

Laboratorio de Datos

Calidad de Datos



Laboratorio de Datos

Calidad de Datos
... por Viviana Cotik (y modificaciones de P. Turjanski)

1º parte de
la materia

Recorrido de la materia (hasta ahora)

- ✓ Lenguaje de programación para trabajar en nuestros proyectos



- ✓ Etapas de un proyecto de Ciencias de Datos

- ✓ Modelado de Datos



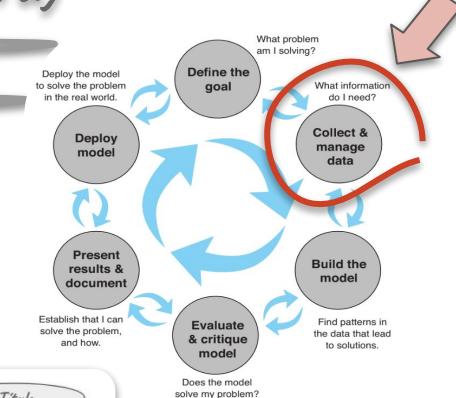
- ✓ Representación de los Datos

Materia	
Código	Nombre
1	Laboratorio de Datos
2	Análisis II
3	Álgebra Lineal

Unidad			
Código	Materia	Título	Descripción
1	Laboratorio de Datos	Administración de datos	Obtención y Manejo de los datos
1	Análisis II	Modelos Explicativos	Construcción de modelos explicativos
1	Álgebra Lineal	Modelos Predictivos	Construcción de modelos predictivos
2	Laboratorio de Datos	Integrales sobre curvas y volúmenes	Integrales en múltiples variables
2	Análisis II	Ecuaciones Diferenciales	Diseño y análisis de sistemas dinámicos

- ✓ Maneras de consultar los Datos AR/SQL

- ✓ Estrategias para mejorar la calidad de los datos desde el diseño de la BD



Introducción a Calidad de Datos

Trabajo individual



Actividad - Consigna



- Ingresar al siguiente *LINK* y completar la encuesta

<https://forms.gle/S1bK4m5NTaMo6QTF7>

Importante: No se pueden hacer preguntas sobre el formulario

- Objetivo. Conocer cuál será el transporte más utilizado (por los alumnos de nuestro curso) el día de hoy para retirarse de Ciudad Universitaria. En caso de que piense que va a utilizar más de un transporte responder con respecto al primero que vaya utilizar.

Trabajo Grupal



- *Solicitar al docente que comparta las respuestas de la encuesta*
- *En grupos de 3 integrantes responder las siguientes preguntas*



1. ¿Cuántos estudiantes respondieron la encuestas?
2. ¿Cuántos estudiantes eran en total?
3. ¿Algún estudiante respondió más de una vez (respuestas repetidas)?
4. A partir de los datos, ¿pueden responder cuál será el transporte más utilizado por los alumnos para retirarse de Ciudad Universitaria? ¿Cuál será?
5. ¿Algunos se retiran caminando? Y en caso afirmativo, ¿existe alguna relación entre estos que se retiran caminando y el número de calzado que utilizan?
6. Las respuestas ¿tenían datos faltantes?
7. ¿Observaron respuestas con cierta inconsistencia?
8. ¿Se encontraron que todas las respuestas tenían el formato correcto?
Mencione al menos 3 casos en se encontraron con respuestas en un formato inesperado
9. ¿En cuántas respuestas obtuvieron “Ciudad Autónoma de Buenos Aires” como respuesta a Provincia?
(www.argentina.gob.ar/pais/provincias)

FIGURAR EN UN LISTADO DE INCUMPLIDORES ARRUINO UN NEGOCIO

Un banco debe pagar \$ 120.000 por incluir mal a un cliente en Veraz

Daniel Gutman

El Banco Río lo incluyó en las listas negras de deudores de la Organización Veraz y solicitó al Banco Central que lo inhabilitara. Pero todo era un error, porque no había existido ningún incumplimiento. El cliente hizo juicio y obtuvo una sentencia de Cámara a su favor. Hasta ahí, un caso igual a muchos otros que ha habido en los últimos años. Lo novedoso es que la Cámara en lo Comercial acaba de establecer la que seguramente sea la indemnización más alta en este tipo de casos: el Banco Río deberá pagarle 120.000 pesos a su ex cliente.

A esa cifra deberán sumársele los intereses a la tasa activa del Banco Nación desde la fecha de inhabilitación en los registros del Central, que es mayo de 1996, lo que llevaría la indemnización a más de medio millón de pesos, según los abogados del demandante.

La importancia de la indemnización —según se explicó en el fallo— tiene que ver con que el damnificado es un empresario que estaba en pleno proceso de ampliación de sus negocios.

El hombre, dueño de una confitería, estaba construyendo un edificio en la avenida Cruz en el cual pensaba instalar una concesionaria de autos, además de una confitería y salón de fiestas en la planta alta. Sin embargo, en mayo de 1996 quedó sin posibilidad de obtener crédito y operar con cheques, por lo que la obra y sus proyectos quedaron inconclusos.

Así, la Sala B de la Cámara —en un voto de la jueza María de Díaz Cordero, al que adhirió Enrique Butty— aplicó el concepto de "pérdida de chance". Es decir, el Río deberá indemnizar al empresario porque lo privó de una oportunidad de ganar dinero.

Las pruebas presentadas y la trayectoria de Eloy Domínguez Alvarez convencieron a la jueza de que él "tenía intención de culminar con la construcción del edificio y ampliar sus negocios" y de que lo hubiera hecho "de no haber existido la arbitraria y errónea decisión adoptada por la entidad bancaria demandada".

Caso

1. Leer el artículo
2. En grupos de 3 integrantes ...
 - a. Describir el problema
 - b. ¿Cuál es la causa del problema?
 - c. ¿Quiénes se benefician al contar con datos de calidad?



Calidad de Datos

¿Cuál es la definición de Calidad de Datos? (discutir 5 min.)



Definiciones

- “Un dato o conjunto de datos X tiene mayor calidad que un dato o conjunto de datos Y , si X satisface las necesidades del usuario mejor que Y ” [Redman, 1996]
- “Satisfacer de manera consistente las expectativas de los usuarios” [English, 1999]

... son definiciones subjetivas

Calidad de Datos

*¿Cuáles son las consecuencias de contar con datos de mala calidad?
(discutir 5 min.)*



Calidad de Datos

¿Cuáles son las consecuencias de contar con datos de mala calidad?

- Desconfianza
- Insatisfacción de los clientes
- Costos innecesarios
- Impacto en la toma de decisiones
- ...



Gráfico tomado de AIT solutions

Problemas de Calidad de Datos

Trabajo en equipo



Ejercicio - Consigna

- ✓ Conformar grupos de 3 integrantes
- ✓ Descargar el registro de Datos de Dengue de 2020 correspondiente al Registro del Sistema Nacional de Vigilancia de la Salud 2.0 (<http://datos.salud.gob.ar/dataset/vigilancia-de-dengue-y-zika>)
-> Vigilancia de Dengue y Zika - 2020 (.xls)
- ✓ Responder las siguientes preguntas:
 1. ¿Detectan problemas de calidad de datos?
En caso afirmativo, mencionar cuáles son y caracterizarlos
 2. ¿Piensan que los datos provistos provienen de una única tabla de una BD relacional?

Ejercicio - Debate

Algunos problemas de Calidad de datos detectados ...

- ✓ Datos nulos: departamento_nombre, provincia_nombre, provincia_id
- ✓ Columnas intercambiadas:
 - grupo_edad_id vs grupo_edad_desc
 - evento_nombre vs semanas_epidemiologicas
- ✓ Códigos incorrectos (departamento_id)
- ✓ Sólo casos de dengue (no de zika)
- ✓ Rangos erróneos en grupo etario
- ✓ Algunos departamentos escritos con comillas (ej. Apostoles vs Apóstoles, etc.)

Trabajo Grupal



- Conformar grupo de 3 estudiantes
- Discutir qué acciones tomarian (en este caso) para mejorar la calidad de los datos

Ejercicio - Solución posible

Possible solución ... armar tablas normalizadas:

1. Chequear y corregir datos (aumentar la calidad de los datos)
2. Crear nuevas tablas con id y descripciones.
En la tabla original sólo deberían figurar los ids (y no las descripciones)
3. Tomar alguna decisión (y documentar) sobre qué hacer con los registros que aparecen dos veces
(pero con distinta cantidad de casos reportados)

Algunos problemas habituales

- ✓ Valores no estandarizados
 - NETOFAGASTA
 - ANMTOFAGASTA
 - ANT0FAGASTA
 - ANTO9FAGASTA
 - ANTOAFAGASTA
 - ANTOFAAGASTA
- ✓ Valores imposibles o poco probables
 - Edad: 200 años
- ✓ Valores faltantes
 - Registros de personas con el campo e-mail vacío
- ✓ Valores desactualizados
 - Ocurrencias duplicadas
 - Falta de datos históricos
 - Inconsistencia entre aplicaciones o en una misma aplicación
 - Datos de pacientes en dos servicios distintos de un hospital
 - Datos de pozos petroleros en dos aplicaciones distintas (perforación, producción)
 - Información crítica que no es confiable
 - Hay personas habilitadas a votar que han fallecido

Causas de problemas en Calidad de Datos

Trabajo Grupal



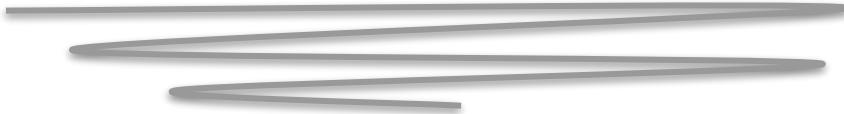
- Conformar grupo de 3 estudiantes
- Discutir cuáles pueden ser las causas por las que los datos tienen problemas de calidad

Posibles causas de problemas

Posibles causas de problemas ...

- ✓ Problemas masivos que reparan un dato, pero no reconstruyen información relacionada
Ej. Recuperar registros de pago (que por un problema de software fueron perdidos) y posteriormente no eliminar de la tabla de deudores a aquellos clientes que saldaron su deuda.
- ✓ Misma información cargada en distintos sistemas
Ej. Datos de Estudiante en SIV-Guarani y datos de estudiante en el Campus (puede diferir el email)
- ✓ Valores predeterminados
Ej. Fecha de Nacimiento 01-01-2021

Posibles causas de problemas



Las causas pueden depender de ...

- *Calidad de software (usabilidad, interfaz [obligatoriedad de carga], seguridad)*
- *Definición de procesos asociados a los datos*
- *Diseño de la BD*
- *Falta de capacitación*

Posibles causas de problemas

Los problemas de calidad los podemos clasificar en ...

Posibles causas de problemas

1. Problemas asociados a la *INSTANCIA*

- ✓ Datos que han cambiado en el mundo real, y que no fueron actualizados
- ✓ Datos que provienen de distintas fuentes, deberían ser consistentes y sin embargo no lo son
- ✓ Datos que no han sido almacenados con la precisión necesaria (por ejemplo, Y2K)

Posibles causas de problemas

2. Problemas asociados al **MODELO DE DATOS**

- ✓ Si se detecta que hay información que no está presente porque no hay forma de almacenarla
-> el modelo de datos físico está incompleto
- ✓ El mundo que se quiere representar evolucionó y no se tradujeron los cambios al modelo
-> pérdida de vigencia del modelo

Posibles causas de problemas

3. Problemas asociados a los **PROCESOS**

- ✓ Distintas personas cargan la misma información haciendo distintas asunciones
- ✓ Se carga con una asunción y se usa con otras
- ✓ Modificaciones manuales por procesos
- ✓ Gente que hace modificaciones pero no debería estar autorizada para hacerlas

Posibles causas de problemas

4. Problemas asociados a **ERRORES DE SOFTWARE**

- ✓ Datos obligatorios que no se asumen como tales y por lo tanto no se cargan
- ✓ Interfaces poco amigables

Posibles causas de problemas

Importante: Sólo el **software** de buena calidad no garantiza la calidad de los datos

Se debe trabajar sobre:

- ✓ La instancia
- ✓ El modelo de datos
- ✓ Los procesos que intervienen en la generación y modificación del dato
- ✓ La consistencia entre las diferentes fuentes de datos

Atributos (o dimensión) de Calidad de Datos

Atributos de Calidad



¿Qué características deberían cumplir los datos para ser de calidad? (5 min.)



Atributos de Calidad

¿Qué características deberían cumplir los datos para ser de calidad?

Deberían ser ...

- ✓ completos
- ✓ oportunos (timeliness) y vigentes
- ✓ consistentes y correctos
- ✓ en cantidad adecuada
- ✓ disponibles/accesibles (ej. medicina), open data.
- ✓ seguros y privados (protección de datos personales)

Atributos de Calidad

Podemos definir los siguiente *Atributos de Calidad* ...

Atributos de Calidad

1. Completitud

- ✓ Están presentes todos los valores para representar la realidad
- ✓ Están presentes todas las instancias existentes en el mundo real

1. Relevancia

- ✓ Los datos son relevantes para representar la realidad

1. Vigencia

- ✓ Los datos se mantienen actualizados con la frecuencia adecuada

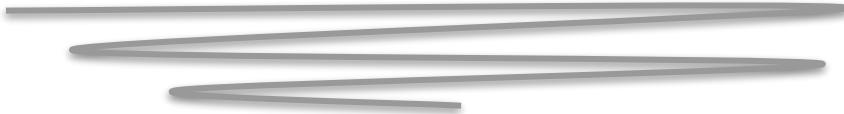
1. Disponibilidad

- ✓ Los datos están accesibles

1. Confiabilidad

- ✓ Se puede considerar que los datos representan información verídica

Atributos de Calidad



6. Consistencia

- ✓ No hay contradicciones entre distintos datos almacenados

6. Corrección

- ✓ Los datos representan la situación real

6. Seguridad/Privacidad

- ✓ Los datos cumplen con los requerimientos de privacidad adecuados de acuerdo a la reglamentación nacional-internacional / criterios éticos
- ✓ Los datos son sólo accesibles por los usuarios autorizados

Generalmente no nos vamos a encontrar con datos perfectos
Es necesario priorizar los atributos de calidad deseados

Cuán buenos son los datos (Calidad)

¿Son buenos los datos?



¿Cómo determinar cuán buena es la calidad de los datos? (5 min.)



¿Son buenos los datos?



¿Cómo determinar cuán buena es la calidad de los datos?

Tenemos que "entender" cuáles son los datos críticos y para ellos determinar los atributos de calidad de interés

Necesitamos seguir una metodología

1. Hacer relevamiento
2. Elaborar métricas de calidad
3. Recolectar valores de dichas métricas

Nos va a permitir cuantificar la calidad de los datos

1. Hacer relevamiento

Objetivo:

Determinar ...

- ✓ Datos críticos
- ✓ Ciclo de vida del dato
- ✓ Atributos de interés

Tareas a realizar:

- ✓ Identificar stakeholders (partes interesadas): CCC (creator, consumer, custodian)
- ✓ Conseguir el compromiso por parte del cliente - la organización
- ✓ Leer documentación sobre los sistemas, sobre el negocio, y estudiar modelos de datos
- ✓ Hacer cuestionarios tendientes a determinar cuáles son los datos críticos, cuál es el ciclo de vida del dato, cuáles son los atributos de interés y los problemas habituales
- ✓ Llevar adelante los cuestionarios con cada uno de los stakeholders identificados

2. Elaborar métricas de calidad

Para los datos críticos y los atributos de interés, armar métricas para cuantificar cuán grave es el problema

Possible Técnica: **Goal Question Metric** (Objetivo, Pregunta, Métrica). Conocida como **GQM**

Goal (Objetivo) Se define un objetivo

Question (Pregunta) Se plantea una pregunta (o más), cuya respuesta permitirá saber si se satisface el objetivo

Metric (Métrica) Se plantea una métricas (o más) -para cada una de las preguntas-, cuya ejecución permitirá responder las mismas

2. Elaborar métricas de calidad

Ejemplo (GQM. Goal Question Metric)

Queremos analizar la Completitud del dato Departamento asociado a los Empleados de la Compañía

Goal (Objetivo) El dato correspondiente al Departamento donde trabaja cada Empleado esté completo

Question (Pregunta) ¿Cuál es la proporción de Empleados que tienen el dato correspondiente a Depto. vacío?

Metric (Métrica) M1: Proporción de registros con campo Departamento vacío en tabla Empleados, es decir,
Cantidad de registros de Empleado con campo id_Departamento vacío

Cantidad total de registros de Empleado

M2: Proporción de id de Departamento que tienen su nombre de departamento vacío
(en Tabla Departamento)

3. Recolectar valores de las métricas

Los valores de las métricas se pueden obtener a través de consultas SQL

*Los objetivos en GQM también podrían establecerse sobre el modelo de datos
Las respuestas podrían ser si o no y como resultado podría concluirse la necesidad
de una revisión del modelo de datos*

Trabajo en equipo



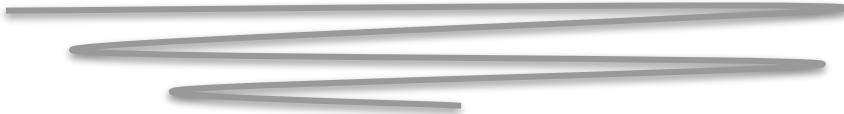
Ejercicio - Consigna



- ✓ Conformar grupos de 3 integrante
- ✓ Descargar el registro de Datos de Dengue de 2020 correspondiente al Registro del Sistema Nacional de Vigilancia de la Salud 2.0 (<http://datos.salud.gob.ar/dataset/vigilancia-de-dengue-y-zika>)
-> Vigilancia de Dengue y Zika - 2020 (.xls)
- ✓ Aplicar la técnica GQM para comenzar a evaluar la calidad de datos de dicha fuente

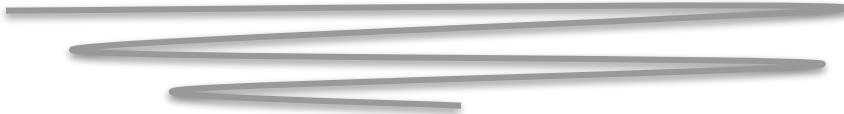
Diagnóstico y Mejoras

Diagnóstico



A partir de los resultados asociados a las ejecuciones de las métricas y del relevamiento, podemos determinar problemas existentes y sus causas.

Mejoras



A partir del diagnóstico: Conclusiones y propuestas de mejora.

No sólo se trata de corregir, sino principalmente de prevenir. Posibles correcciones en:

- ✓ *Instancia*
- ✓ *Modelo de datos*
- ✓ *Procesos*
- ✓ *Capacitación*
- ✓ *Software*

Herramientas de detección de problemas de Calidad de Datos



Existen muchas herramientas para automatizar la detección de:

- ✓ Textos parecidos (soundex, keyboard distance, edit distance, ..., uso de diccionarios).
- ✓ Datos nulos
- ✓ Problemas de integridad referencial

Tareas para la próxima clase

- 1. Resolver la guía de ejercicios de “Calidad de Datos”*

Bibliografia

- ✓ English, 'Improving Data Warehouse and Business Information Quality', John Wiley & Sons (1999)
- ✓ Piattini, Calero, Genero (eds.): 'Information and Database Quality', Kluwer (2001)
Cap. 7: Bobrowski, Marré, Yankelevich, 'A NEAT Approach for Data Quality Assessment'
- ✓ Redman, 'Data Quality for the Information Age', Artech House (1996)
- ✓ Wang, Strong, Guarascio, 'Beyond Accuracy: What data quality means to data consumers', Total Data Quality Management Program (1996)