Calidad de datos en las organizaciones

Un método analítico para la evaluación de la calidad de las bases de datos

Jorge Villalobos Alvarado
Escuela Colombiana de Ingeniería
jorge.villalobos@escuelaing.edu.co

Contenido

- El caso de la calidad de datos
- Procesos que afectan la calidad de datos
- La definición de exactitud
- *Data Profiling* – evaluación de la calidad

El caso de la calidad de datos

I

La calidad de los datos en las organizaciones

- Los datos son activos corporativos o institucionales importantes pero es un hecho que en la mayoría de las organizaciones estos no se administran con el mismo rigor que otros activos.
- Lograr y mantener calidad en los datos requiere esfuerzo planeado, permanente y cuesta.
- Los datos, en la mayoría de las organizaciones, son deficientes en calidad.

Es un problema general …

- Los problemas de calidad de datos son universales – existen en todas las organizaciones.
- Por lo general la baja calidad obedece, no a una mala gestión en particular, sino a la ejecución normal de los procesos asociados con el manejo de información en la organización.

¿Qué dificulta controlar la calidad de los datos?

- Los cambios continuos y las rápidas implementaciones de sistemas.
- Los métodos, estándares, técnicas y herramientas para controlar la calidad no se han desarrollado al mismo ritmo que los de diseño, construcción e implementación de sistemas.
- Falta reconocimiento de la situación real.
- Falta conciencia sobre la importancia del tema.

Actividades desfavorables para la calidad de los datos …

- Tres principales:
  - Nuevos usos (o nuevas aplicaciones)
  - Replicación (o duplicación)
  - Integración

Cambios continuos …

- Los sistemas de información de una organización evolucionan y cambian continuamente.
- Los cambios los impulsan las necesidades del negocios – del mejor uso de la información
- De modo inexorable los cambios conducen a **nuevos usos** de los datos

Nuevos usos de los datos

- Los datos son de calidad si son adecuados para lo que se necesitan.
- La calidad depende tanto de los datos como del uso de los mismos.
- Con alta probabilidad, nuevos usos, o usos diferentes de los previstos en el diseño original, degradan la calidad de la base de datos.

Cambios de uso …

- Esto representa uno de los mayores problema de las bases de datos.
- Por muchas razones:
  - El diseño puede no incluir todos los campos necesarios.
  - Se acomodan los datos a un diseño inadecuado.
  - Las aplicaciones y los datos están fuertemente acoplados.
  - La metadata no refleja la realidad del contenido de la base de datos.
  - Con frecuencia hay replicación (duplicación) de datos.
  - …
- Es muy difícil anticipar los usos futuros de los datos al construir una base de datos (salvo que su contenido sea insignificante)

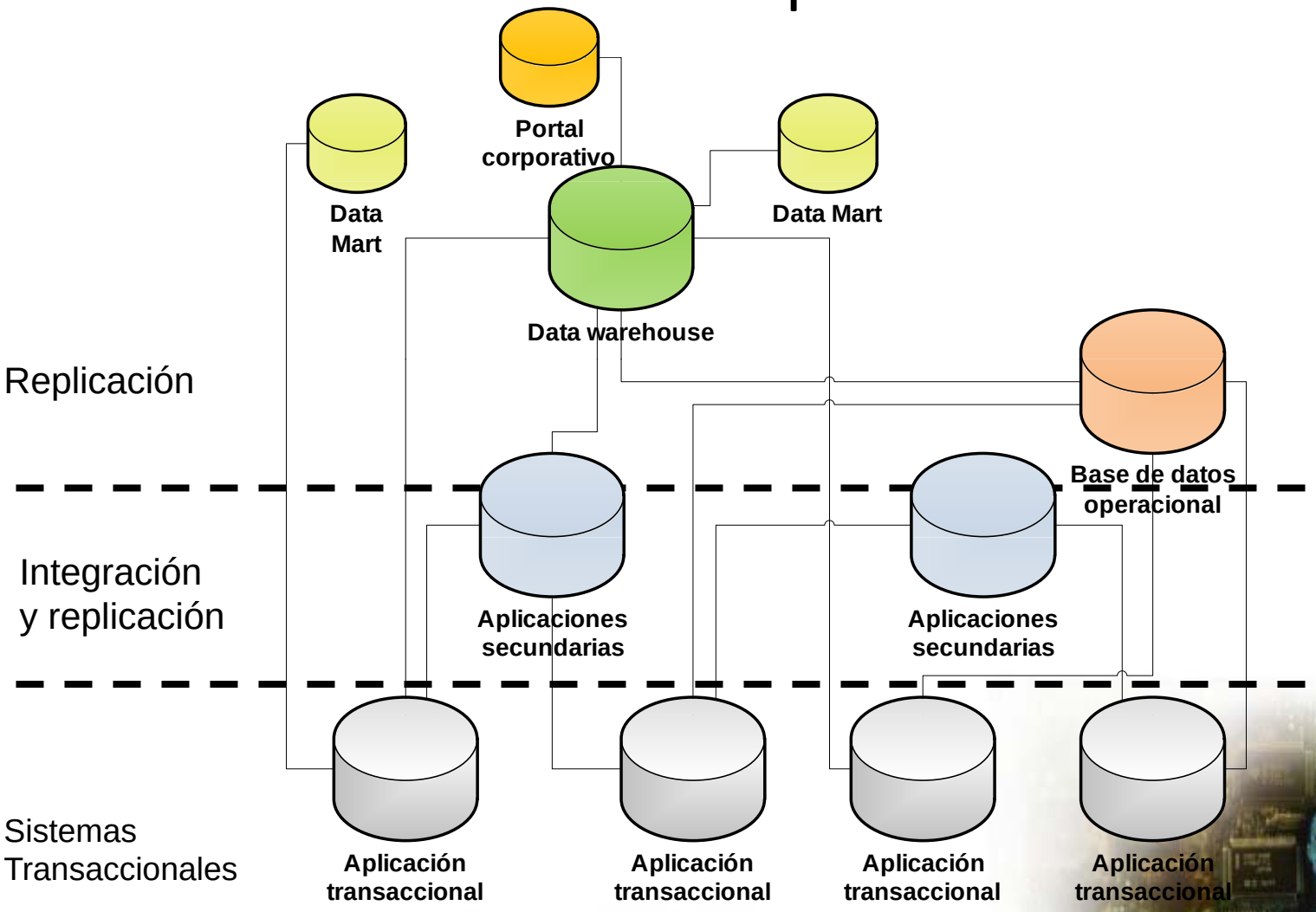
Replicación

- En las nuevas maneras de utilizar los datos existe la tendencia a **replicar** (o duplicar) los datos para satisfacer las nuevas necesidades.
- Replicación incluye agrupación de datos, combinación de diversas fuentes, migración a estructuras de datos diferentes de las originales y adición de series históricas (o de tiempo).
- Datos replicados son fuente de error.

Integración

- Además de replicar existe la necesidad de integrar los datos de diversas bases de datos en aplicaciones interactivas.
- La integración usualmente implica traslado a una estructura de base de datos diferente
- En todos estos procesos, de nuevos usos, replicación o integración, existe el riesgo de dañar la calidad de las bases de datos.

Integración y replicación de bases de datos operacionales



Errores en sistemas transaccionales

- Todos los sistemas transaccionales, en una forma u otra y en mayor o menor grado, contienen defectos en sus datos.
- Por lo general las organizaciones administran estos errores reduciendo los efectos negativos en los clientes y en las operaciones.

El efecto en los sistemas de soporte de decisiones

- En los sistemas transaccionales, un valor errado tiene muy poco, o no tiene, impacto.
- Pero esos valores errados se propagan a los sistemas de soporte de decisiones y su efecto es mucho mayor.
  - El efecto acumulativo de muchos valores errados en el mismo atributo puede causar resultados indeseados.

Resumen

- El problema de calidad de datos es universal y resulta de la naturaleza cambiante de los procesos de información.
- Nuevas aplicaciones (nuevos usos), integración de datos y replicaciones afectan la calidad de datos.
- Los errores de datos en los sistemas transaccionales se trasladan con efectos impredecibles a los sistemas de soporte de decisiones.

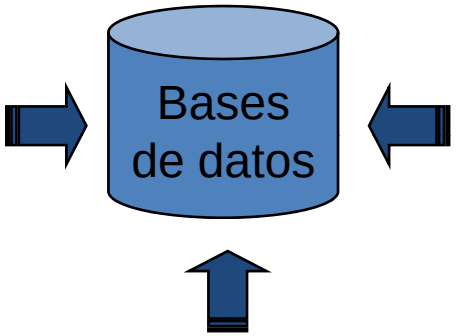
Procesos que afectan la calidad de datos

II

Procesos que afectan la calidad de datos

Procesos externos

- Migración de datos
- Consolidación de sistemas
- Entrada manual de datos
- Alimentación por lotes
- Interfases en tiempo real



Procesos internos

- Procesamiento de datos
- Limpieza de datos
- Depuración de datos

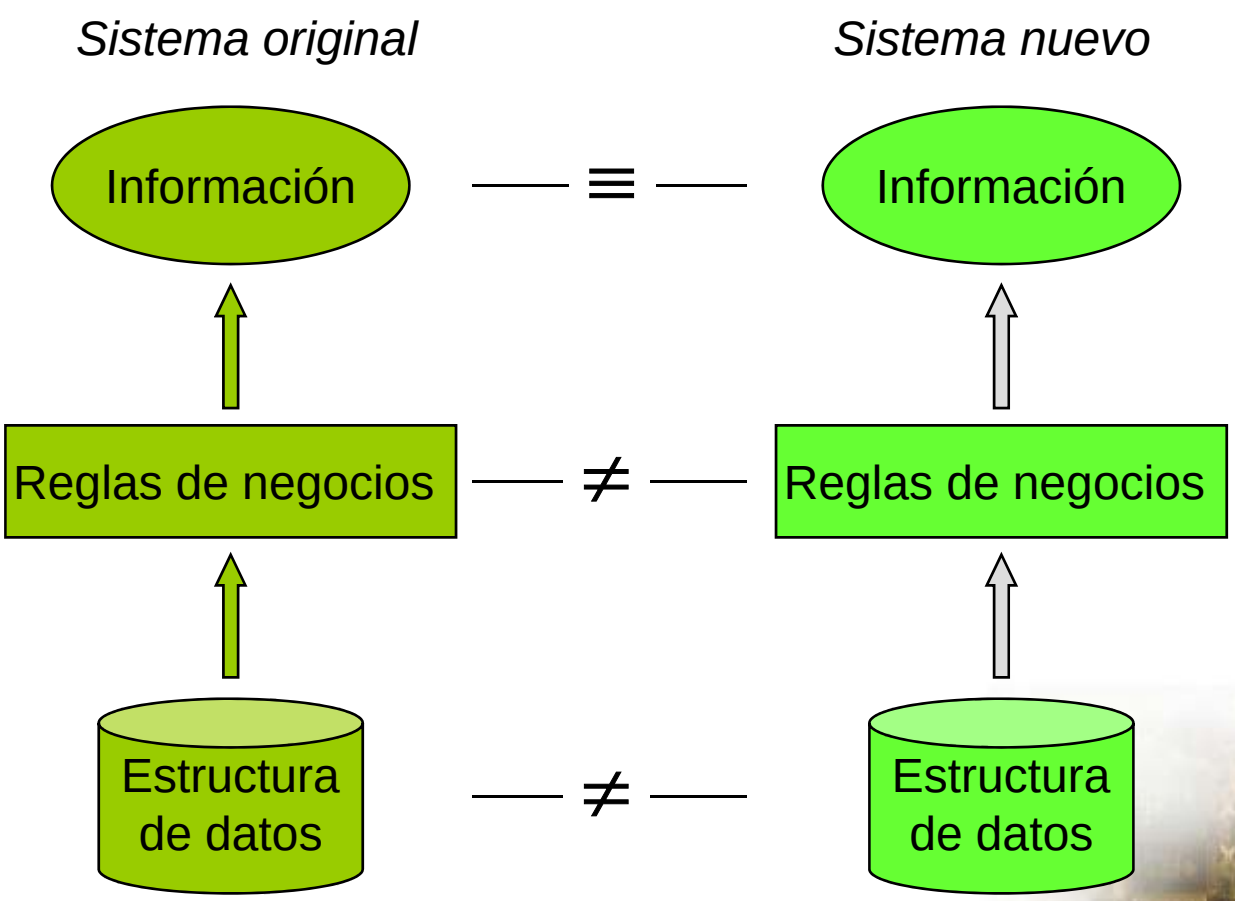
Deterioro natural

- Nuevos usos de los datos
- Cambios no registrados
- Actualización de los sistemas
- Pérdida de conocimiento *expertise*)
- Automatización de procesos

Migraciones

- Migración de los datos de un sistema, legado o antiguo, a un nuevo sistema.
- El proceso requiere establecer la correspondencia entre la estructura original y la nueva estructura. En teoría el problema es trivial pero en la práctica presenta muchas dificultades por algunas de las siguientes razones:
  - **Metadata** incompleta
  - Condiciones específicas incorporadas en el código (del programa)
  - Valores faltantes, o nulos
  - Las reglas de negocios del sistema nuevo seguramente son diferentes a las del sistema antiguo.
  - Con frecuencia hay duplicación de datos

Dificultades en la conversión de datos



Consolidaciones

- Las consolidaciones son parecidas a las migraciones, pero de mucha mayor complejidad:
  - Usualmente los datos de la fuente se trasladan a una BD que ya contiene información, lo cual genera toda clase de conflictos de datos (duplicados, series de tiempo, etc.)
- Cuando ocurren, son una de las principales causas de problemas de calidad de datos

Entrada manual

- Una cantidad significativa de los datos de una organización entra a las BD en forma manual, por formularios o interfases.
- Algunas de las principales causas de error son:
  - Captura errada del valor
  - Formularios e interfases Web con fallas en el diseño que inducen a registrar errores.
  - Valores faltantes
  - Valores por defecto (*default*)
  - Falta de instrucciones adecuadas (metadata)

Cargas por lotes [1]

- Los procesos “batch” se utilizan regularmente para intercambiar (o cargar) datos entre sistemas.
  - Mucha información entra a las bases de datos de la organización de esta manera.
- Después de consolidaciones y migraciones, estos procesos generan la mayor cantidad de problemas de calidad de datos.

Cargas por lotes [2]

- Las razones son las siguientes:
  - Los procesos “batch” sufren frecuentes cambios estructurales, actualizaciones y mejoras.
  - Usualmente no se someten a pruebas regresivas (*regression testing)* y aseguramiento de calidad (QA) porque no hay tiempo suficiente y por la dificultad de hacerlo.
  - Los procesos “batch” propagan los errores por múltiples bases de datos (más o menos como un virus)

Interfases en tiempo real

- En la actualidad los sistemas intercambian muchos datos con interfases en tiempo real.
- Esto permite tener la información sincronizada y es de alto valor para la organización pero no da tiempo para verificar que los datos sean correctos.
  - En tiempo real, la transacción (o el dato) se acepta o se rechaza.
  - Además, no es fácil determinar si el dato recibido es correcto porque usualmente se intercambian bloques pequeños de datos, fuera de contexto y sin suficiente información para detectar errores.
  - El potencial para generar errores es mayor que en los procesos “batch” y debe ser evaluado cuando se cambian los sistemas.

Pérdida de saber (expertise, know-how)

- Muchos detalles importantes sobre el significado de los datos, particularmente en aplicaciones legadas, no están documentados y sólo los conocen una pocas personas (fallas en la metadata)
- Ausencia temporal o permanente de los expertos en los datos conduce al uso inapropiado y afecta la calidad de los datos.
- Es una forma de deterioro de datos.

Resumen

- Las causas de problemas de calidad de datos son muy variadas y continuas.
- Las de mayor impacto son las consolidaciones y migraciones, pero la entrada, las interfases y el deterioro actúan continuamente.

Calidad de datos y la definición de exactitud

Las dimensiones de la calidad de datos

Dimensiones básicas *

- Para satisfacer su propósito, los datos deben ser:
  - Exactos (correctos)
  - Oportunos
  - Relevantes
  - Completos
  - Entendibles (inteligibles)
  - Confiables

* [OLSO02]

Exactitud de los datos

- La **exactitud** de los datos es sólo una de las dimensiones de la calidad de datos, pero es una condición necesaria (aunque no suficiente) y el componente más importante.
- Si los datos están errados, faltan o presentan inconsistencias, no es posible lograr calidad de datos.
- Cualquier programa de mejoramiento de la calidad debe iniciar asegurando la exactitud de los datos.

Exactitud de los datos

- Para ser exacto, un dato debe tener el valor correcto y estar representado de manera consistente e inequívoca:
  - **Correcto**
  - **Consistente**
  - **Inequívoco**

Características de la exactitud

- Exactitud tiene dos características:
  - Forma y
  - Contenido
- La forma es importante porque elimina ambigüedades sobre el contenido.
- Un valor no es exacto si el usuario del valor no puede determinar que es o que significa.

Consistencia en la representación del valor

- La consistencia es parte de la exactitud.
- Inconsistencia se refiere a valores diferentes que representan lo mismo.
- Los valores inconsistentes no se pueden agregar o comparar correctamente.

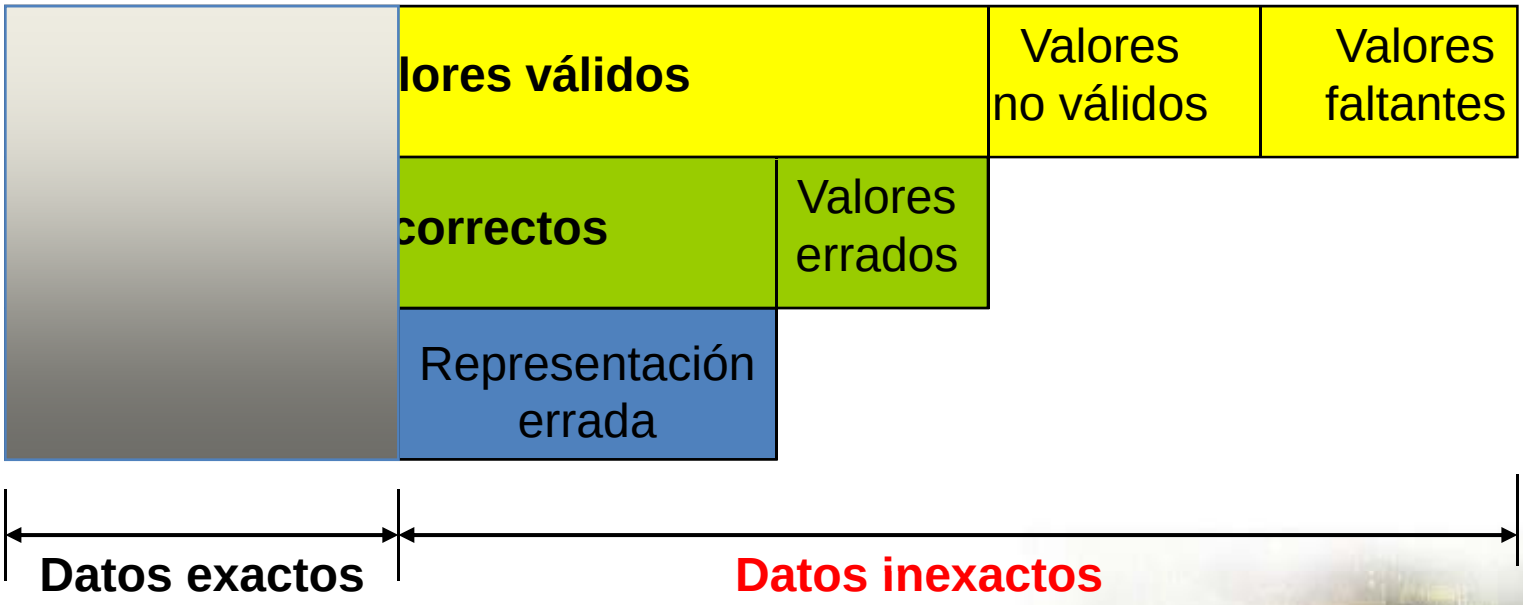
Valores válidos

- Un valor es válido si es elemento del conjunto de posibles valores correctos y se representa en forma consistente e inequívoca.
- Un valor válido no es necesariamente correcto, pero el valor correcto siempre es válido.

Valores faltantes

- Los valores faltantes son causa de errores en los datos; su significado es ambigüo.
- Un dato sin valor puede ser correcto o errado.
- Los valores faltantes se deben evitar en los procesos de creación de datos.
- Lo correcto es distinguir entre “blanco” (no hay valor) y “nulo” (no se conoce el valor).

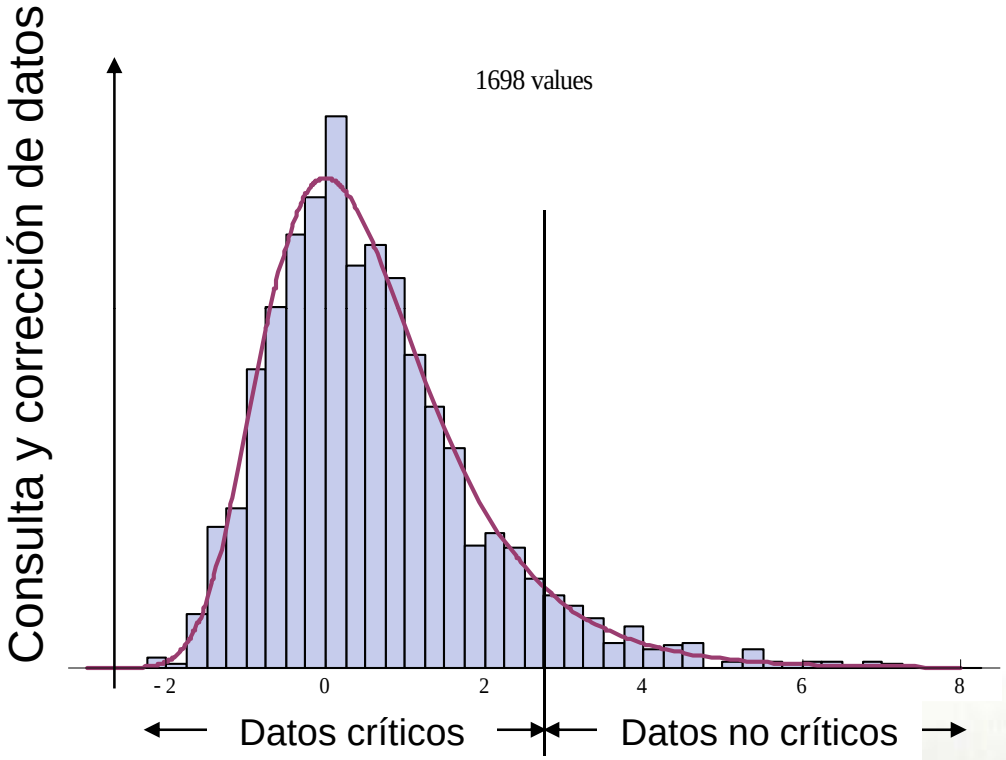
Datos exactos e inexactos



Distribución de los errores

- La distribución de los errores en la base de datos no es uniforme
  - Unos datos son más importantes que otros
  - Hay tendencia a corregir los datos importantes errados más que otros datos
  - El uso de un dato errado mejora la probabilidad de que el error sea detectado y corregido.
  - Fallas en la captura de los datos no es igual para todos.

Distribución de errores



El efecto de la distribución …

- La tendencia de datos más importantes a ser más exactos es la razón principal por la cual los problemas de calidad de datos no son [tan] evidentes en las aplicaciones transaccionales.
  - La calidad es aceptable para satisfacer los requerimientos del negocio
- Los problemas de inexactitud se manifiestan cuando los datos se mueven y se utilizan para tomar decisiones (en sistemas de soporte de decisiones - DSS)
  - Muchos datos utilizados para registrar información "secundaria" sobre la transacción ahora cobran importancia.

¿Cómo identificar los valores errados?

- La mayoría de los errores se pueden identificar. No es probable hallar la totalidad.
- Hay dos alternativas para encontrar los datos errados:
  - Verificación manual
    - Sólo verificación manual puede, en teoría, localizar la totalidad de los errores.
  - Análisis automático

Verificación manual

- Manualmente, con base en la fuente original de la información, se verifican todos y cada uno de los valores.
  - Es la única manera de determinar que valores son correctos y cuales incorrectos
  - Las técnicas analíticas no pueden determinar si un valor es correcto al menos que puedan consultar una fuente alterna para confirmar el valor

Revisión manual

- El proceso manual es susceptible de error y no garantiza la detección total.
- Es muy demorado y costoso.
- En algunos casos no es posible aplicarlo.
- Para la mayoría de los casos no es práctico.
- Se puede hacer verificación selectiva para mejorar la confiabilidad de la calidad de los datos.

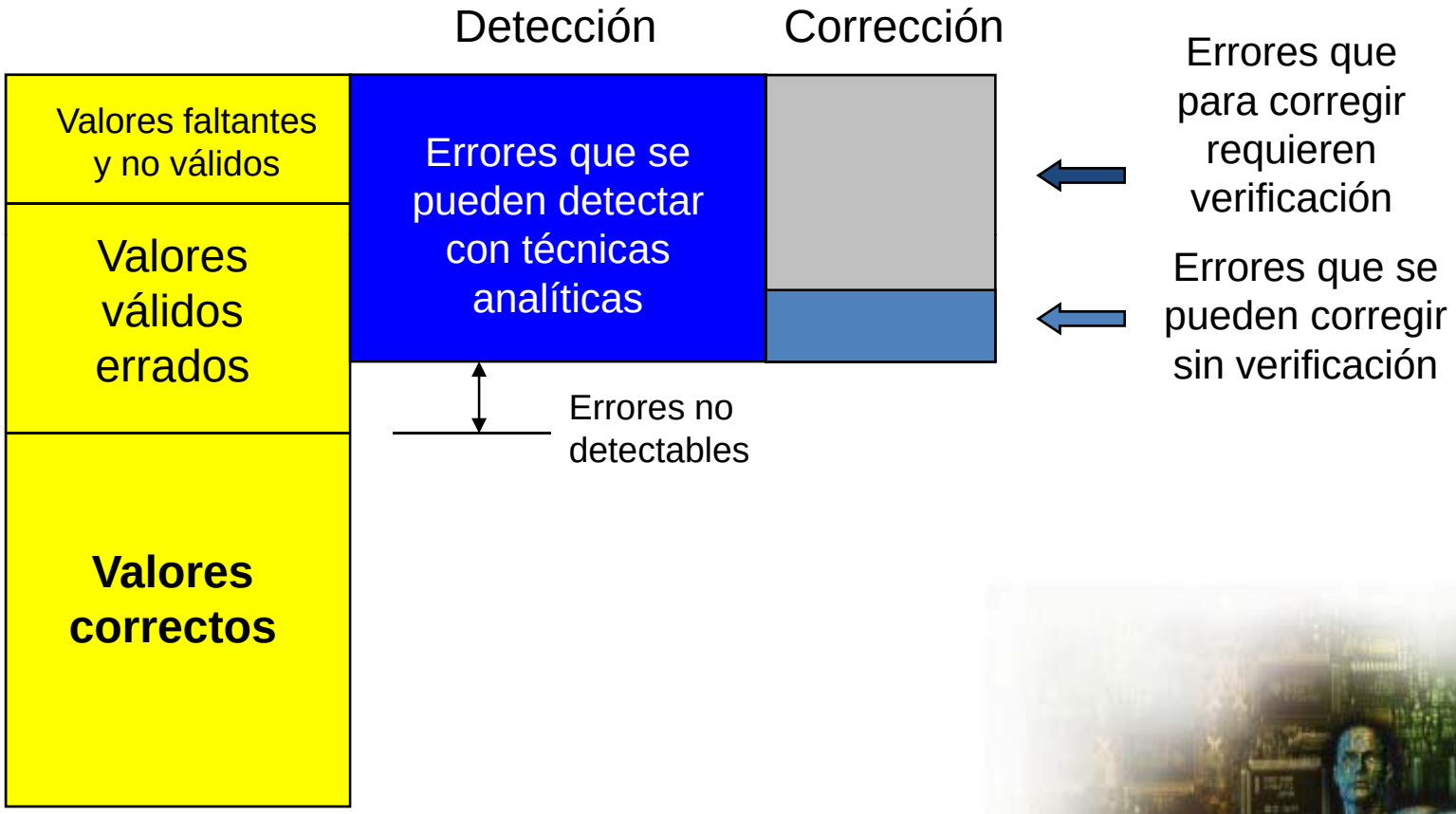
Técnicas analíticas

- Utilizan software y la habilidad del analista de calidad de datos para detectar los datos inexactos.
- Las técnicas analíticas se pueden aplicar a:
  - Transacciones que están ocurriendo
  - Bases de datos que están cambiando
  - Bases de datos en producción, periódicamente
- Existen 4 categorías de análisis que se pueden aplicar a los datos:
  - Análisis de elementos (datos)
  - Análisis estructural
  - Análisis de reglas de negocio
  - Análisis estadístico

Aplicación de técnicas analíticas

- Las técnicas analíticas, bien aplicadas, identifican suficientes errores para dar una idea clara del estado de calidad de los datos.
- No pueden detectar todas las inexactitudes en los datos de una BD.
- Sin embargo, un programa continuo de mejoramiento de la calidad de los datos logra resultados satisfactorios.

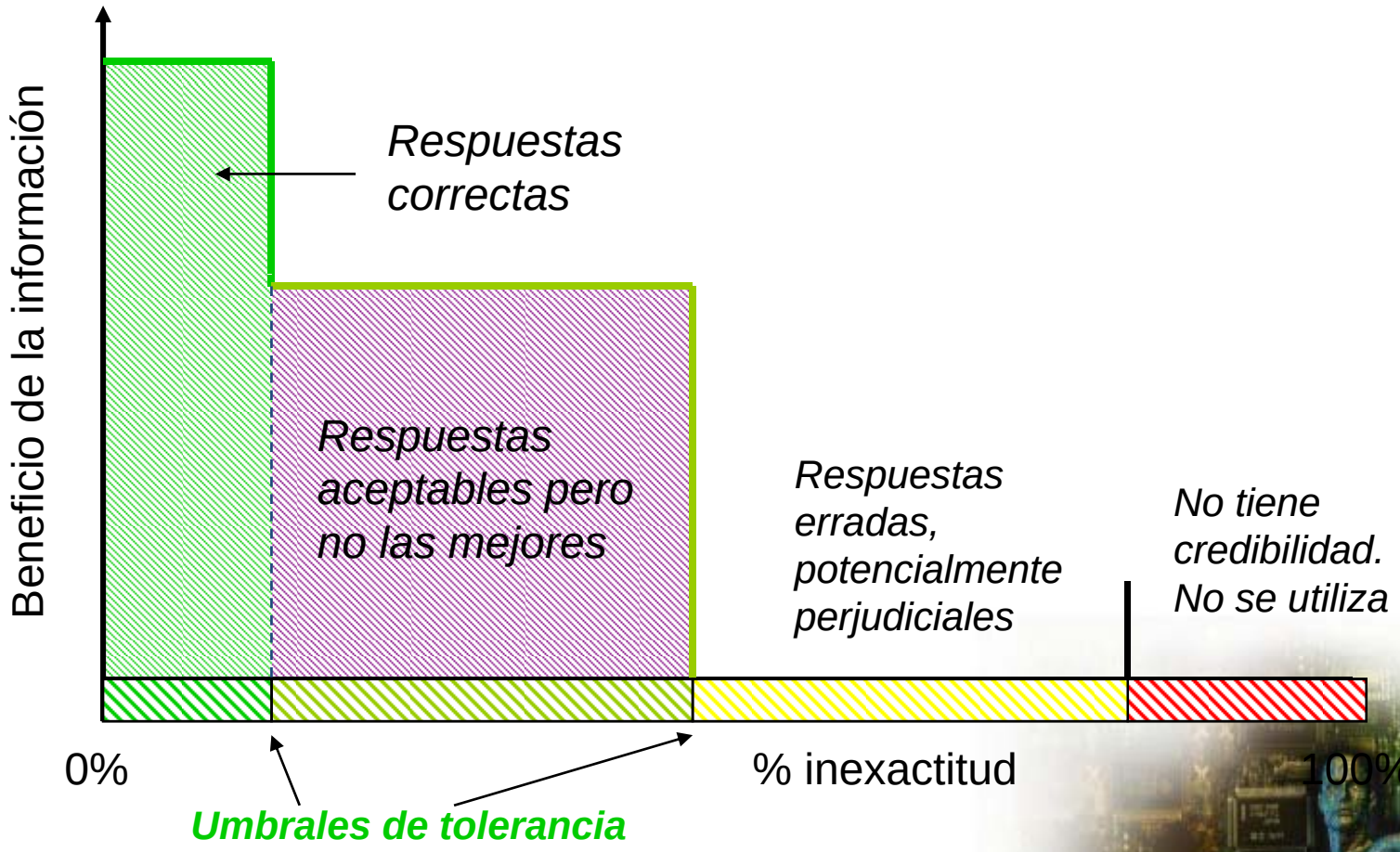
Detección de errores



Grados de tolerancia de la calidad de datos

- La mayoría de las aplicaciones, incluyendo los sistemas de soporte de decisiones, tienen algún grado de tolerancia a la inexactitud de los datos.

Los umbrales de tolerancia



Márgenes de tolerancia

- Inexactitudes hasta el umbral de tolerancia permiten tomar decisiones de alta calidad.
- No es necesario lograr exactitud del 100%
- Si la calidad de los datos excede el umbral de tolerancia, los datos pueden causar decisiones erradas, pero difíciles de notar porque las decisiones no son “tan malas”. Esta es una situación precaria.
- A mayores niveles de inexactitud, los datos pierden credibilidad y no se usan para tomar decisiones.

La toma de decisiones y la calidad de los datos

- La eficiencia de la toma de decisiones depende de la calidad de datos, de tal manera que pequeñas mejoras en la exactitud de los datos puede conducir a mejoras sustanciales en la información para toma de decisiones.
  - Esto representa beneficios importantes para la organización.

Resumen

- La exactitud de los datos es la más visible e importante dimensión de calidad de datos.
  - Es la más tangible de tratar,
  - Más fácil de mejorar,
  - Usualmente no requiere reingeniería de procesos
  - No requiere reestructuración de la organización
- No se puede lograr calidad total, pero sí se puede mejorar la calidad al punto que la información sea adecuada para la toma de decisiones.

Data Profiling

El proceso de evaluación

¿Qué es?

- *Data profiling* es el proceso de reconstruir el conjunto de rasgos particulares que caracterizan los datos de una base de datos
  - Se examinan y se documentan las características de los datos
- Consiste en la aplicación de técnicas analíticas a repositorios de datos con el propósito de determinar:
  - el **contenido** actual,
  - la **estructura** y
  - la **calidad** de los datos.

¿Cómo lo hace?

- *Data Profiling* utiliza dos métodos diferentes para analizar los datos:
  - Descubrimiento: con software, se revelan las características de los datos a partir de los mismos.
    - Es análogo a hacer *data mining* para reconstruir la metadata.
  - Pruebas asertivas: se formulan condiciones verdaderas (reglas) sobre los datos y se prueban con el software.
    - Permite determinar donde difieren los datos de la metadata y corregirla

Aplicación a calidad de datos

- La técnica se utiliza para deducir información sobre los propios datos.
- En el contexto de aseguramiento de calidad de datos, es el proceso utilizado para descubrir (o detectar) errores o inexactitudes en una base de datos.
- Es la herramienta esencial para evaluar o diagnosticar la calidad de una base de datos.

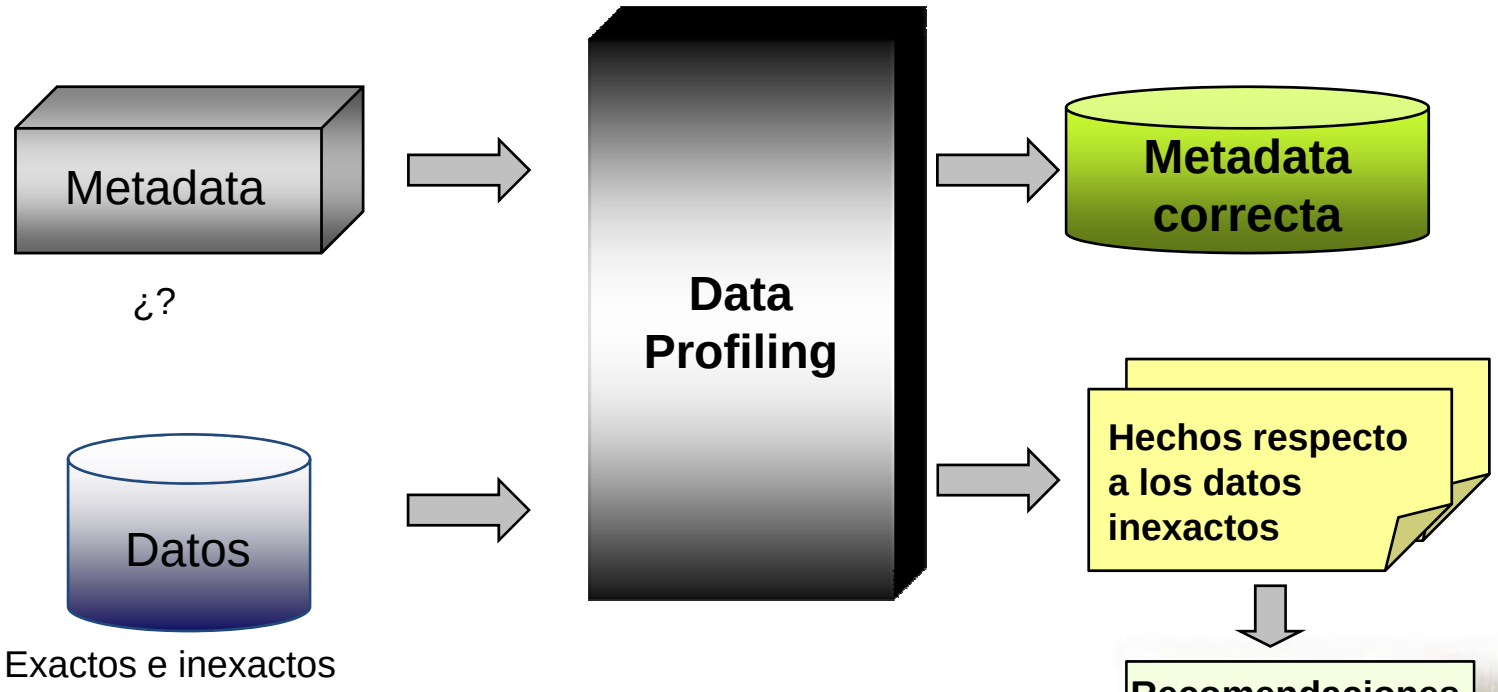
Tradicionalmente …

- Los analistas de datos han utilizado por muchos años métodos *ad hoc* (no formales, con un propósito específico) para examinar y evaluar los datos.
  - Sin una metodología formal y apropiada, y sin herramientas analíticas diseñadas específicamente para hacer el diagnóstico, el proceso es muy dispendioso y no es efectivo.

Tecnología formal

- El proceso de *data profiling* ha evolucionado y madurado a una tecnología formal y efectiva que utiliza un método inductivo para la evaluación de la calidad de datos.

El proceso


Metadata
¿?
Datos
Exactos e inexactos
Data Profiling
Metadata correcta
Hechos respecto a los datos inexactos
Recomendaciones sobre el estado de la calidad de datos pendientes de resolver

Resultados

- El proceso reconstruye la metadata a partir del contenido real de la base de datos.
- Estado de la calidad de los datos en la base de datos, sobre lo cual se formulan recomendaciones.
- No corrige datos; sólo diagnostica e identifica anomalías.
  - Documentadas en el repositorio de metadata

Metodología para la evaluación

- Utiliza 4 pasos:
  1. Análisis de elementos (propiedades de columnas)
  2. Análisis de la estructura (dependencias funcionales, sinónimos, reglas de integridad)
  3. Verificación de reglas de negocios
     - Simples
     - Compuestas
  4. Análisis estadístico

Pasos del proceso

1. Análisis de elementos

- Se examinan los valores individuales de cada columna de cada tabla para determinar si son válidos.
  - Requiere una definición de qué es válido y que no es válido.
- Analizando los tipos, longitud, rangos, valores discretos, patrones, formatos, etc. se determinan los rasgos de las columnas.
- El proceso automático se complementa con inspecciones visuales que pueden detectar errores imposibles de hallar por software.
- La técnica **sólo identifica valores no válidos**. No puede determinar si un valor es correcto.

2. Análisis de la estructura

- Consiste en identificar
  - las dependencias funcionales en cada tabla,
  - hallar sinónimos (pares de columnas que representan el mismo objeto de negocios), en cada tabla y entre tablas;
  - examinar llaves primarias y llaves foráneas (verificar reglas de integridad).
- Construir modelo de datos en 3NF (tercera forma normal).
- Este análisis permite aislar el **error en un subconjunto de registros**, pero no identifica los valores errados (para eso es necesaria la verificación manual)

3. Análisis de reglas de negocio simples

A. Análisis de reglas de negocio aplicables a un objeto de negocios (usualmente varias columnas de una tabla).

- Consiste en analizar conjuntos de valores con una regla específica que aplica para varios datos.
  - Cuando la regla detecta inconsistencia no se puede saber donde está el error salvo que se identifique (por lo menos) un dato errado
  - Si la regla compara dos datos y muestra inconsistencia, no indica cual es el dato incorrecto; los dos pueden estar errados.
    - O los datos son correctos pero la violación resulta de una actividad del negocio que no cumple con la regla.
  - Por lo general se formulan muchas (cientos) reglas para correlacionar los valores y asegurar que el conjunto es coherente y válido.
- **No permite determinar cual es el valor errado**

4. Análisis de reglas de negocio compuestas

B. Análisis de reglas de negocio aplicadas a varios objetos de negocios

- Se formulan reglas que se utilizan para identificar la presencia de errores en valores agregados sobre grandes volúmenes de datos.
  - Violación de las reglas indican que faltan datos o que estos tienen errores.
  - O los datos pueden estar errados, o los datos son correctos pero la violación resulta de una actividad del negocio que no cumple con la regla.
- No identifica los valores errados.

5. Análisis estadístico

- Aplicable a casos donde no es posible formular una regla concreta y complementa los análisis anteriores.
- Con base en estadísticas (distribución de frecuencias, conteos, sumas, promedios, valores extremos, etc.) se puede determinar si los resultados son razonables o ilógicos.

En síntesis …

- Análisis de elementos sólo permite hallar valores no válidos.
- Análisis estructural, análisis de reglas de negocio y análisis estádistico permiten hallar inexactitudes entre valores válidos.
  - No se pueden identificar los valores errados pero sí determinar, con certeza, que existen valores errados.
- Nota: los datos pueden pasar todas las pruebas y aún así estar errados!

¿Cuándo se debe hacer *Data Profiling*?

- En todos los proyectos de diagnóstico, evaluación o mejoramiento de calidad de datos.
- En todos los proyectos de TI que trasladan datos a otras estructuras, migran o consolidan datos.
- Las bases de datos importantes de la organización se deben “perfilar” periódicamente.

Conclusiones

- El proceso de *data profiling*, si se hace correctamente, es una técnica efectiva que contribuye significativamente a mejorar la calidad de los datos de la organización.
- Utilizada adecuadamente puede reducir los ciclos de implementación de proyectos críticos en varios meses y mejorar el conocimiento de los usuarios respecto a los datos.
- Debe ser una competencia central de tecnología en la organización (*core competency technology)*