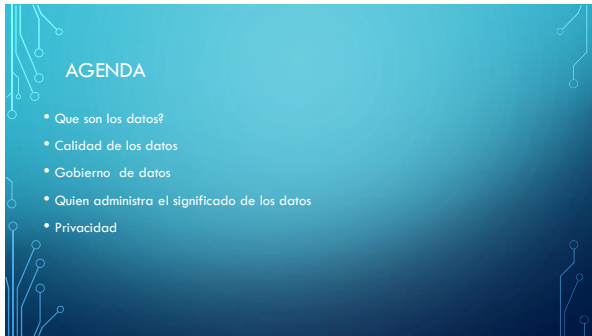
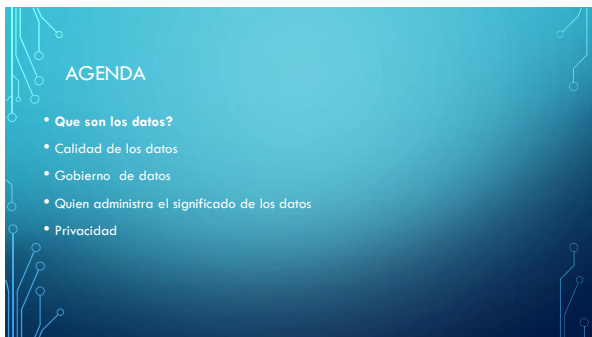




1



2



3

TIPOS DE DATOS

- Registro
 - Matriz de datos
 - Documentos
 - Datos de transacciones
- "Semi estructurada"
 - XML, JASON
- Grafos
 - Redes sociales
- Ordenados
 - Datos secuenciales
 - Datos Espacio - Temporales
 - Stream Data

4

MATRIZ DE DATOS

- Los datos consisten en un conjunto de registros, cada uno de los cuales contiene un conjunto fijo de atributos

Registro	Nombre	Fecha Nac.	Fecha Abandono
1	Juan	Castro	01-01-70
2	Maria	Castro	03-08-88
3	Pedro	Saltero	15-07-98
4	José Luis	Separado	23-04-75
5	Isma	Separado	09-05-93

5

DOCUMENTOS

- Cada documento se representa como un vector de términos
 - Cada término es un elemento del vector.
 - El valor de cada componente es la cantidad de veces que el término aparece en el documento
- El término documento es muy amplio, pueden ser comentarios en una red social, opiniones de productos, email, etc.

	Opinión	Comentario	Denuncia	Producto	Calidad	Exclamante	Defectuoso
Documento 1	3		1	2			
Documento 2			5		3		6
Documento 3		1		4	1	2	

6

DATOS DE TRANSACCIONES

- Un tipo especial de registro
 - Cada registro (transacción) involucra un conjunto de ítems
 - Por ejemplo los productos adquiridos en una compra

TID	Items
1	Pan, coca,
2	Cerveza, pan
3	Cerveza, coca, pañales, leche
4	Cerveza, pan, pañales,
5	Coca, pañales, cerveza

7

SEMI ESTRUCTURADOS, XML, JSON

- ▶ XML es un lenguaje de marcación desarrollado por la WWW.
- ▶ Es de tipo jerárquico y se utiliza mucho para el intercambio de información.
- ▶ Existe una manera de «validar» el contenido mediante el uso de .xsd
- ▶ JSON es similar
- ▶ Esto se puede usar, por ejemplo para "enriquecer" la información de un cliente llamando a alguna API que al recibir una dirección nos devuelva las coordenadas geográficas

8

REDES SOCIALES



Patryk de Intermedia del área de el laboratorio de Investigación de Harvard Harvard experimento con la estructura de la organización. Image from <http://www.patrykdeintermedia.com/> (source: <http://patrykdeintermedia.com/>)

9

DATOS ORDENADOS

- Secuencia de transacciones

Items / eventos

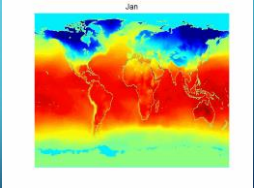
(A B)	(D)	(C E)
(B D)	(C)	(E)
(C D)	(B)	(A E)

10

DATOS ORDENADOS

- Datos espacio temporales

Temperatura media de la tierra y el oceano



11

DATOS ORDENADOS

Stream Data

- Los datos de tipo stream fluyen por un sistema de computadora en forma continua y con distintas velocidades. Son datos capturados por sensores.
- Están ordenados temporalmente, cambian rápidamente, son masivos y potencialmente infinitos
- IoT (Internet of things) : la estimación de Gartner para 2021 es de 25.000 millones de dispositivos conectados.
- Actualmente se habla de loE (Internet of Everything)

12

AGENDA

- Que son los datos?
- Calidad de los datos
- Gobierno de datos
- Quien administra el significado de los datos
- Privacidad

13

NUESTROS PRECONCEPTOS

- Si los datos están guardados en una tabla están bien
- Tenemos todos los campos de la tabla "con datos"
- Los datos "valen" para siempre.

14

ALGUNAS SITUACIONES QUE ME PASARON

- Tienda de Electrodomésticos 1, al darme de alta como cliente, indicaron en el campo "email" "No posee"
- En un banco privado las cartas que mandaba el sector tarjetas (ya nadie manda cartas) llegaban a la casa de mis padres, las que me mandaba el sector comercial a mi casa
- Tienda de Electrodomésticos 2, al darme de alta como cliente, indicaron en el campo "email" nosotros@domestico.com (o algo similar)
- En una tabla con categorías el 80% de los casos tenía categoría "otros"
- Todos los montos están informados en 0
- Hace 25 años cuando abrí la cuenta en el banco era soltera, pero ahora hace 20 años que estoy casada. Cuanto tiempo valen los datos que tenemos cargados?
- Un ente del estado que tiene 7 pisos y tenía 7 copias de su "tabla" principal, que por supuesto no coincidían ni en atributos ni en cantidad de registros

15

CALIDAD DE DATOS

- Si no se invierte dinero y esfuerzo la calidad de datos es mala
- Es necesaria monitorearla permanentemente
- Esta aceptado que el 70% del trabajo de un proyecto de minería de datos se invierte en "acomodar" y "cruzar" los datos.

16

EJEMPLOS DE ERRORES CLÁSICOS

- Fuera de Rango: Edad del Paciente= 185 ()
- No-Standard: Data Main Str, Main Street, Main ST, Main St.
- Datos inválidos: El dato puede ser "A" o "B" pero el valor es "C"
- Reglas culturales diferentes:
 - Fecha= Enero1, 2002 o 1-1-2002 o 1 Ene 02
 - Montos en diferentes monedas
- Distintos Formatos: [919]674-2153 o [919]6742153 o 9196742153
- Cosméticos: Jon | Jones transformado en Jon J Jones
- Informar el cuit como monto de la operación
- Completar campos requeridos con espacios en blanco , o puntos o parecidos
-

17

EJEMPLOS MAS SOFISTICADOS

- Tengo 3 bases de clientes y cada una tiene una dirección diferente , cual es la válida?
- El 30% de los clientes cumple años el mismo día
- La dirección no se corresponde con la localidad. O la localidad con la provincia o todo el paquete con el código postal. O la altura no existe en esa calle...
- Tiene 12 años pero esta casado
- Emitió una factura por \$1.000 millones
- Algunas cosas que "son sospechosas"
 - Tiene 7años y nivel de estudio de doctorado
 - Gasta \$ 2.000.000 por mes de tarjeta

18

QUE ES UN OUTLIER?

- Que pasa con el monto de \$ 2.000.000 de gasto de la tarjeta de crédito?
- Estos datos se conocen como "outliers", pueden ser
 - Errores
 - Valores atípicos (valores reales , pero poco frecuentes)



19

OUTLIERS PELIGROSOS



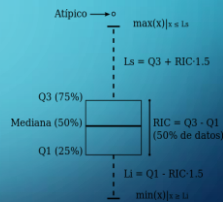
En 1985 tres investigadores (Farman, Gardiner y Shanklin) fueron desconcertados por un cierto dato recopilado por el "examen antártico británico" que demostraba que los niveles del ozono para la Antártida habían caído el 10% debajo de los niveles normales de enero. El problema era, porque el satélite Nimbo 7, que tenía instrumentos a bordo para medir con precisión los niveles del ozono, no había registrado concentraciones de ozono semejantemente bajas. Cuando examinaron los datos del satélite no les tomó mucho darse cuenta de que el satélite de hecho registraba estas niveles de concentraciones bajas y lo había estado haciendo por años. Pero como las concentraciones de ozono registradas por el satélite fueron tan bajas eran tratadas como outliers por un programa de computadora y desechadas! El satélite Nimbo 7 de hecho había estado recolectando la evidencia de los niveles bajos de ozono desde 1976. El dato a la capa de ozono pasó desapercibido y no fue tratado por nueve años porque los outliers fueron desechados sin ser examinados.

Moraleja: No tirar los outliers sin examinarlos, porque pueden ser los datos más valiosos de un dataset.

20

COMO PUEDO ANALIZAR LA CALIDAD?

- Lo primero es un Análisis univariado
 - Cuanto es el valor mínimo, y el máximo!
 - Media, Mediana, Moda, Cuartiles
 - Histogramas
 - Tablas de frecuencia
 - Gráficos
- Después el análisis bivariado
 - Coeficiente de correlación
 - Tablas de contingencia
 - Diagramas de dispersión de puntos
 - Etc.
- Puedo seguir con el perfilado de los datos
 - Que tipo de información "leo" de este sitio, esta nueva tanda que estoy leyendo es consistente con los datos previamente leídos?



21

AGENDA

- Que son los datos?
- Calidad de los datos
- **Gobierno de datos**
- Quien administra el significado de los datos
- Privacidad

22

GOBIERNO DE DATOS

- De acuerdo a la Data Management Association (DAMA, <http://www.dama.org>), la data resource management (administración de datos) es el "Desarrollo y ejecución de arquitecturas, practicas y procedimientos que manejan adecuadamente las necesidades del ciclo de vida de los datos de una empresa"
- Incluye aspectos de calidad, arquitectura, seguridad y meta data de los datos.
- No es un tema de **SISTEMAS** es un tema de **TODA** la organización
- Lo datos se consideran cada vez mas un **ACTIVO** de la compañía

23

NIVEL DE MADUREZ DEL GOBIERNO DE DATOS



Fuente, presentación "Ciencia, gobierno y monetización de datos" María del Rosario Brusera

24

IMPLEMENTACIÓN

- No se puede empezar con un mega proyecto, conviene elegir un objetivo no muy ambicioso, pero que sirva para mostrar la utilidad
- Como se apreciaba en el grafico de niveles de madurez el gobierno de datos es un camino sin fin...
- Se necesita "sponsoreo" del mas alto nivel.

25

AGENDA

- Que son los datos?
- Calidad de los datos
- Gobierno de datos
- **Quien administra el significado de los datos**
- Privacidad

26

ALGUNAS CUESTIONES PRACTICAS

- A quien le pregunto cuando tengo que agregar un campo a una tabla?
- Cual es la dirección actualizada de los empleados?
- Como calculo el saldo de un cliente?
- En que moneda están expresados los precios?
- Quién es el «dueño» de la tabla de cliente?
- Cuantas tablas de país tengo? Existe una equivalencia entre las mismas?
- Dado una venta como se cuales son las entregas asociadas?

27

UNA TABLA CUALQUIERA....

- Esta imagen muestra una porción de una tabla contenida en una base de datos de una compañía. La tabla contiene los datos de los clientes de la compañía. La primera columna es el código del cliente (si el número es negativo se refiere a un cliente "ficticio"), las columnas 2 y 3 identifican el periodo de validez del registro. La columna ID_GROUP indica a que grupo pertenece el cliente (si el valor de FLAG_CP es "S", entonces el cliente es el líder del grupo y si FLAG_CF es "S", entonces el cliente es el controlador del grupo), FATURATO es la ganancia anual (pero el valor es válido solo si FLAG_FATT es "S") (fuente: <http://wp.sigmod.org/?p=871>)

IDC	TL1_START	TL1_END	TL2_START	TL2_END	FLAG_CP	FLAG_CF	FATURATO	FLAG_FATT
-114524	26-Aug-2003	1-Sep-2003	-127786	-	S	N	175000.00	N
-148764	15-Aug-2003	17-Jun-2004	20503	-	N	N	230000.00	N
-118249	15-Aug-2003	26-Aug-2004	-127786	-	N	S	180000.00	S
-151293	15-Aug-2003	27-Aug-2004	-127779	-	S	N	810000.00	N
-139909	17-Aug-2003	1-Sep-2003	-128119	-	N	S	187000.00	S
-171204	17-Aug-2003	1-Sep-2003	-128119	-	S	N	0.00	N
1102272	1-Aug-2004	9-Aug-2005	170441	-	-	-	-	-

Figure 1: A portion of the Customer table in a database of a large organization.

28

QUE ES UN ADMINISTRADOR DE DATOS?

- Es una persona o un conjunto de personas responsables de la administración de datos. Es un perfil netamente funcional.
- **NO ES UN DBA.**
 - El dba es un especialista en un motor de base de datos , mientras que un administrador de datos es un especialista en los "datos" de una organización.

29

TAREAS PRINCIPALES ADMINISTRADOR DE DATOS⁽¹⁾

"Diseño lógico"

- Recolectar y analizar los requerimientos
- Modelar el negocio basado en los requerimientos (tanto conceptual como lógico)
- Definir standars (referidos a la forma de nombrar los objetos, abreviaciones, etc.) y asegurar su cumplimiento
- Conducir sesiones de *definición de datos* con los usuarios
- Manejar y administrar los *repositorios de metadata* y las herramientas de modelado
- Asistir al administrador de base de datos en la creación de los modelos físicos a partir de los modelos lógicos "

(1) IBM: Data Administration VS. Database Administration, <http://www.ibm.com/csicw/articles/4192>

30

DEFINICIÓN DE DATOS⁽²⁾

- En las organizaciones hay dos lugares donde típicamente se encuentran las definiciones de los datos desde el punto de vista del negocio
 - La cabeza de las personas. Estas son reglas no escritas y existen en todas las áreas de las empresas que interactúan con datos. Si las definiciones se encuentran solo en este lugar las empresas son vulnerables a la baja calidad de los datos, originada en falta de consistencia y de confianza
 - En los modelos de datos. Las herramientas de modelado de datos hacen un trabajo aceptable en recolectar este tipo de información. El problema es que suelen reflejar solo el estado inicial y no los cambios.

(2) Selecting the "Right" Data Data to Manage, <http://www.idata.com/news-articles/5069/>

31

AGENDA

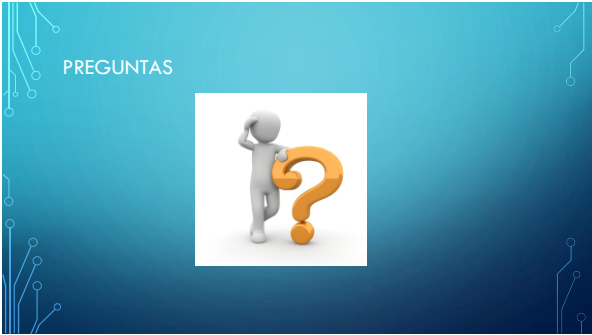
- Que son los datos?
- Calidad de los datos
- Gobierno de datos
- Quien administra el significado de los datos
- **Privacidad**

32

LA PRIVACIDAD ES UNA PREOCUPACIÓN CRECIENTE

- <https://youtu.be/i-ifnYR811w>
- Existen numerosas regulaciones internacionales al respecto
- Las organizaciones deben cumplir las normas locales y , ahora, la nueva ley de la Unión Europea referida a protección de datos garantiza la protección de los mismos para todos los ciudadanos, independientemente de donde estén
- En la Argentina existen numerosos "secretos", estadístico, fiscal, educativo

33



34
