiones de calidad

Singularidad (o unicidad)

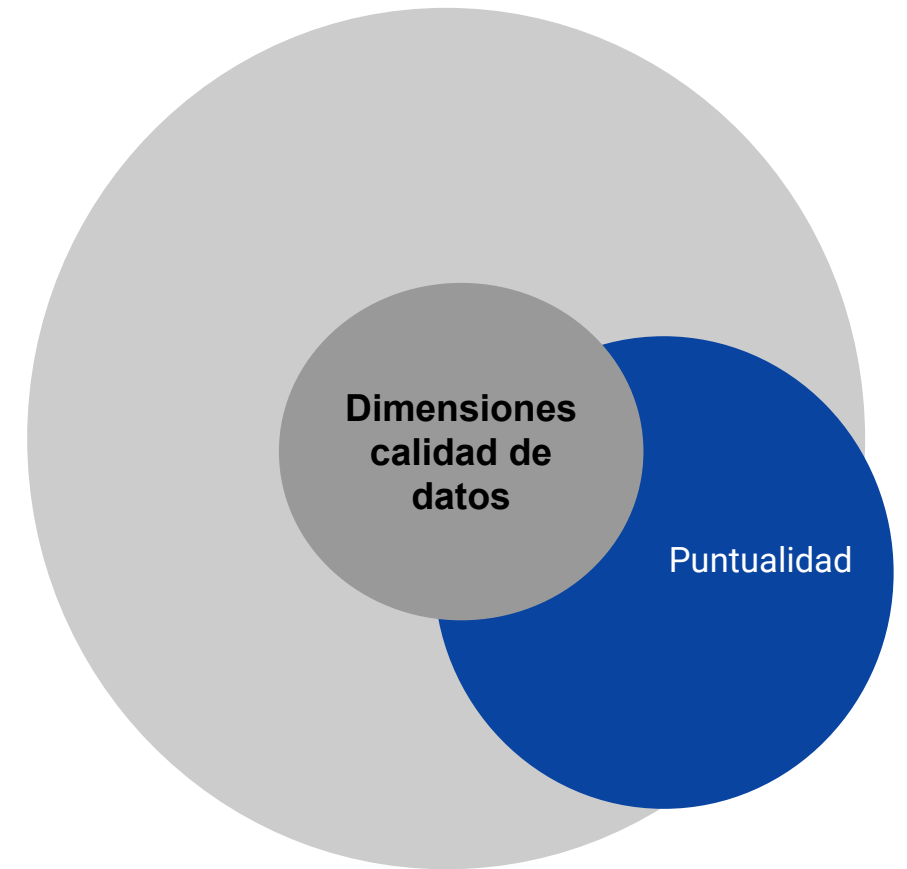
Definición	Nada se almacena más de una vez en función de cómo se identifica ese elemento.
Medida	Análisis del número de cosas evaluadas en el "mundo real" en comparación con el número de registros de cosas en el set de datos. El número real de cosas podría determinarse a partir de un conjunto de datos diferente y más confiable (maestro) o de un comparador externo relevante. Unidad de medida: %
Ejemplo	El colegio tiene 120 estudiantes actuales y 380 ex alumnos (es decir, 500 en total); la base de datos de estudiantes muestra 520 registros de estudiantes diferentes. Esto podría incluir a Fred Smith y Freddy Smith como registros separados, a pesar de que sólo hay un estudiante en la escuela llamado Fred Smith. Esto indica una singularidad de $500/520 \times 100 = 96.2\%$
Pseudo código	$(\text{Numero de elementos reales}) / (\text{Numero de elementos almacenados})$



Dimensiones de calidad

Puntualidad

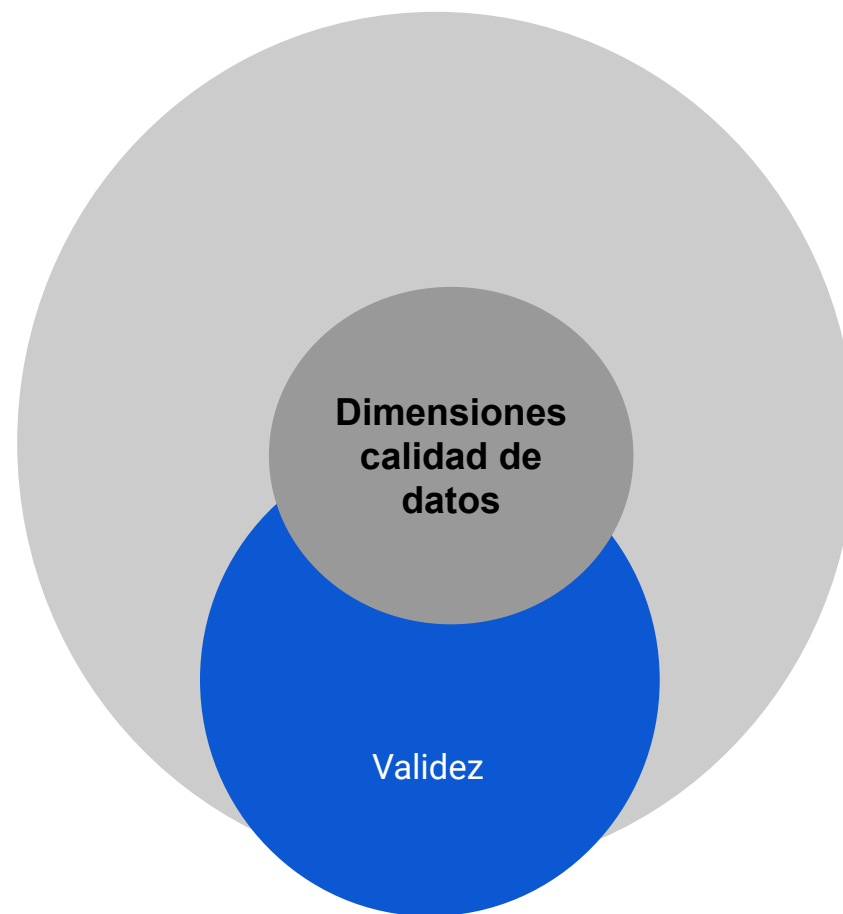
Definición	El grado en que los datos representan la realidad desde el momento requerido.
Medida	Diferencia de tiempo. Unidad de medida: tiempo
Ejemplo	Tina Jones proporciona detalles de un número de contacto de emergencia actualizado el 1 de junio de 2013, que luego se guarda en la base de datos del estudiante el 4 de junio de 2013. Esto indica un retraso de 3 días. Este retraso infringe la restricción de puntualidad ya que el acuerdo de nivel de servicio para los cambios es de 2 días.
Pseudo código	Fecha modificación del registro (4th June 2013) menos fecha de información del dato (1st June 2013) = 3 días



Dimensiones de calidad

Validez

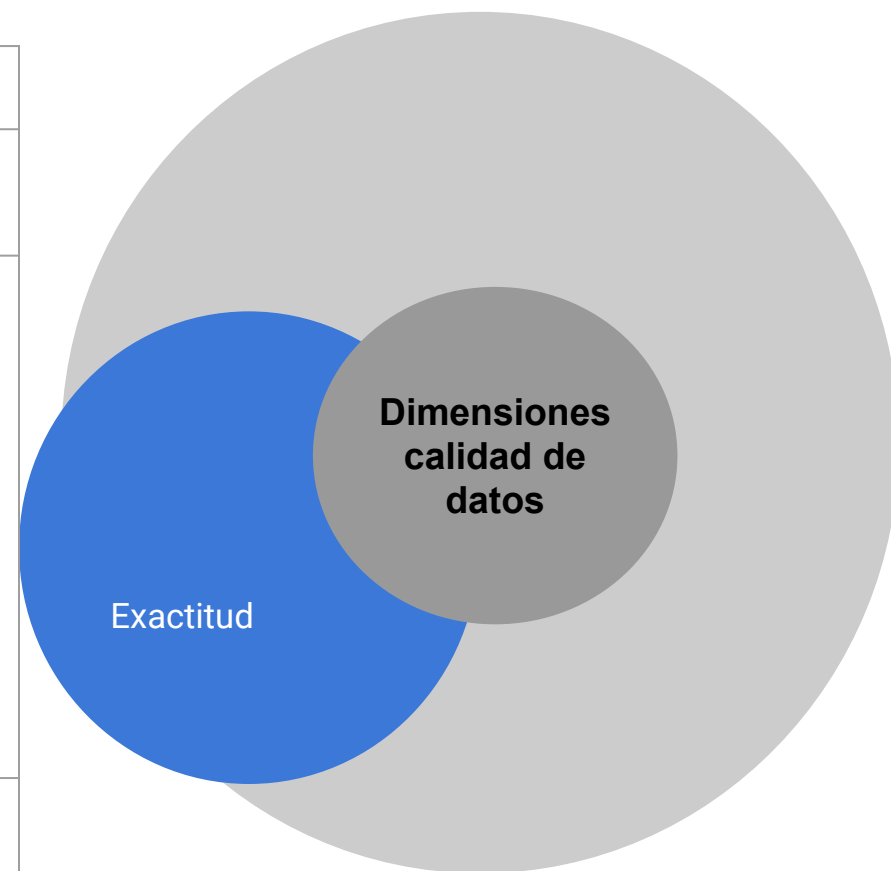
Definición	Los datos de definición son válidos si se ajustan a la sintaxis (formato, tipo, rango) de su definición.
Medida	Comparación entre los datos y los metadatos (definición). Unidad: %
Ejemplo	<p>A cada clase en el colegio se le asigna un identificador de clase; Consiste en las 3 iniciales del maestro más un número de grupo de año de dos dígitos de la clase. Se declara como AAA99 (3 caracteres alfabéticos y dos caracteres numéricos).</p> <p>Situación:</p> <p>Se nombra a una maestra de noveno año, Sally Hearn (sin segundo nombre), por tanto, solo hay dos iniciales. Se debe decidir sobre cómo representar dos iniciales o la regla fallará y la base de datos rechazará el identificador de clase de "SH09". Se decide que se agregará un carácter adicional "Z" para rellenar las letras a 3: "SZH09", sin embargo, esto podría romper la regla de exactitud. Una mejor solución sería modificar la base de datos para aceptar 2 o 3 iniciales y 1 o 2 números.</p>
Pseudo código	Evaluar que el identificador de clase es de 2 o 3 letras a-z seguido de 1 o 2 números



Dimensiones de calidad

Exactitud

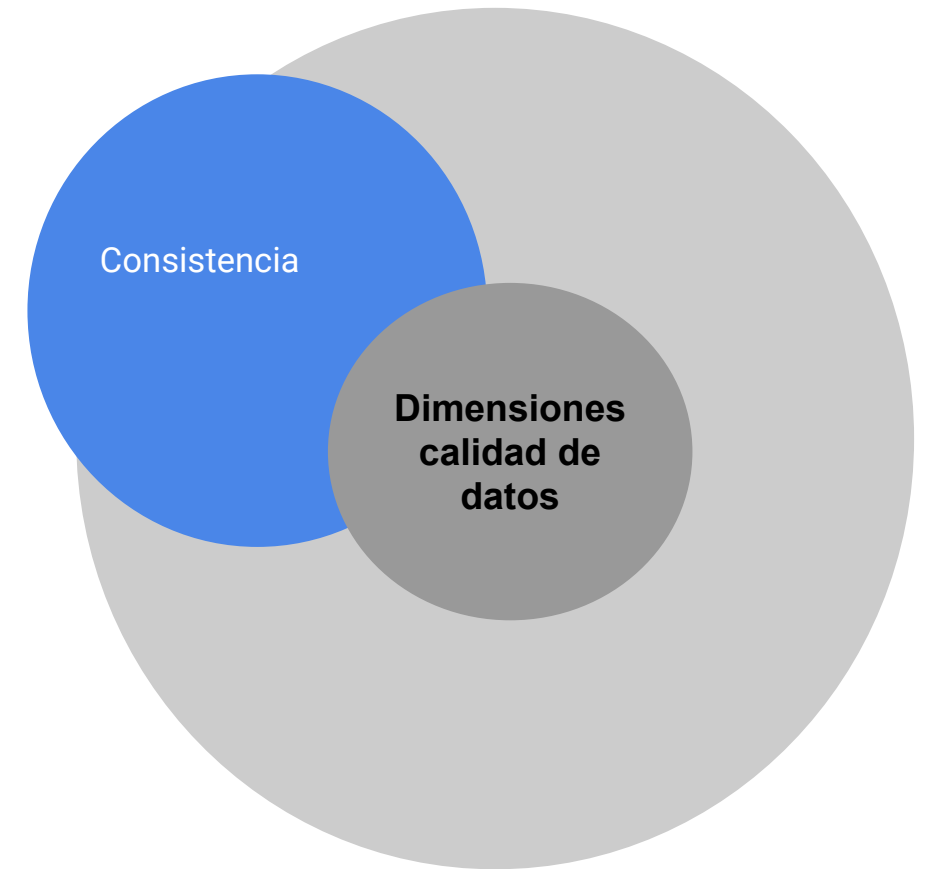
Definición	El grado en que los datos describen correctamente el objeto del "mundo real"
Medida	El grado en que los datos reflejan las características del objeto u objetos del mundo real que representa. Unidad: %
Ejemplo	Un colegio europeo está recibiendo solicitudes para su admisión anual de septiembre y requiere que los estudiantes tengan 5 años antes del 31 de agosto del año de admisión. En este escenario, el padre, un ciudadano americano, completa la fecha de nacimiento (DOB) en el formulario de solicitud en el formato de fecha de EEUU, MM/DD/AAAA en lugar del formato europeo DD / MM / AAAA, haciendo que se revierta la representación de días y meses. Como resultado, 09/08/2014 realmente significaba 08/09/2014, haciendo que el estudiante sea aceptado como la edad de 5 años el 31 de agosto en 2019. La representación de la DOB. del alumno, aunque válida en su contexto estadounidense, significa que en Europa la edad no se obtuvo correctamente y, por lo tanto, el valor registrado no fue exacto.
Pseudo código	$\frac{(\text{Número de niños que presentaron solicitud de 5 años antes de agosto})}{(\text{Número de niños que presentaron una solicitud de 5 años antes del 31 de agosto} + \text{Número de niños que presentaron una solicitud de 5 años después de agosto y antes del 31 de diciembre})} \times 100$



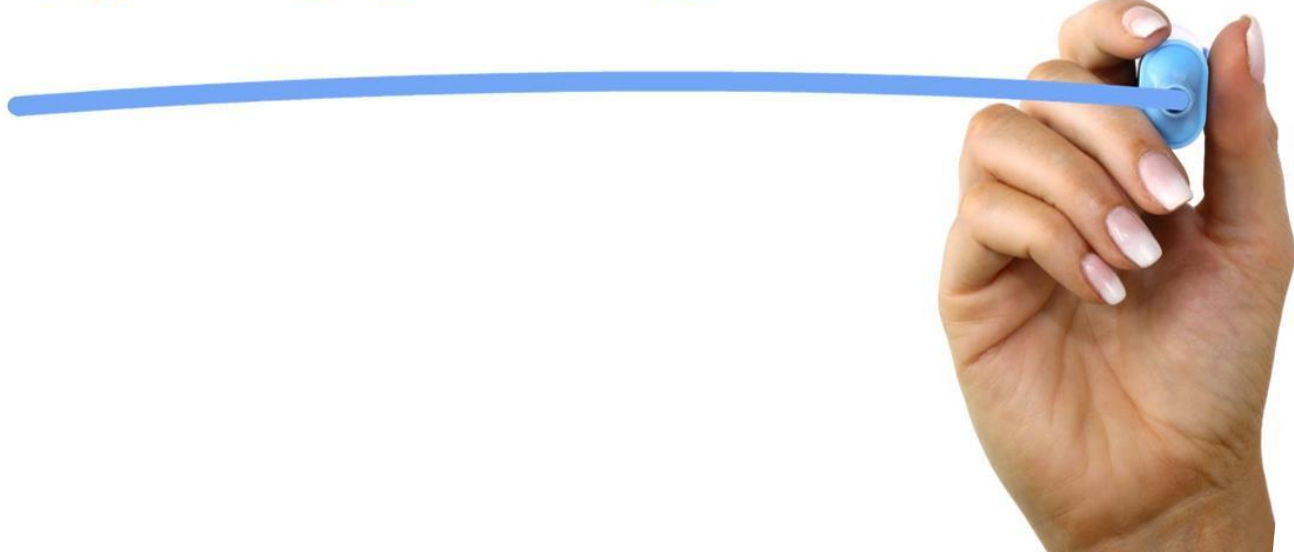
Dimensiones de calidad

Consistencia

Definición	La ausencia de diferencia, al comparar dos o más representaciones de una cosa con una definición
Medida	Análisis de patrón y / o frecuencia de valor. Unidad: %
Ejemplo	Administrador del colegio: la fecha de nacimiento de un estudiante tiene el mismo valor y formato en el registro escolar que el almacenado en la base de datos de estudiantes.
Pseudo código	Select count distinct on 'fecha nacimiento'



INDEX



Índice

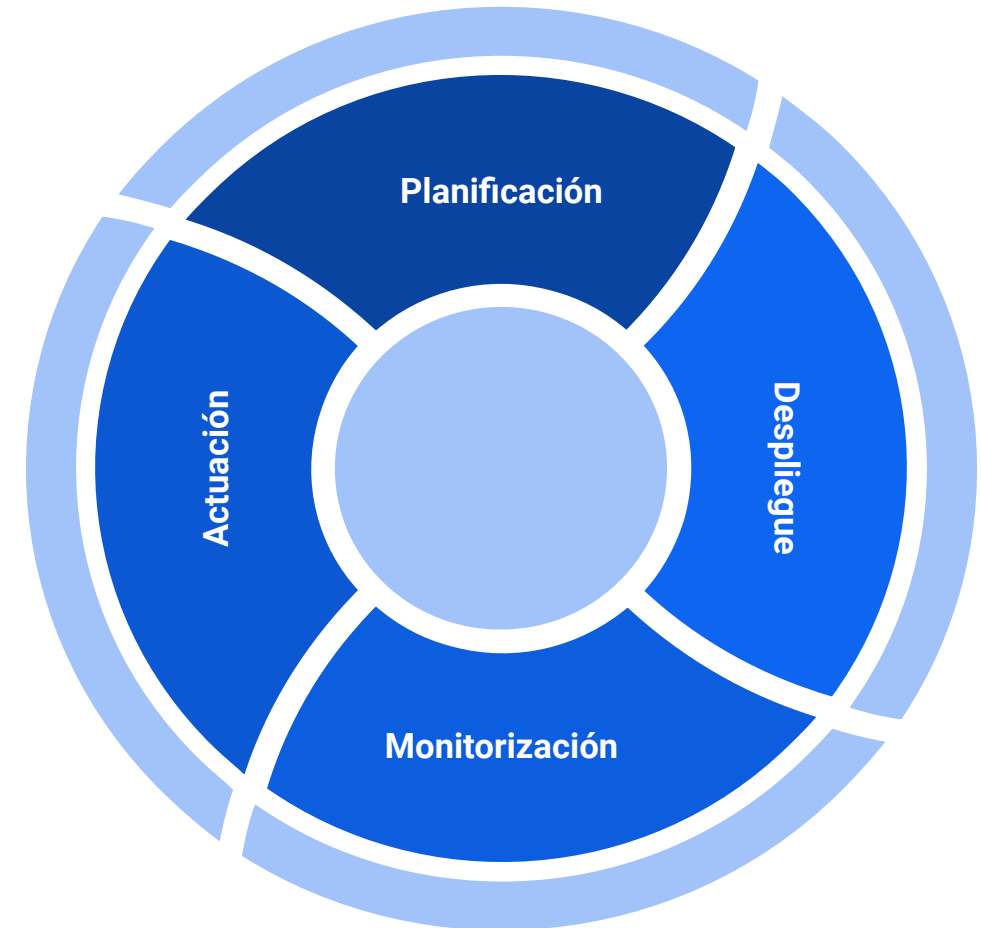
1. Necesidades de calidad de datos
2. ¿Qué es calidad de datos?
3. Dimensiones de calidad
- 4. Ciclo de vida**
5. Herramientas
6. Caso práctico
7. Conclusiones

Ciclo de vida

DAMA

DAMA define un ciclo basado en 4 etapas:

1. Planificación
2. Despliegue
3. Monitorización
4. Actuación



Ciclo de vida

Planificación

Plan para la evaluación del **estado actual** y la identificación de **métricas clave** para medir la calidad.

Se debe establecer el coste e **impacto** de los problemas de calidad conocidos y evaluar **soluciones** y alternativas para abordar o minimizar el impacto.



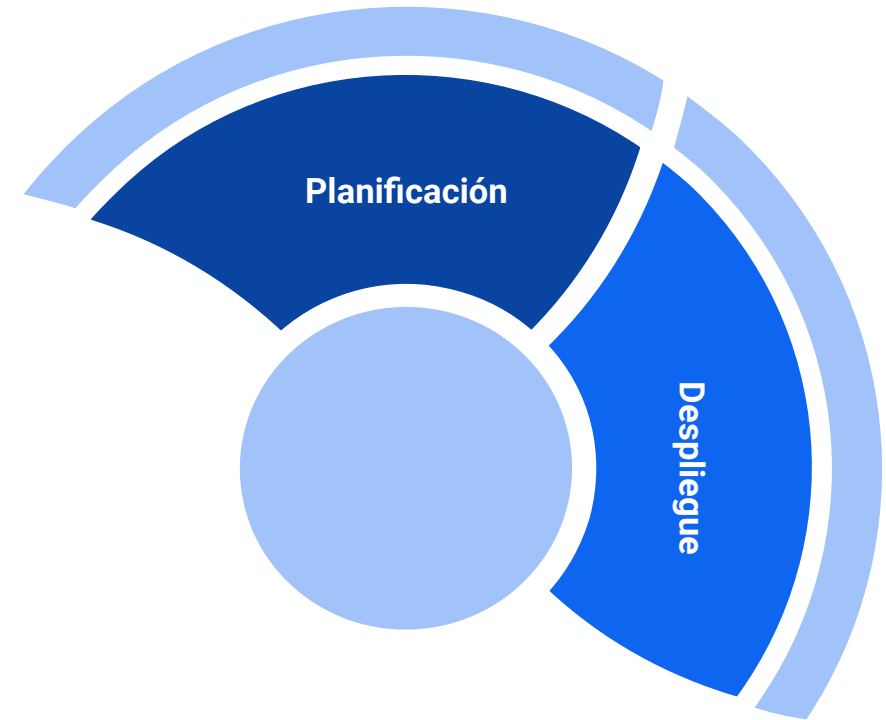
Ciclo de vida

Despliegue

Implementar procesos para **medir y mejorar** la calidad de los datos:

Perfilado de datos (**Data profiling**):

- Implantar **inspecciones** y monitoreo para **identificar** problemas de datos cuando ocurran
- **Corregir** los procesos defectuosos que son la causa raíz de los errores de datos o corregir los errores posteriores (remediación de datos)
- Cuando no es posible corregir errores en su origen, corregirlos en su punto **más temprano** en el flujo de datos

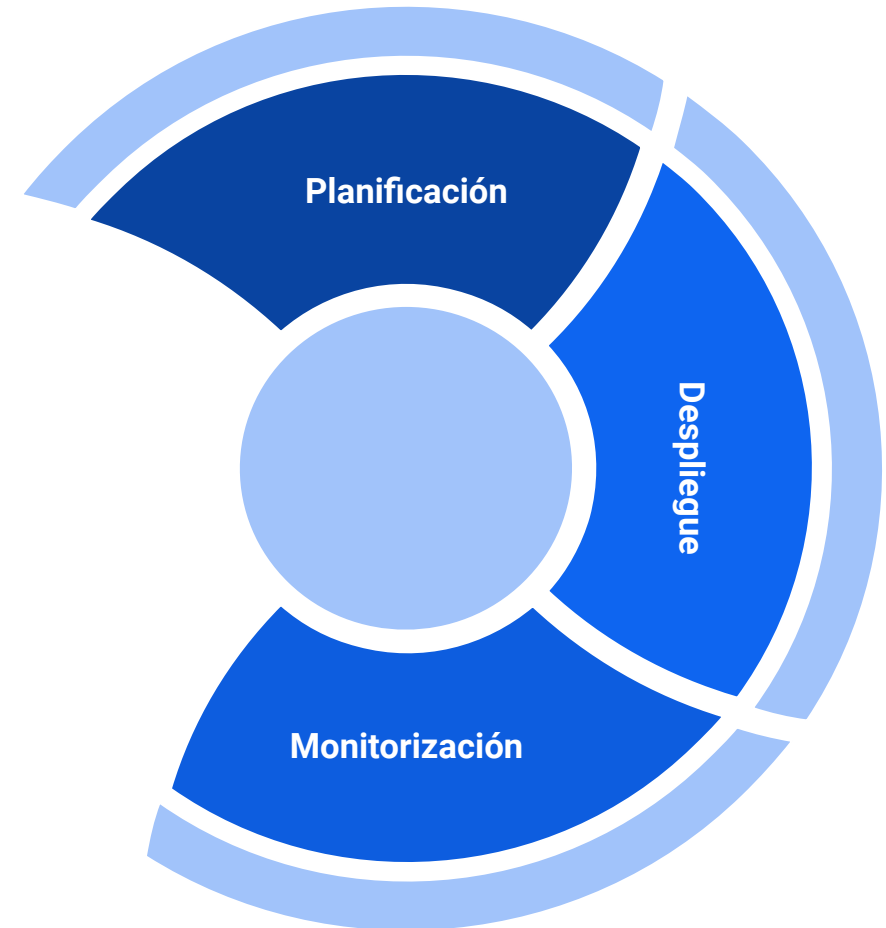


Ciclo de vida

Monitorización

Monitorizar la calidad de los datos medidos contra las reglas de negocio (business rules) definidas

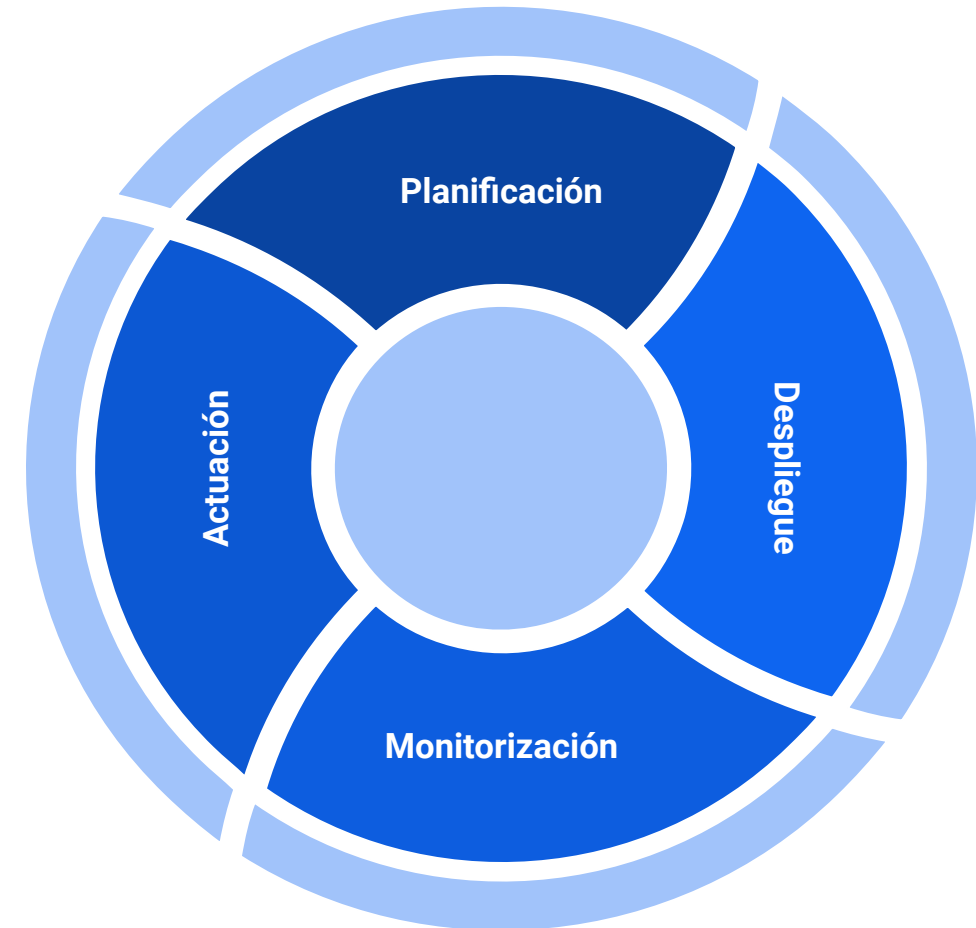
- Si la calidad de los datos **cumple** con los umbrales definidos para la aceptabilidad, los procesos están bajo control y el nivel de calidad de los datos cumple con los requisitos del negocio.
- Si la calidad de los datos **cae por debajo** de los umbrales de aceptabilidad, **notifique** a los administradores de datos para que puedan tomar medidas durante la siguiente etapa



Ciclo de vida

Actuación

- Actuar para **resolver** cualquier problema identificado para mejorar la calidad de los datos y cumplir mejor con las expectativas del negocio.
- Los nuevos ciclos comienzan cuando se investigan **nuevos conjuntos** de datos o cuando se identifican **nuevos requisitos** de calidad de datos para los conjuntos de datos existentes



Data profiling - Ejemplos típicos

Profiling de columna



- Count, unique count, null count, blank count, pattern count
- Minimum, maximum, mean, mode, median, standard deviation, standard error
- Completeness (% of non-null records)
- Data type (defined v actual)
- Primary key candidates

Análisis de frecuencia



- Count/percentage each distinct value
- Count/percentage each distinct character pattern

Análisis de PK y FK



- Candidate primary/foreign key relationships
- Referential integrity checks between tables

Análisis duplicados



- Identification of potential duplicate records (with variable sensitivity)

Conformidad de reglas de negocio



- Using a preliminary set of business rules

Análisis de outliers (valores atípicos)



- Identification of possible out of range values or anomalous records

Ciclo de vida

Data cleansing (remediación)

El **análisis detallado** de cada problema de calidad de datos es vital para garantizar que se identifique la causa raíz y, siempre que sea posible, se elimine para que no se repitan los incidentes. Además de esto, los problemas de calidad existentes deben **resolverse** mediante uno de los siguientes mecanismos:

- Corrección **automatizada**: los defectos obvios a menudo pueden identificarse y repararse activando una rutina de limpieza de datos automatizada, sin intervención manual (por ejemplo, estandarización de direcciones o sustitución de campos)
- Corrección **dirigida**: los defectos menos obvios a menudo se pueden identificar automáticamente, pero pueden requerir intervención manual para determinar si la solución sugerida es adecuada (por ejemplo, resolución de identidad y deduplicación)
- Corrección **manual**: en algunos casos, aunque un defecto puede identificarse automáticamente, la única forma de resolverlo es mediante la inspección y corrección manual (por ejemplo, una combinación no válida de campos en los que no está claro qué campo tiene la culpa)

Las herramientas de calidad de datos utilizan un sistema de puntuación para reflejar el nivel de confianza en la aplicación de una corrección; esto se puede usar para decidir qué defectos deben corregirse automáticamente (la opción más barata y, a menudo, la preferida) y cuáles deben marcarse como dirigidos o manuales corrección

Ciclo de vida

Data cleansing (remediación)

The *quïck* fox jump's over the the lazy dog

Ciclo de vida

Data cleansing (remediación)

STANDARDISATION

The quiick fox jump's over the the lazy dog

Ciclo de vida

Data cleansing (remediación)

STANDARDISATION

SUBSTITUTION

The quick fox jumps over the the lazy dog

Ciclo de vida

Data cleansing (remediación)

STANDARDISATION

SUBSTITUTION

DE-DUPLICATION

The quick fox jumps over the the lazy dog

Ciclo de vida

Data cleansing (remediación)

STANDARDISATION

SUBSTITUTION

DE-DUPLICATION

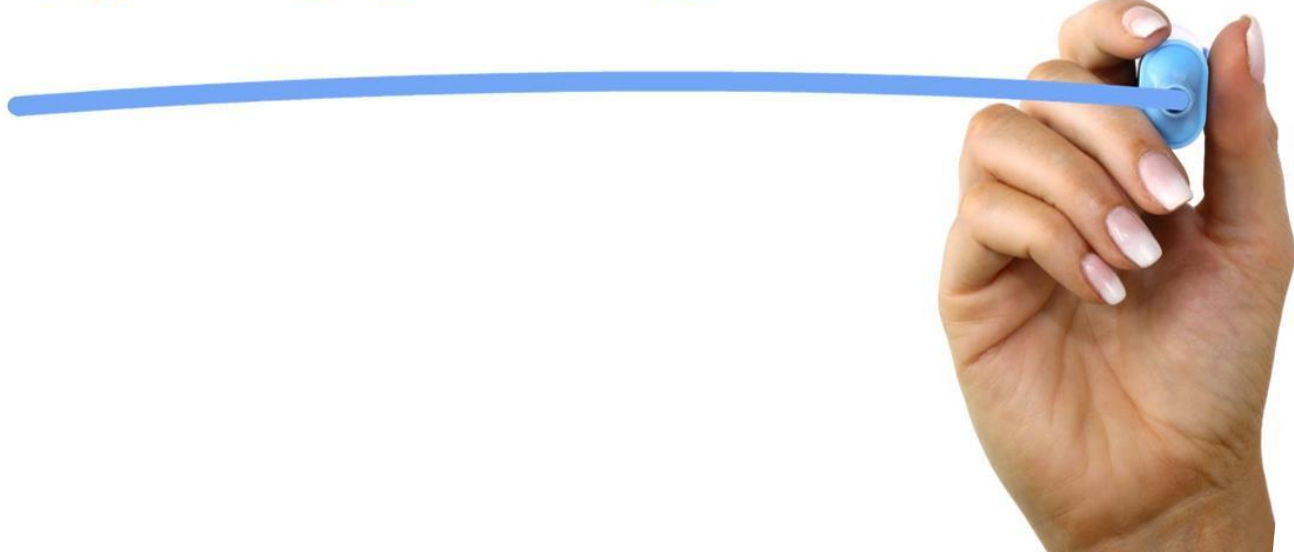
ENRICHMENT

brown

The quick fox jumps over the lazy dog



INDEX



Índice

1. Necesidades de calidad de datos
2. ¿Qué es calidad de datos?
3. Dimensiones de calidad
4. Ciclo de vida
- 5. Herramientas**
6. Caso práctico
7. Conclusiones

Herramientas

Informatica



Informatica es reconocida como un proveedor destacado en el mercado de software de gestión de datos. La plataforma permite a las organizaciones acceder, integrar, limpiar, masterizar, gobernar y asegurar big data.

La herramienta tiene conectores especialmente diseñados para cientos de fuentes de datos, tratamiento en tiempo real e ingestión masiva. La interfaz visual de desarrollador también garantiza que se puedan adaptar a las mejores plataformas open source sin sacrificar la usabilidad.

El soporte de la nube pública para Big Data Management está disponible en AWS y Microsoft Azure.



Herramientas

SAP



SAP ofrece sus capacidades de gestión de datos en una sola plataforma. SAP HANA permite a los usuarios recopilar y combinar todo tipo de datos en tiempo real, así como mejorar el gobierno, la supervisión y la orquestación de datos.

Los usuarios también pueden crear una vista unificada de datos con integración “smart data” que permite aplicaciones y gestión de datos avanzados. La plataforma es flexible y se puede implementar en las instalaciones, en la nube o mediante implementaciones híbridas. HANA es una herramienta en memoria con procesamiento de datos rápido y análisis avanzado con procesamiento OLAP y OLTP.



Herramientas

IBM



IBM Db2 Hybrid Data Management ofrece a las organizaciones la opción de seleccionar cualquier tipo de base de datos, almacén de datos o software open source.

La solución recopila, gestiona y proporciona información sobre datos en la nube local, privada y pública, e integrada con tipos de datos estructurados y no estructurados.

IBM proporciona aprendizaje automático integrado y ciencia de datos para que los usuarios puedan ejecutar análisis de datos en su entorno nativo, y el motor SQL con virtualización de datos incorporada, permite una gestión de datos escalable.



Herramientas

SAS



SAS Data Management es especialmente útil cuando se analizan grandes y complejos volúmenes de datos. Sin embargo, la solución tiene un poco de curva de aprendizaje y es mejor para los usuarios con experiencia en software y lenguaje SAS.

SAS está construido sobre un marco de calidad de datos, y el business glossary incorporado, así como las capacidades de visualización de linaje y gestión de metadatos de terceros y de terceros mantienen a todos los usuarios sincronizados.



Herramientas

Talend



Talend ha reforzado sus capacidades de gestión de datos durante el último año, y recientemente anunció el lanzamiento de una nueva solución de gestión de metadatos que proporciona a las organizaciones un acceso más fácil a los data lakes y otros proyectos de big data.

Talend Metadata Manager ofrece un marco de gobierno para crear, controlar, atribuir, definir y administrar datos empresariales para que los usuarios puedan extraer y propagar valor adicional. Contiene funcionalidades de preparación de datos y el soporte de la compañía para entornos multi-nube.



Herramientas

Oracle

ORACLE

El conjunto de capacidades de gestión de datos de Oracle permite a los usuarios administrar conjuntos de datos tanto tradicionales como nuevos en su plataforma en la nube.

La compañía también ofrece una nube de almacenamiento de datos autónomo con más de 2,000 aplicaciones SaaS.

La plataforma abarca la gama de funcionalidades de big data, con soporte para integración de datos y análisis también. Sus otras ofertas de gestión de datos incluyen Oracle Big Data Cloud, Oracle Big Data Cloud Service, Oracle Big Data SQL Cloud Service y Oracle NoSQL Database.



Herramientas

Precisely



Antiguamente conocido como Syncsort. Ofrece sus capacidades de integración de datos a través de tres plataformas distintas, Syncsort DMX, DMX-h e Ironstream.

La herramienta principal de la compañía es DMX, una solución que lleva todas las transformaciones de datos a un motor ETL de alto rendimiento. Syncsort DMX permite a los usuarios acelerar las consultas y aplicaciones de la base de datos al utilizar mejor las bases de datos relacionales. La función de ejecución inteligente selecciona dinámicamente los algoritmos más eficientes en función de las estructuras de datos y los atributos del sistema que encuentra en tiempo de ejecución.



Herramientas

Stratio

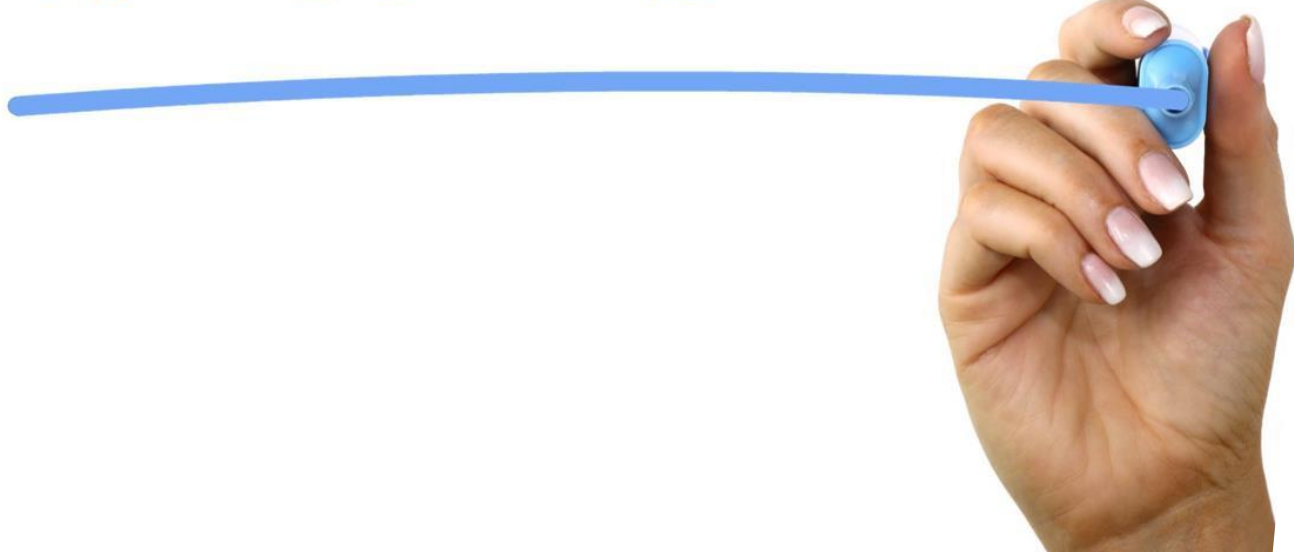


Stratio ofrece dentro de su plataforma Data Centric la solución para las empresas para abordar la transformación digital y desplegar cualquier caso de uso de negocio en una misma plataforma, de forma sencilla, para que los propios usuarios de negocio puedan ser autónomos en las soluciones a desplegar utilizando inteligencia artificial y tecnología distribuida. Dando capacidades de knowledge graph y virtualización de datos semánticos.

Su módulo de Gobierno (Stratio Governance) está orientado en facilitar la toma de decisiones para la gestión de datos. Unifica la gobernanza para garantizar la integridad de los datos, la coherencia y la validez de la información en todo el proceso de extremo a extremo, vocabulario común de términos de negocio, etc.



INDEX

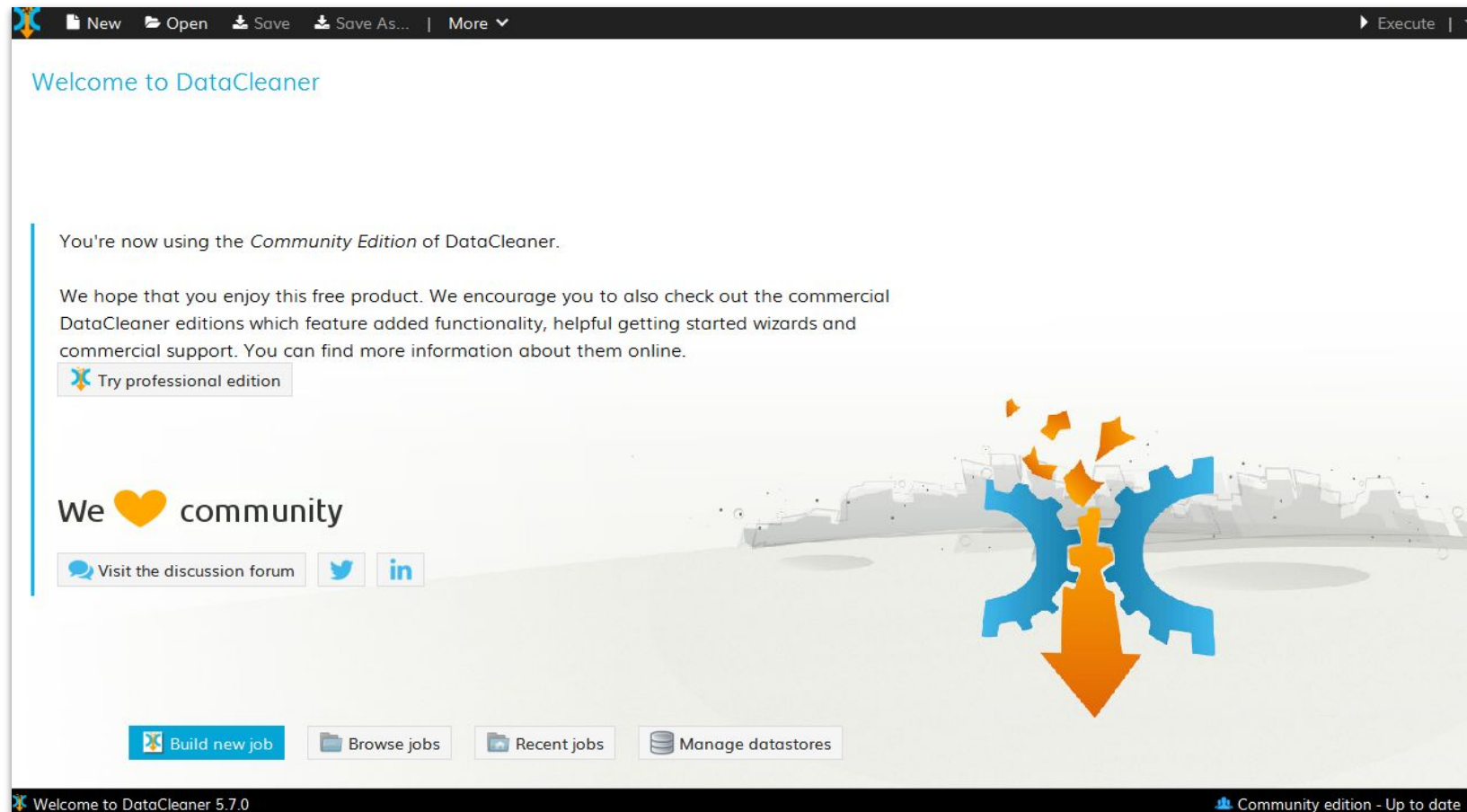


Índice

1. Necesidades de calidad de datos
2. ¿Qué es calidad de datos?
3. Dimensiones de calidad
4. Ciclo de vida
5. Herramientas
6. **Caso práctico**
7. Conclusiones

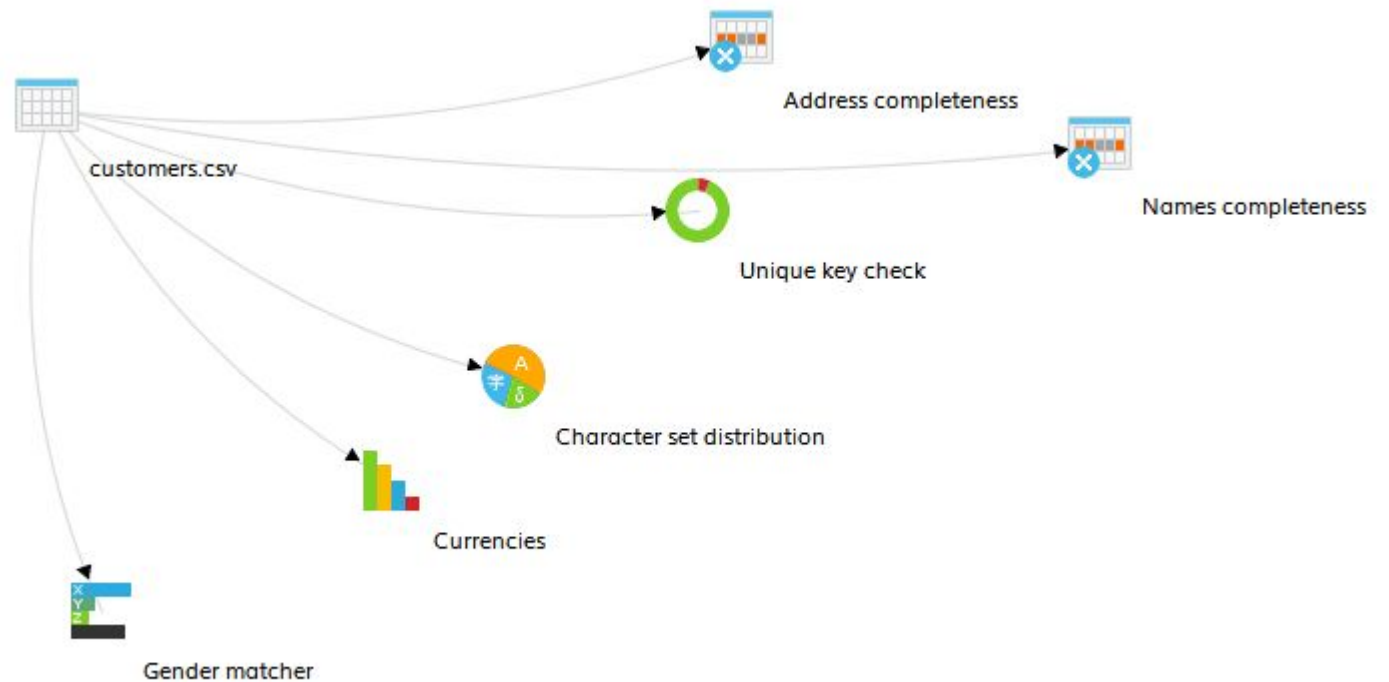
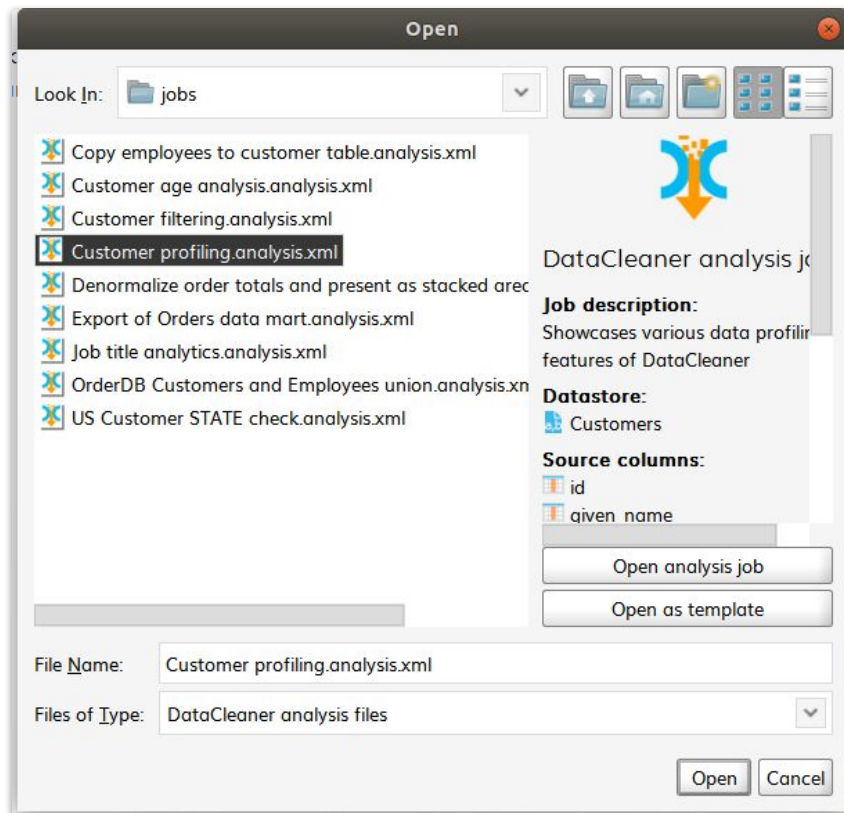
Caso práctico

Data Cleaner <https://datacleaner.github.io/downloads>



Caso práctico

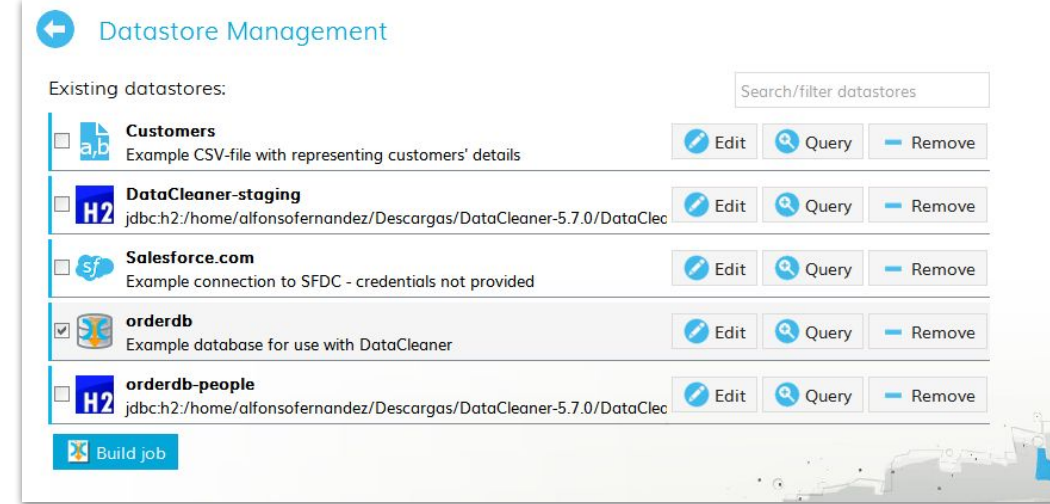
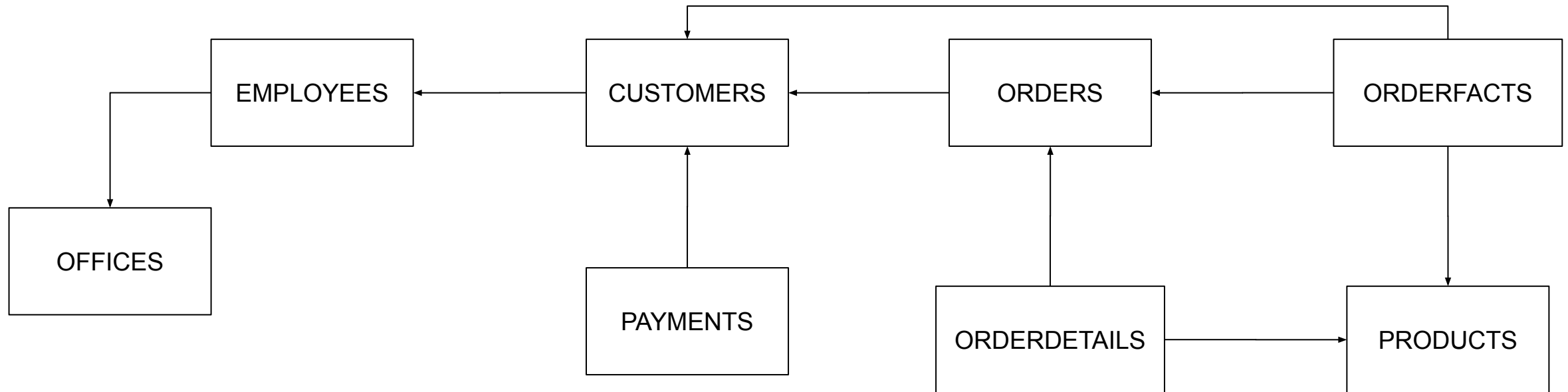
Profiling - Ejemplo



Caso práctico

OrderDB - Ciclo de vida de DQ

Somos el nuevo equipo del CDO de “Rayo McQueen motorland”. Tenemos una base de datos con información de oficinas, empleados, clientes, productos, pedidos y pagos. Este es el modelo relacional:



Caso práctico

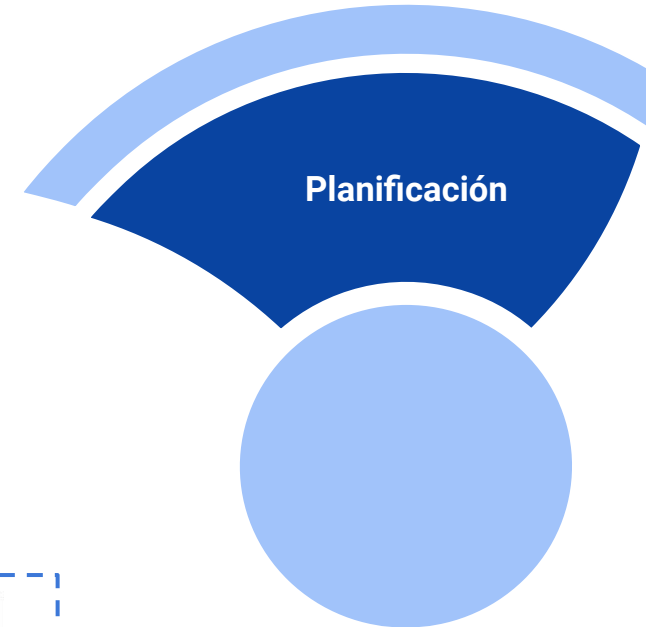
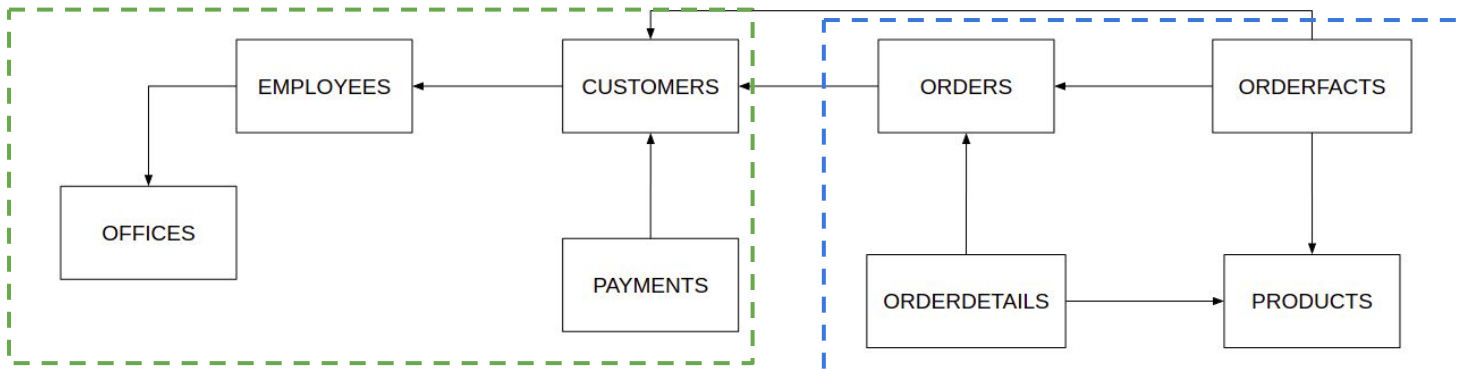
OrderDB - Ciclo de vida de DQ

Planificación:

Primero, entender los datos. 2 equipos:

Uno se encargará de las tablas Office, Employee, Customer y payments.

El otro del resto: Orders, Orderdetails, Orderfacts, y Products



Caso práctico

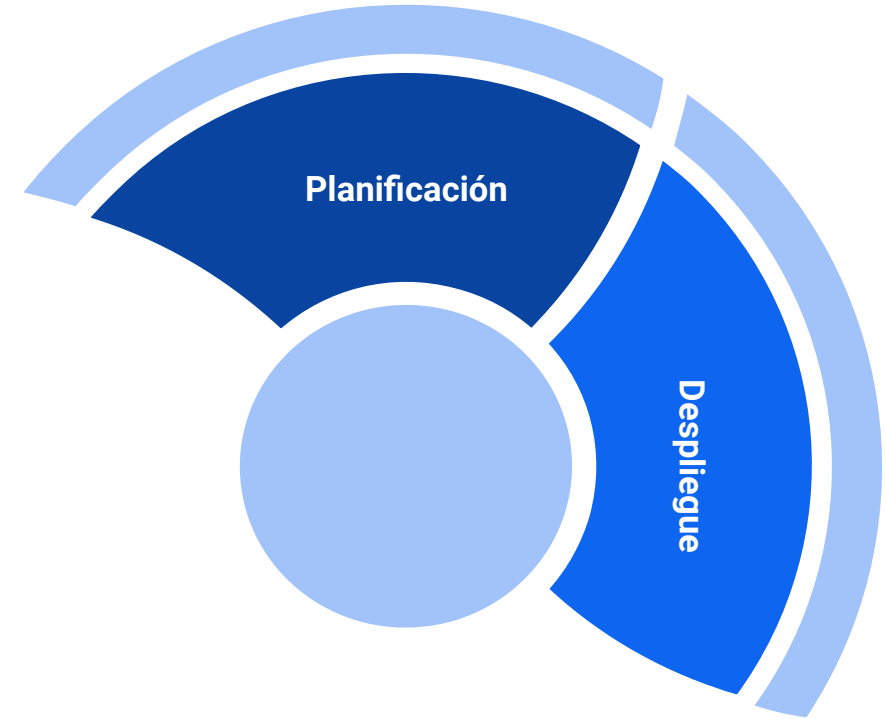
OrderDB - Ciclo de vida de DQ

Despliegue:

Sobre esas tablas, realizad un profiling para ver posibles problemas de datos.

...

¿Qué habéis visto?



Caso práctico

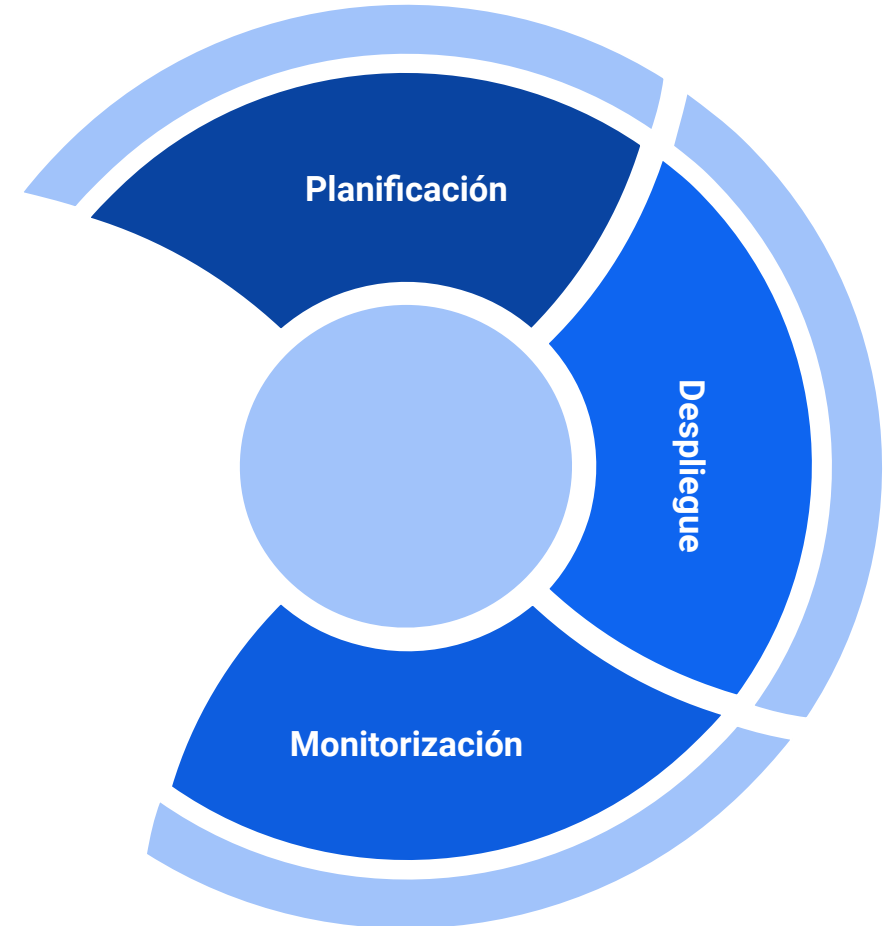
OrderDB - Ciclo de vida de DQ

Monitorización:

Seleccionad 2 o más reglas a implementar que sean de al menos 2 dimensiones distintas. Sacad el pseudocódigo

1. **Compleitud**
2. **Singularidad**
3. **Puntualidad**
4. **Validez**
5. **Exactitud**
6. **Consistencia**

(*) Hay alguna dimensión que no se podrá aplicar

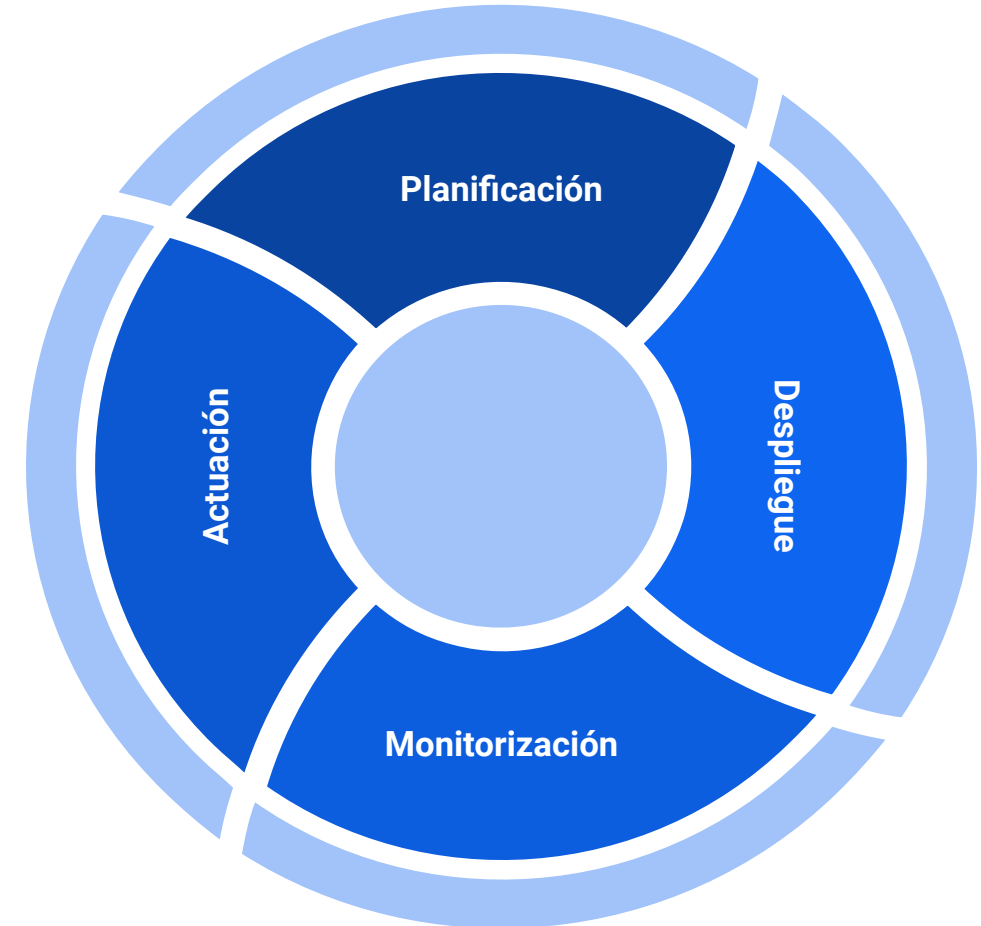


Caso práctico

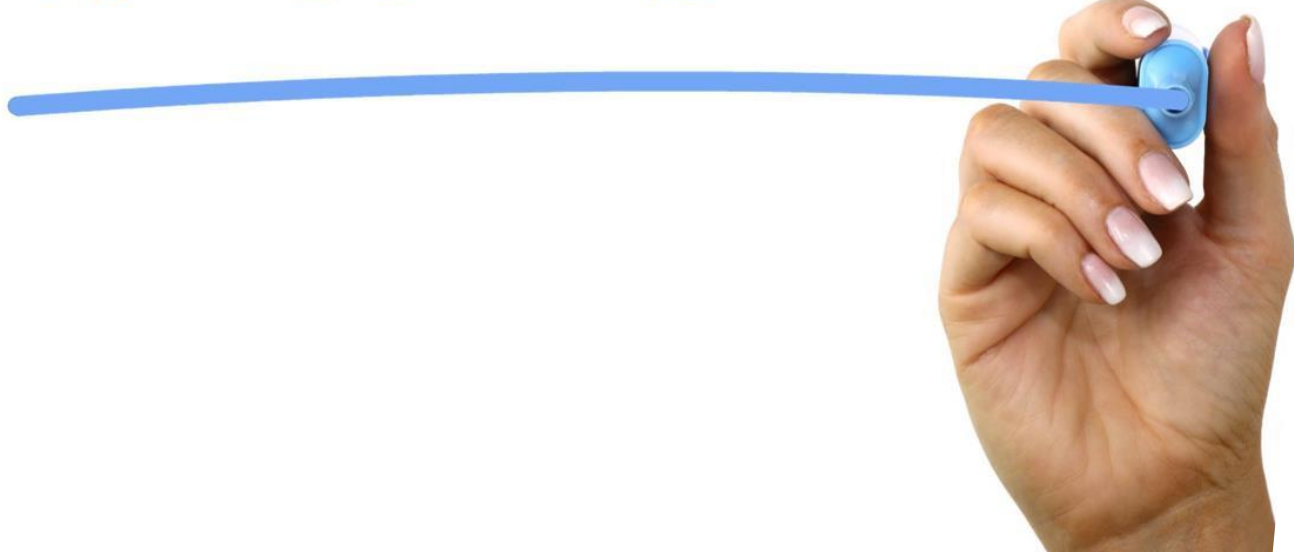
OrderDB - Ciclo de vida de DQ

Actuación:

Implementad vuestras reglas y sacad una versión de la tabla limpia.



INDEX



Índice

1. Necesidades de calidad de datos
2. ¿Qué es calidad de datos?
3. Dimensiones de calidad
4. Ciclo de vida
5. Herramientas
6. Caso práctico
7. **Conclusiones**

Conclusiones

Siete errores en la gestión de la calidad

No tener en cuenta el uso previsto de los datos.

Los datos deben ser adecuados para su propósito, ni más, ni menos.

1

Confundir Validez con precisión.

La validez es solo el primer paso hacia la precisión.

2

Tratar la gestión de calidad de datos como una actividad única.

Los datos de calidad sólo pueden garantizarse mediante un ciclo continuo

3

Arreglar los datos en los data warehouse en vez de en la fuente origen.

Los datos limpios para los informes no resuelven los problemas operativos de una calidad deficiente

4

Aplicar de principios de calidad de software a la calidad de los datos.

Los datos son infinitamente más volátiles que el SW y requieren un enfoque diferente

5

Culpar a los sistemas por los malos datos

Las personas y los procesos son los culpables de la mayoría de los problemas de DQ

6

Crear que el objetivo final es obtener datos de buena calidad

Obtener un beneficio a través de la explotación de la información es el objetivo final

7

Conclusiones

Lecciones aprendidas

- ✓ Administrar datos como un **activo** organizacional **central**.
- ✓ Identificar una **fuentes de verdad** para todos los elementos de datos.
- ✓ Todos los elementos de datos tendrán una **definición** de datos estandarizada, un **tipo** de datos y un **dominio**.
- ✓ Aprovechar el **gobierno de datos** para el control y el rendimiento de DQM
- ✓ Utilizar **estándares** de datos **internacionales** y de la industria siempre que sea posible.
- ✓ Los **consumidores** de datos especifican las **expectativas** de calidad de los datos.
- ✓ Definir **reglas de negocio** para afirmar **conformidad** con las **expectativas** de calidad de datos.
- ✓ Los **propietarios** de procesos de negocio **aceptarán y cumplirán los SLA** de calidad de datos
- ✓ Aplicar correcciones de datos en la **fuentes original**, si es posible.
- ✓ Si no es posible corregir los datos en la fuente, **enviar las correcciones** de datos al propietario de la fuente original.
- ✓ **Informar** de los niveles medidos de calidad de datos a los administradores de datos apropiados, propietarios de procesos de negocio y gestores de los SLAs de calidad.

¡Muchas gracias!