**Laboratorio de Datos - Guía de ejercicios**

**Calidad de Datos**

**Consignas:**

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 una 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 [Un banco debe pagar $ 120.000 por incluir mal a un cliente en Veraz](https://www.clarin.com/economia/banco-debe-pagar-120000-incluir-mal-cliente-veraz_0_rJ4juMlAFg.html)

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:

i. Identificar quiénes fueron los afectados en cada situación (usuarios o clientes, managers que hacen uso de los datos, desarrolladores o encargados de mantenimiento de los sistemas, otros)

ii. ¿Qué impacto identifica en estos casos (además del económico descrito)? (descreimiento en la organización, causa de costos innecesarios, impacto en toma de decisiones, disminución de satisfacción de usuarios y clientes, etc.)

b) Describir algún inconveniente de Calidad de Datos que lo haya afectado en su vida

Ejercicio 2

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

b. Interfaces poco amigables

c. Rangos de valores que no se respetan

d. Distintas personas cargan la misma información haciendo distintas asunciones

e. Gente que hace modificaciones pero no debería estar autorizada para hacerlas

f. Hay información que no está presente porque no hay forma de almacenarla

g. El mundo que se quiere representar evolucionó, pero esta situación no se ve reflejada en el sistema.

h. Datos que han cambiado en el mundo real, y que no fueron actualizados

i. Datos que provienen de distintas fuentes y que no son consistentes

j. Datos correspondiente al año, que han sido almacenados con dos dígitos en lugar de cuatro.

k. Posibles valores completados en el campo región:

● ANETOFAGASTA

● ANMTOFAGASTA

● ANT0FAGASTA

● ANTO9FAGASTA

● ANTOAFAGASTA

● ANTOFAAGASTA

Ejercicio 3

Dados los siguientes problemas, i) clasificarlos en función del atributo de calidad que se ve afectado; ii) determinar si el problema es de modelo o de datos.

a. No se cargan unidades de medida en que se midió la profundidad de un pozo petrolero

b. No es posible almacenar el sistema de referencia

c. Hay inconsistencias entre nombres de un mismo pozo petrolero en distintos sistemas

d. La ubicación de una central telefónica no coincide con la ubicación real

e. El nombre de un pozo petrolero no corresponde con el 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 3 originados por instancia o modelo, mencione qué atributos de calidad se ven afectados.

Ejercicio 4

A modo de repaso de temas vistos anteriormente

a. ¿Qué problemas de diseño encuentra en la tabla que figura a continuación?

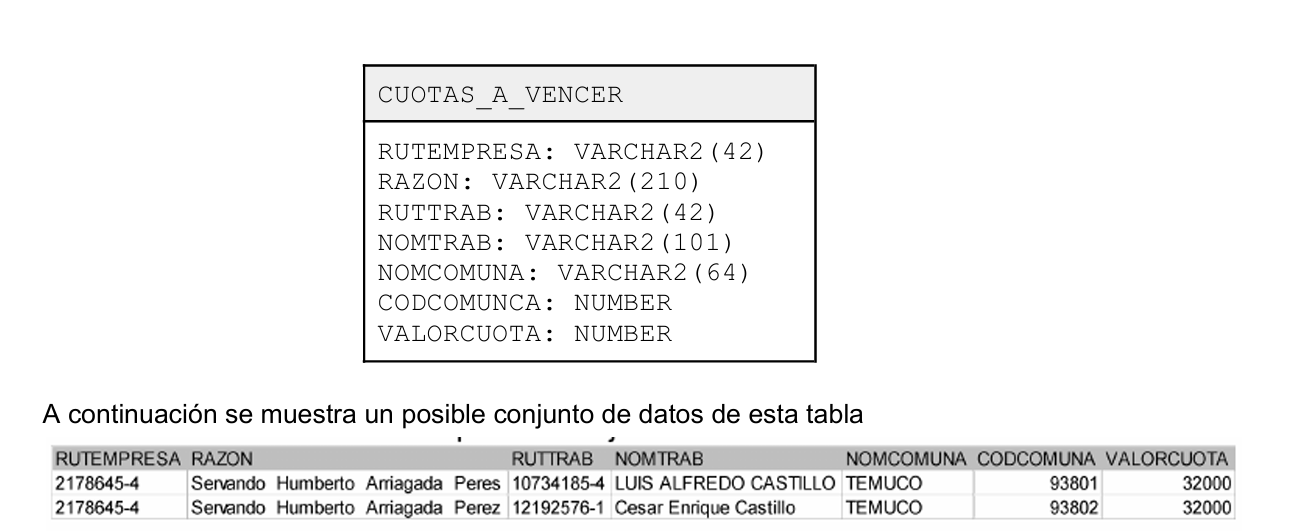
b. ¿Qué tipo de anomalías produce?

c. ¿Qué problemas de calidad de datos puede acarrear?

d. ¿Qué otros problemas de diseño (de la base de datos) cree que pueden afectar la calidad de los datos? Relacionarlos con los atributos de calidad del modelo

e. ¿En los casos en que se deje información redundante en una tabla adrede, cómo recomienda proceder para evitar problemas en la calidad de los datos?

Nota: RUT es el Rol Único Tributario (el número con el cuál se identifica a las personas físicas y jurídicas en Chile, similar a nuestro CUIT/CUIL).



Ejercicio 5

Un cliente desea conocer la calidad de sus datos en cuanto a:

a. Empresas sin dirección que posee almacenadas en su sistema

b. Empresas que parecerían estar almacenadas más de una vez en su sistema (puede asumir que se identifican por nombre más dirección)

1. Definir métricas según el modelo Goal Question Metric (GQM) para dar soporte al cliente.

Recuerde que en todos los casos debe identificar el objetivo, la pregunta y la métrica.

2. ¿En algún caso le puede ser de utilidad el uso de algún algoritmo de matching para ejecutar la métrica?

**Respuestas:**

Ejercicio 1

Primer Caso:

1. Los afectados fueron los dueños de la empresa petrolífera, así como también los del equipo de planificación estratégica (pues si se contaba con la disponibilidad del recurso, entonces habrá problemas en los planes originales); y dependiendo de la locación del pozo, la generalidad del Estado-Nación por impacto/daño en zonas públicas/estatales.
2. Además del gran costo innecesario producido, hay un detraimiento de la imagen de la empresa pues se le podrían quitar permisos (o se la inhabilite en ciertos espacios), o no se la quiera contratar (pérdida de clientela) por un mal accionar.

Segundo caso:

1. El afectado fue el cliente pues se le privó de la realización de variadas actividades económicas y financieras tras ser puesto en el Veraz; pero además los desarrolladores y administradores de la BD, pues ésta situación implica un análisis profundo del historial de cada cliente para verificar o cerciorar que se esté guardando y manipulando su información correctamente, y también genera una necesidad de renovación/rediseño de la BD.
2. Además de la pérdida por indemnización, el banco se halló con una disminución de satisfacción de usuarios y clientes.

Tercer caso:

1. Se ven afectados el Estado-Nación por la generación de un costo innecesario, las empresas cuya norma les alcanzaba pero no fueron notificadas debidamente y los encargados de mantenimiento de la BD pues esto implica la modificación de un 30% de cierta parte de la información almacenada.
2. Además del gran costo innecesario, esto genera en las empresas un rechazo hacia la administración pública/gobierno de turno, lo que implicará menor inversión en el país; así como también decae la satisfacción de los ciudadanos a con su dirigencia por el mal uso de sus recursos.

Ejercicio 2

Problemas asociados a errores de SOFTWARE: en esta categoría, entran aquellos problemas de calidad de datos derivados de fallos técnicos en las aplicaciones o sistemas encargados del manejo de la información. Es entonces que los siguientes problemas:

a. Datos obligatorios que no se asumen como tales y por lo tanto no se cargan

b. Interfaces poco amigables

están asociados a ésta categoría.

Problemas asociados al MODELO DE DATOS: en ésta categoría, entran aquellos problemas de calidad de datos derivados de una deficiencia en el diseño del esquema que define cómo se organizan y relacionan los datos. Es entonces que los siguientes problemas:

f. Hay información que no está presente porque no hay forma de almacenarla

g. El mundo que se quiere representar evolucionó, pero esta situación no se ve reflejada en el sistema.

están asociados a ésta categoría.

Problemas asociados a los PROCESOS: en ésta categoría, entran aquellos problemas de calidad de datos derivadas de protocolos de generación, transformación y gestión de datos defectuosos; errores en los procesos de captura, depuración, integración y control de los datos. Es entonces que los siguientes problemas:

c. Rangos de valores que no se respetan

d. Distintas personas cargan la misma información haciendo distintas asunciones

e. Gente que hace modificaciones pero no debería estar autorizada para hacerlas están asociados a ésta categoría.

Problemas asociados a la INSTANCIA: en ésta categoría, entran aquellos problemas de calidad de datos derivados de errores que afectan a registros individuales en la BD, enmarcando incidencias como datos faltantes, valores duplicados, errores tipográficos o formatos inadecuados. Es entonces que los siguientes problemas:

h. Datos que han cambiado en el mundo real, y que no fueron actualizados

i. Datos que provienen de distintas fuentes y que no son consistentes

j. Datos correspondiente al año, que han sido almacenados con dos dígitos en lugar de cuatro.

k. Posibles valores completados en el campo región: ANETOFAGASTA, ANMTOFAGASTA, ANT0FAGASTA, ANTO9FAGASTA, ANTOAFAGASTA, ANTOFAAGASTA.

están asociados a ésta categoría.

Ejercicio 3

Atributo afectado: COMPLETITUD

Éste atributo se define como la presencia de:

1. todos los valores para representar la realidad.
2. todas las instancias existentes en el mundo real.

Es por ello que:

“b. No es posible almacenar el sistema de referencia” afecta éste atributo; y el problema surge del modelo de datos.

En el ejercicio 3, “g. El mundo que se quiere representar evolucionó, pero esta situación no se ve reflejada en el sistema.” y “f. Hay información que no está presente porque no hay forma de almacenarla” son problemas que también afectan éste atributo; donde los problemas nuevamente surgen del modelo de datos.

Atributo afectado: RELEVANCIA

Éste atributo se define como la importancia de los datos almacenados para la representación de la realidad.

Es por ello que:

“a. No se cargan unidades de medida en que se midió la profundidad de un pozo petrolero” afecta éste atributo; y el problema surge del modelo de datos.

En el ejercicio 3, “a. Datos obligatorios que no se asumen como tales y por lo tanto no se cargan” es un problema que también afecta este atributo; donde el problema nuevamente surge del modelo de datos.

Atributo afectado: VIGENCIA

Éste atributo se define como la actualización de los datos con una frecuencia adecuada.

Es por ello que: “g. Las direcciones de los clientes no están actualizadas” afecta éste atributo; y el problema surge del modelo de datos.

En el ejercicio 3, “h. Datos que han cambiado en el mundo real, y que no fueron actualizados” es un problema que también afecta este atributo, y el problema nuevamente surge del modelo de datos.

Atributo afectado: CONFIABILIDAD

Éste atributo se define como la certeza de que los datos representan información verídica.

Es por ello que: “f. Hay personas fallecidas que figuran como empleados participantes de cursos (por los cuáles la empresa que los informa consigue una exención impositiva)” afecta éste atributo; y el problema surge de los datos.

Atributo afectado: CONSISTENCIA

Éste atributo se define como la ausencia de contradicciones entre distintos datos almacenados.

Es por ello que: “c. Hay inconsistencias entre nombres de un mismo pozo petrolero en distintos sistemas” afecta éste atributo; y el problema surge de los datos.

En el ejercicio 3, “d. Distintas personas cargan la misma información haciendo distintas asunciones” y “i. Datos que provienen de distintas fuentes y que no son consistentes” son problemas que también afectan este atributo; y los problemas surgen de los datos (aunque el caso d. podría deberse a un problema del modelo de datos si es que no se ha determinado bien la funcionalidad de la información).

Atributo afectado: CORRECCIÓN

Éste atributo se define la autenticidad de la representación de los datos para con la situación real.

Es por ello que: “d. La ubicación de una central telefónica no coincide con la ubicación real” y “e. El nombre de un pozo petrolero no corresponde con el que debería ser, de acuerdo a la ley” afectan éste atributo; y los problemas surgen de los datos.

En el ejercicio 3, “k. Posibles valores completados en el campo región: ANETOFAGASTA, ANMTOFAGASTA, ANT0FAGASTA, ANTO9FAGASTA, ANTOAFAGASTA, ANTOFAAGASTA” es un problema que también afecta este atributo; y el problema surge nuevamente de los datos.

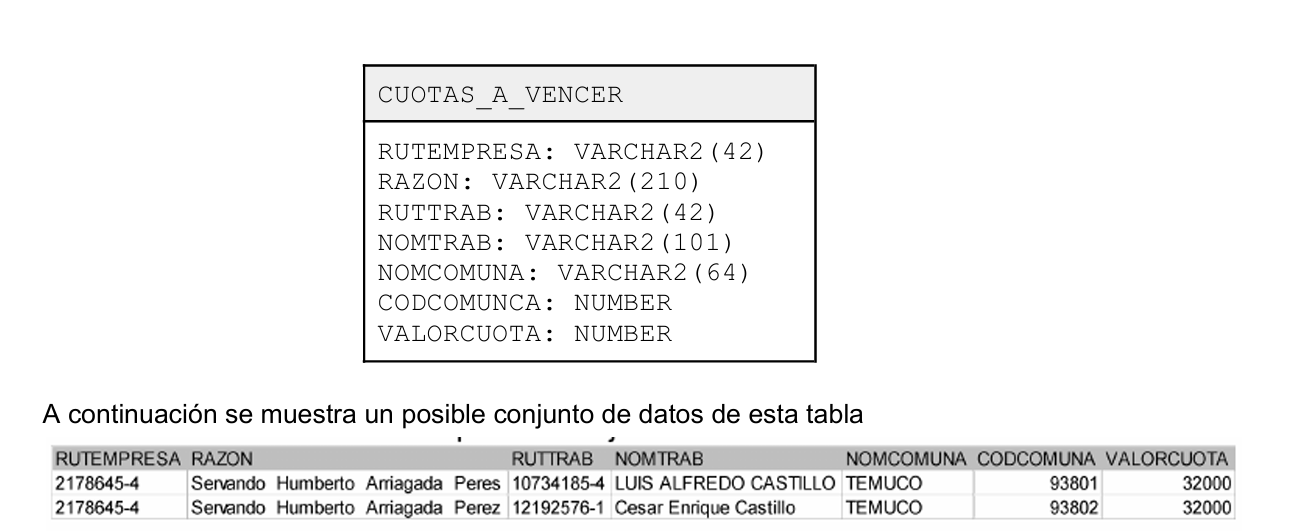
Atributo afectado: SEGURIDAD/PRIVACIDAD

Éste atributo se define como el cumplimiento de los requerimientos de privacidad adecuados de acuerdo a la reglamentación nacional-internacional/criterios éticos de los datos presentes. Esto implica, por ejemplo, la no accesibilidad a los datos de usuarios no autorizados.

En el ejercicio 3, “e. Gente que hace modificaciones pero no debería estar autorizada para hacerlas” es un problema que afecta éste atributo; y el problema surge del modelo de datos.

Ejercicio 4:

1. Los problemas de diseños son varios:
2. La semántica de los atributos del esquema no es clara. No se logra comprender la funcionalidad de la tabla; y hay nombres de ciertos atributos que podrían ser más claros.
3. Hay redundancia, pues ciertos valores se repiten/pueden repetir, y eso genera un coste mayor de almacenamiento del que podría ser necesario.
4. Dependiendo de cómo se realicen ciertos JOINs, hay alta probabilidad de generación de tuplas espúreas al contar con mucha información repetida, unión de atributos no fuertemente enlazados/relacionados, etc.
5. Para guardar información en el esquema evitando valores NULLs, será necesario que en cada paso se inserten todos los valores de cada atributo del esquema, aún si lo único que se quisiera guardar sea, por ejemplo, RUTEMPRESA
6. Ésta tabla puede generar anomalías de:



. **Inserción** pues se observa que hubo una inconsistencia en el valor del atributo RAZON para dos tuplas que comparten la misma RUTEMPRESA; lo mismo con el valor del atributo CODCOMUNA para dos tuplas que comparten el mismo NOMCOMUNA.

. **Modificación** pues en caso de que, por ejemplo, se deba corregir la RAZON asociada a la RUTEMPRESA 2178645-4 a “Servando Humberto Arraigo Pérez”, se deberá realizar la modificación en cada tupla que cumpla con el número de Rol Único Tributario afectado; este procedimiento, si no se hace con cuidado, puede generar anomalías.

. **Deleción** pues en caso de que se deba eliminar la primera tupla porque, por ejemplo, la persona ya pagó la cuota, estaría perdiendo a la vez los datos de la comuna; que, en el caso de su código, podría haber sido el correcto y, por ende, al quitar la tupla, se desvanece dicha pieza de información.

1. La tabla podría acarrear problemas de:

* Valores no estandarizados: en el atributo NOMTRAB el valor de la primera tupla está completamente en mayúsculas; mientras que en la segunda tupla, solo en las iniciales.
* Información crítica no confiable: como he observado anteriormente, no es posible indicar cuál es el código de la comuna TEMUCO.
* Valores desactualizados: al momento de querer actualizar el VALORCUOTA asociada a cierta RUTEMPRESA, podría suceder que no se observe y modifique cada una de las tuplas en las que se encuentre mencionada.
* Inconsistencias entre aplicaciones/en una misma aplicación: en la tabla se halla una inconsistencia en el valor del atributo RAZON para dos tuplas que comparten la misma RUTEMPRESA.
* Ocurrencias duplicadas: si no hay estandarización en los valores y se presentan inconsistencias, entonces esto puede provocar además la presencia de tuplas repetidas, o que su razón de ser es, en definitiva, expresar lo mismo/brindar la misma información.

1. Otros problemas del diseño de una BD que pueden afectar los atributos de calidad del modelo son la *falta de datos históricos*,que podrían ser de fundamental relevancia si se quiere, por ejemplo, analizar el pago de cuotas o los montos de cuota que los clientes de cierto servicio han pagado, para la posterior determinación de un modelo de ventas acorde a las posibilidades del mercado; o la aparición de *valores imposibles o poco probables* que, siguiendo el mismo contexto, puede traer fallas en las expectativas o en los modelos de predicción que se generen a partir de los datos.
2. En los casos en que se desee dejar información redundante en una tabla adrede, implementaría dependencias funcionales para lograr consistencia, vigencia y correctitud de los datos, atributos clave en lo que respecta a la calidad del modelo de una BD.

Ejercicio 5:

1. **Goal**: el dato correspondiente destinado a la Dirección de cada empresa no esté vacío.

**Question**: ¿Cuál es la proporción de empresas que tiene Dirección vacía?

**Metric**:

En este caso, los algoritmos de matching no son muy útiles, pues suponiendo que las empresas que aparecen son distintas entre sí, no habría necesidad de verificar direcciones pues lo que se observa es que, justamente, tienen valor nulo.

1. **Goal**: eliminar tuplas repetidas (contienen misma información en {NombreEmpresa, Dirección}.

**Question1**: ¿cuántos pares {NombreEmpresa, Dirección} distintos hay?

**Metric**:

En este caso, los algoritmos de matching sí son muy útiles pues permitirían validar si cada nombre y dirección registrados son distintos.