Materia Laboratorio de Datos

Práctica Calidad de Datos

Ejercicio 1

He aquí algunas situaciones reales en que la mala calidad de los datos trajo pérdidas económicas, algunas de ellas fácilmente cuantificables (una vez producidas).

Caso 1. Pozo petrolero perforado en ubicación errónea por interpretación equivocada del sistema de coordenadas en uso. La empresa fue multada.

Caso 2. Un banco local fue condenado a pagar a un cliente indemnizaciones por cientos de miles de pesos por haber sido incluido erróneamente en bases de datos de morosos.

Fuente: Diario Clarín 14/02/2003
http://www.clarin.com/diario/2003/02/14/e-01001.htm

Caso 3. En un organismo del gobierno de un país latinoamericano se mandaron cartas a todas aquellas empresas beneficiadas por una norma. El 30% de la correspondencia volvió rechazada por problemas en la dirección.

a) Para cada uno de los casos:

a. Identifique quiénes fueron los afectados en los problemas (usuarios o clientes, managers que hacen uso de los datos, desarrolladores o encargados de mantenimiento de los sistemas, otros).
b. ¿Qué impacto identifica en estos casos (además del económico descripto)? (descreimiento en la organización, causa de costos innecesarios, impacto en toma de decisiones, disminución de satisfacción de usuarios y clientes).

b) Describa algún inconveniente de Calidad de Datos que lo haya afectado en su vida personal y/o alguno que haya detectado a nivel laboral.

hay múltiples afectados. Primeramentela empresa resulta la directamente afectada, ya que debidoa su error no so

Ejercicio 2

De al menos dos ejemplos de sistemas o conjuntos de sistemas con pocos bugs, pero que permitan el almacenamiento de información con problemas de calidad.

Del gobierno

Instancia: Como están los datos?; consistencia de los datos, precisión y actualización de estos-->Relevancia,Confiabilidad?consistenci

Ejercicio 3

Dados los siguientes inconvenientes clasifíquelos según el origen de los mismos (instancia, proceso, modelo, software):

a) datos obligatorios que no se asumen como tales y por lo tanto no se cargan SOFTWARE
b) interfaces poco amigables SOFTWARE
c) rangos de valores que no se respetan INSTANCIA
d) distintas personas cargan la misma información haciendo distintas asunciones PROCESOS
e) gente que hace modificaciones pero no debería estar autorizada para hacerlas PROCESOS
f) hay información que no está presente porque no hay forma de almacenarla MODELO
g) el mundo que se quiere representar evolucionó, pero esta situación no se ve reflejada en el sistema. MODELO
h) datos que han cambiado en el mundo real, y que no fueron actualizados INSTANCIAA
i) datos que provienen de distintas fuentes y que no son consistentes INSTANCIA
j) datos de años que han sido almacenados con dos dígitos en lugar de cuatro. INSTANCIA
k) Posibles valores completados en el campo región: INSTANCIA
   - ANETOFAGASTA
   - ANMTOFAGASTA
   - ANT0FAGASTA
   - ANTO9FAGASTA
   - ANTOAFAGASTA
   - ANTOFAAGASTA

Ejercicio 4

Dados los siguientes problemas clasifíquelos en función del atributo de calidad que se ve afectado y a si es problema de modelo o de datos.

a) No se cargan unidades de medida en que se midió la profundidad de un pozo.
b) No es posible almacenar el sistema de referencia.
c) Hay inconsistencias entre nombres de un mismo pozo en distintos sistemas.
d) La ubicación de una central telefónica no coincide con la ubicación real.
e) El nombre de un pozo no corresponde al que debería ser de acuerdo a la ley.
f) Hay personas fallecidas que figuran como empleados participantes de cursos (por los cuáles la empresa que los informa consigue una exención impositiva)
g) Las direcciones de los clientes no están actualizadas

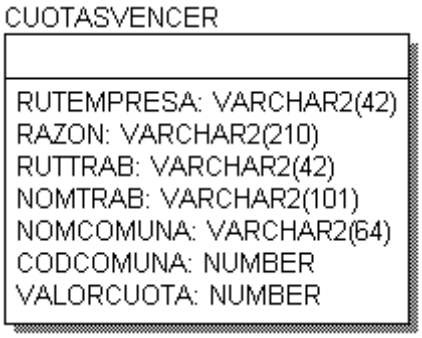
En los casos del ejercicio anterior originados por instancia y modelo, mencione qué atributos de calidad se ven afectados

Ejercicio 5

A modo de repaso de temas vistos anteriormente

a) ¿Qué problemas encuentra en el diseño de la tabla que figura a continuación?
b) ¿Qué tipo de anomalías produce?
c) ¿Qué problemas de calidad de la información puede traer?
d) ¿Qué otros problemas de diseño de una base de datos cree que pueden afectar la calidad de la información? Relaciónelos con los atributos de calidad del modelo.
e) ¿En los casos en que adrede se deje información redundante en una tabla, cómo recomienda proceder para evitar problemas de calidad de la información?

Nota: RUT es el Rol único Tributario (el número con el cuál se identifica a las personas físicas y jurídicas en Chile).



A continuación se muestra un posible conjunto de datos de esta tabla.

RUTEMPRESA	RAZON	RUTTRAB	NOMTRAB	NOMCOMUNA	CODCOMUNA	VALORCUOTA
2178645-4	Servando Humberto Arriagada Peres	10734185-4	LUIS ALFREDO CASTILLO	TEMUCO	93801	32000
2178645-4	Servando Humberto Arriagada Perez	12192576-1	Cesar Enrique Castillo	TEMUCO	93802	32000

Ejercicio 6

a) Defina métricas según el modelo Goal Question Metric (GQM) para identificar cantidad de:
   - empresas sin dirección almacenada en el sistema,
   - empresas que parecerían estar más de una vez en la base de datos (identificando por nombre y dirección)

   En todos los casos identifique el objetivo, la pregunta y la métrica.
b) ¿En algún caso le puede ser de utilidad el uso de algún algoritmo de matching para ejecutar la métrica?

-Determinar en qué casos es conveniente el uso de algoritmos específicos (por ej. soundex y keyboard distance) para la ejecución de las métricas.

Ejercicio 7

Tomar el dataset corregido de Dengue (del campus virtual) y listar los nombres de departamentos y sus ids, nombres provincia e id provincia para todos aquellos departamentos con mismo nombre, pero distinto id de departamento y distinto id de provincia. Ordernarlos por nombre de departamento.