Laboratorio de Datos – Guía Calidad de Datos

**Ejercicio 1**

**a) Análisis de los casos**

**Caso 1: Pozo petrolero perforado en una ubicación errónea**

**i. Afectados:**

* La empresa petrolera (por multas y costos innecesarios).
* Ingenieros y geólogos que tomaron decisiones basadas en datos incorrectos.
* Accionistas y directivos (por pérdida de confianza y reputación).

**ii. Impacto:**

* Pérdida económica por multas y costos de perforación en una ubicación errónea.
* Impacto ambiental si se produjo daño en la zona perforada.
* Descreimiento en la calidad de los datos geoespaciales.
* Posible retraso en la producción de petróleo.

**Caso 2: Cliente del banco incluido erróneamente en bases de datos de morosos**

**i. Afectados:**

* Cliente perjudicado (por problemas financieros y reputacionales).
* Banco (por demanda y pago de indemnización).
* Otros clientes que pueden perder confianza en la institución.

**ii. Impacto:**

* Daño a la reputación del cliente, afectando su acceso a créditos o productos financieros.
* Pérdida de confianza en el banco.
* Costos legales para la entidad bancaria.
* Posible sanción de organismos reguladores.

**Caso 3: Correspondencia del gobierno con errores en direcciones**

**i. Afectados:**

* Empresas que no recibieron la comunicación.
* Organismo gubernamental (por ineficiencia en la gestión).
* Ciudadanos que dependen de medidas o beneficios relacionados con la norma.

**ii. Impacto:**

* Costos innecesarios por envío fallido de correspondencia.
* Ineficiencia administrativa.
* Posible descreimiento en la gestión pública.
* Impacto en la toma de decisiones de las empresas afectadas.

**b) Experiencia personal o laboral con problemas de calidad de datos**

En lo personal me ha pasado que me han cambiado el apellido en algunos boletines de notas del Colegio mientras que en lo laboral o académico no tengo recuerdo que me haya pasado algo con respecto a mis datos personales. En esta materia estamos viendo que hay datos duplicados, mala información, datos incorrectos, etc.

**Ejercicio 2**

**Ejemplo 1: Un CRM (Customer Relationship Management) en una empresa**

Un sistema CRM puede funcionar correctamente desde un punto de vista técnico (sin errores de código o fallas en la interfaz), pero aun así contener información con problemas de calidad, como:

* Datos de clientes duplicados porque distintos empleados los ingresan con ligeras variaciones en el nombre.
* Información desactualizada, como direcciones o números de teléfono que ya no están en uso.
* Campos obligatorios llenados con valores incorrectos o genéricos (“N/A”, “000000” en lugar de un número válido).

📌 **Impacto:** Mala segmentación de clientes, pérdida de oportunidades de venta y envío de información a contactos incorrectos.

**Ejemplo 2: Un sistema de gestión de historias clínicas electrónicas en un hospital**

El software puede ser robusto y sin fallas en su código, pero permitir que los médicos ingresen información de manera inconsistente o incompleta:

* Diferentes médicos pueden registrar diagnósticos en formatos distintos (Ejemplo: “Diabetes tipo II” vs. “DM2”).
* Datos faltantes en registros de pacientes (como alergias o antecedentes familiares).
* Información ingresada con errores tipográficos, lo que puede afectar tratamientos o investigaciones médicas.

📌 **Impacto:** Riesgos para la salud de los pacientes, dificultades para analizar datos clínicos a nivel poblacional y errores en la facturación de seguros.

**Ejercicio 3**

**Clasificación según el origen del problema**

| **Inconveniente** | **Origen** |
| --- | --- |
| **a.** Datos obligatorios que no se asumen como tales y por lo tanto no se cargan. | **Modelo** (El modelo de datos no define correctamente restricciones de obligatoriedad). |
| **b.** Interfaces poco amigables. | **Software** (La usabilidad del sistema dificulta la correcta carga de datos). |
| **c.** Rangos de valores que no se respetan. | **Instancia** (Se cargan datos fuera de los valores esperados en registros individuales). |
| **d.** Distintas personas cargan la misma información haciendo distintas asunciones. | **Proceso** (Falta de normas o guías claras para el ingreso de datos). |
| **e.** Gente que hace modificaciones pero no debería estar autorizada para hacerlas. | **Software** (Falta de control de accesos y permisos en el sistema). |
| **f.** Hay información que no está presente porque no hay forma de almacenarla. | **Modelo** (El diseño del modelo de datos no contempla ciertos atributos necesarios). |
| **g.** El mundo que se quiere representar evolucionó, pero esta situación no se ve reflejada en el sistema. | **Modelo** (El modelo no ha sido actualizado para reflejar cambios en la realidad). |
| **h.** Datos que han cambiado en el mundo real, y que no fueron actualizados. | **Instancia** (Datos desactualizados en registros individuales). |
| **i.** Datos que provienen de distintas fuentes y que no son consistentes. | **Proceso** (No hay integración ni normalización de datos de diferentes fuentes). |
| **j.** Datos correspondientes al año, que han sido almacenados con dos dígitos en lugar de cuatro. | **Modelo** (Definición incorrecta del formato de almacenamiento de fechas). |
| **k.** Posibles valores completados en el campo región con errores tipográficos. | **Instancia** (Errores de carga en registros individuales por falta de validación). |

📌 **Resumen de cada origen:**

* **Instancia:** Errores en datos específicos en registros individuales.
* **Proceso:** Problemas en la forma en que los datos son ingresados, manipulados o combinados.
* **Modelo:** Errores en el diseño de la estructura de datos o la lógica de representación del mundo real.
* **Software:** Problemas en la interfaz, controles de acceso o funcionalidades del sistema.

**Ejercicio 4**

**Clasificación de los problemas según atributo de calidad y tipo de problema**

| **Problema** | **Atributo de calidad afectado** | **¿Modelo o datos?** |
| --- | --- | --- |
| **a.** No se cargan unidades de medida en que se midió la profundidad de un pozo petrolero. | **Completitud** (Falta información clave para interpretar los datos). | **Datos** (La información no ha sido ingresada correctamente). |
| **b.** No es posible almacenar el sistema de referencia. | **Completitud** (El modelo no permite guardar una información relevante). | **Modelo** (El diseño de la base de datos no contempla este dato). |
| **c.** Hay inconsistencias entre nombres de un mismo pozo petrolero en distintos sistemas. | **Consistencia** (Distintas versiones del mismo dato en diferentes sistemas). | **Datos** (Error en la gestión y unificación de la información). |
| **d.** La ubicación de una central telefónica no coincide con la ubicación real. | **Exactitud** (Los datos no reflejan la realidad). | **Datos** (El valor ingresado no es correcto). |
| **e.** El nombre de un pozo petrolero no corresponde con el que debería ser, de acuerdo a la ley. | **Conformidad** (No cumple con las reglas externas que regulan los datos). | **Datos** (Carga incorrecta o no validada con la normativa). |
| **f.** Hay personas fallecidas que figuran como empleados participantes de cursos (por los cuales la empresa consigue una exención impositiva). | **Validez** (Los datos no representan información real o lógica). | **Datos** (Los registros no han sido actualizados correctamente). |
| **g.** Las direcciones de los clientes no están actualizadas. | **Actualización/Temporalidad** (La información quedó desfasada con el tiempo). | **Datos** (Falta de mantenimiento y actualización). |

**Ejercicio 3: Atributos de calidad afectados en problemas de instancia o modelo**

| **Problema del Ejercicio 3** | **Atributo de calidad afectado** |
| --- | --- |
| **a.** Datos obligatorios que no se asumen como tales y por lo tanto no se cargan. | **Completitud** |
| **c.** Rangos de valores que no se respetan. | **Validez** |
| **f.** Hay información que no está presente porque no hay forma de almacenarla. | **Completitud** |
| **g.** El mundo que se quiere representar evolucionó, pero esta situación no se ve reflejada en el sistema. | **Actualización/Temporalidad** |
| **h.** Datos que han cambiado en el mundo real, y que no fueron actualizados. | **Actualización/Temporalidad** |
| **j.** Datos correspondientes al año, que han sido almacenados con dos dígitos en lugar de cuatro. | **Precisión** |
| **k.** Errores tipográficos en nombres de regiones. | **Exactitud** |

📌 **Resumen de los atributos de calidad más afectados:**

* **Completitud** → Falta de datos esenciales.
* **Consistencia** → Mismos datos con valores distintos en distintos sistemas.
* **Exactitud** → Datos incorrectos respecto a la realidad.
* **Validez** → Información no lógica o que no cumple con reglas.
* **Conformidad** → Datos que no cumplen con normativas externas.
* **Actualización/Temporalidad** → Datos desactualizados o fuera de contexto.
* **Precisión** → Formatos incorrectos o poca granularidad en los datos.

**Ejercicio 5**