QUE SON LOS DATOS? COMO LOS ADMINISTRAMOS? GOBIERNO DE DATOS

MG. CECILIA ANA RUZ

AGENDA

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

AGENDA

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

TIPOS DE DATOS

- **Registro**
  - Matriz de datos
  - Documentos
  - Datos de transacciones
- **“Semi estructurado”**
  - XML, Jason
- **Grafos**
  - Redes sociales
- **Ordenados**
  - Datos secuenciales
  - Datos Espacio - Temporales
  - Stream Data

MATRIZ DE DATOS

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

Numero	Nombre	Estado Civil	Fecha Nacimiento
1	Juan	Casado	01-01-70
2	Maria	Casado	03-08-88
3	Pedro	Soltero	15-07-98
4	Jose Luis	Separado	23-04-75
5	Silvia	Separado	06-05-82

DOCUMENTOS

- Cada documento se representa como un vector de términos
  - Cada término es un atributo 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, emails, etc.

	Queja	Contento	Demora	Producto	Calidad	Excelente	Defectuoso
Documento 1	3		1	2			
Documento 2			5		3		6
Documento 3		1		4	1	2	

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

SEMI ESTRUCTURADOS, XML, JSON

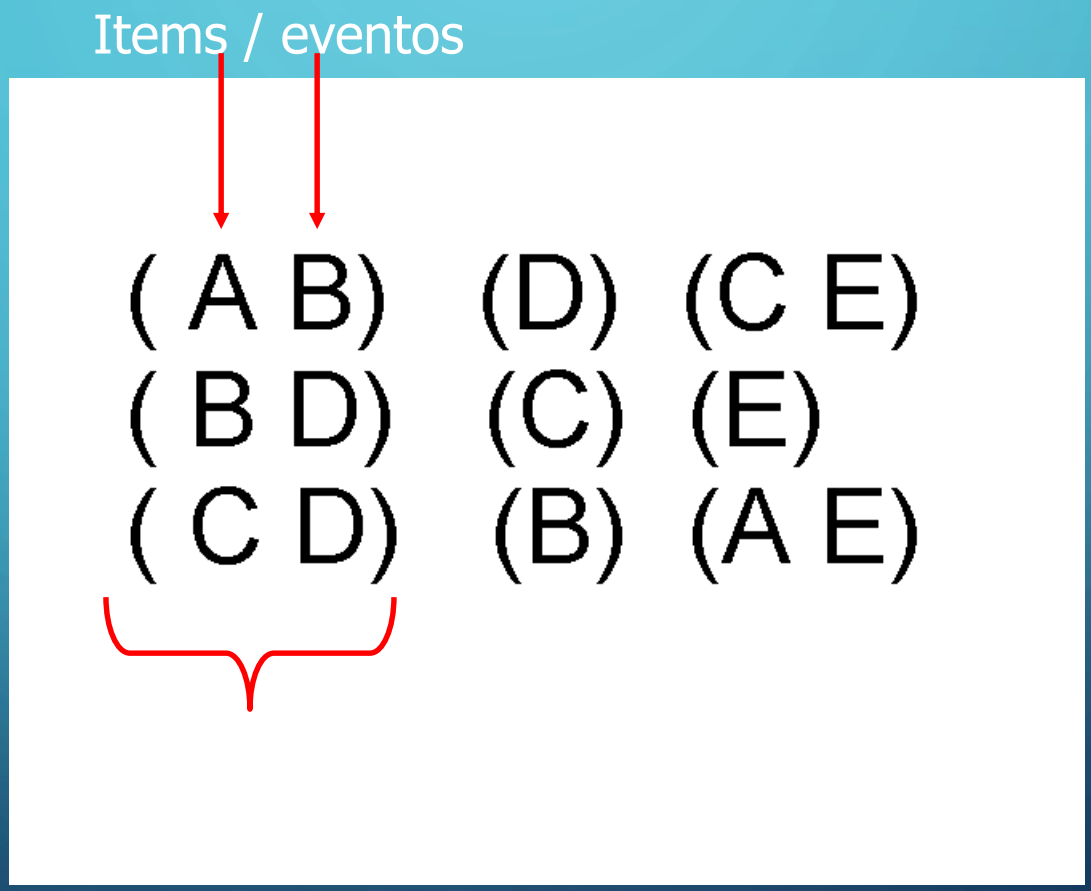
- XML: es un lenguaje de marcación desarrollado por la WWW.
- Es de tipo jerárquico y se utiliza mucho para el intercambio de información.
- Existe una manera de «validar» el contenido mediante el uso de .xsd
- JSON es similar
- Esto se puede usar, por ejemplo para “enriquecer” la información de un cliente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

REDES SOCIALES

Patrón de intercambio del email en el laboratorio de investigación de Hewlett Packard superpuesto con la estructura de la organización . (Image from http://wwwpersonal.umich.edu/ ladamic/img/hplabsemailhierarchy.jpg)

DATOS ORDENADOS

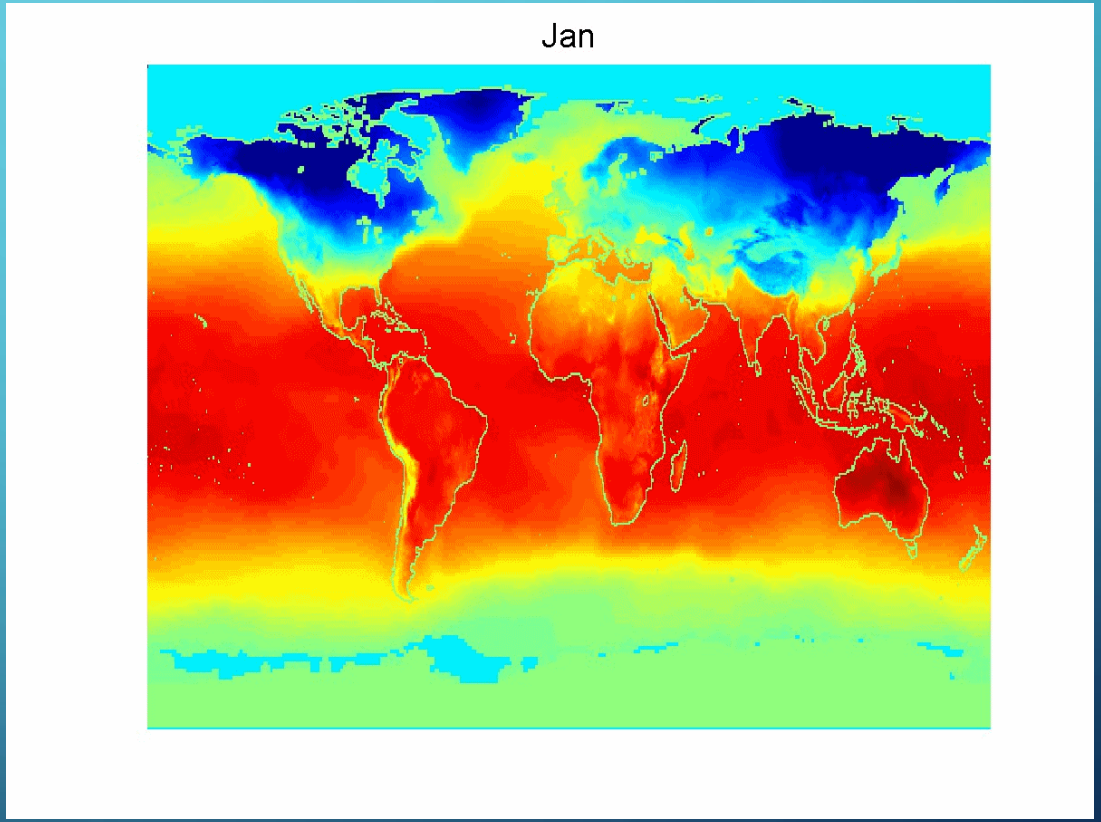
- Secuencia de transacciones



DATOS ORDENADOS

- Datos espacio temporales

Temperatura media de la tierra y el oceano



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) : para el año 2020 se calcula que va a haber 75 billones de dispositivos conectados.
- Actualmente se habla de IoE (Internet of Everything)

AGENDA

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

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.

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” nada@hotmail.com...(o algo similar)
- En una tabla con categorías el 80% de los casos tenia 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 tenia 7 copias de su “tabla” principal, que por supuesto no coincidían ni en atributos ni en cantidad de registros

CALIDAD DE DATOS

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

EJEMPLOS DE ERRORES CLASICOS

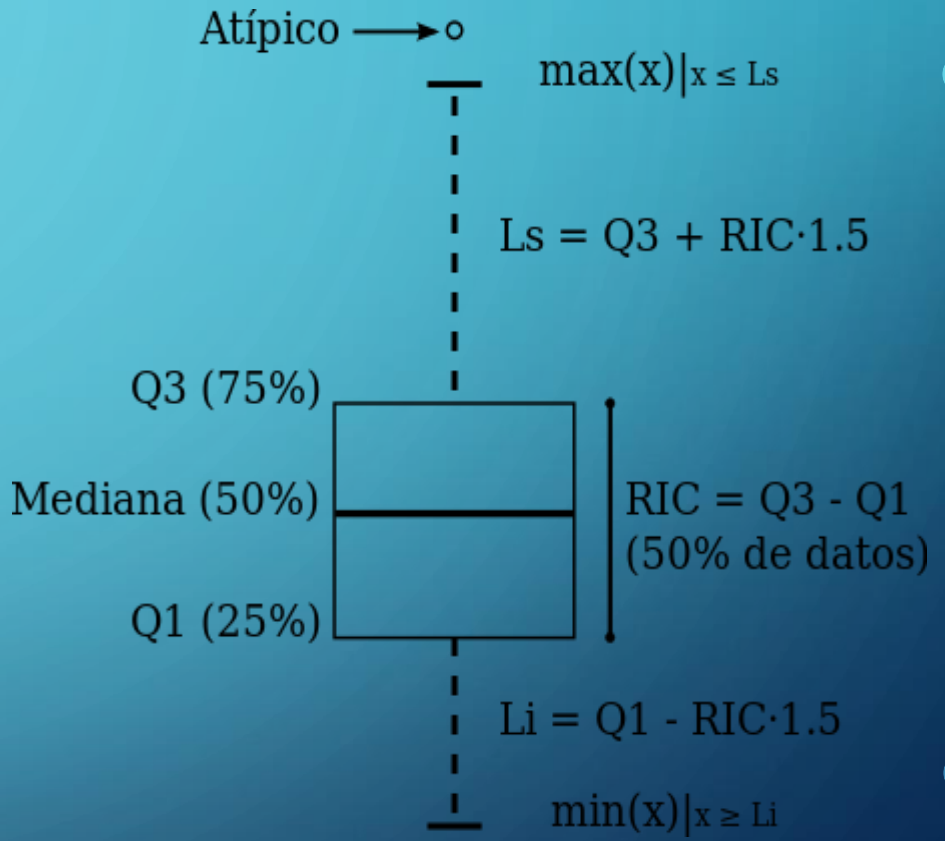
- Fuera de Rango: Edad del Paciente= 185 ()
- No-Standard: Data Main Str, Main Street, Main ST, Main St.
- Datos inválidos: El dato puede ser “A” o “B” pero el valor es “C”
- Reglas culturales diferentes:
  - Fecha= Enero1, 2002 o 1-1-2002 o 1 Ene 02
  - Montos en diferentes monedas
- Distintos Formatos: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

EJEMPLOS MAS SOFISTICADOS

- Tengo 3 bases de clientes y cada una tiene una dirección diferente , cual es la valida?
- 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
- Algunas cosas que “son sospechosas”
  - Tiene 7años y nivel de estudio de doctorado
  - Gasta $ 2.000.000 por mes de tarjeta

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?



AGENDA

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

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 compañia

NIVEL DE MADUREZ DEL GOBIERNO DE DATOS

Fuente, presentación "Ciencia, gobierno y monetización de datos" Maria del Rosario Bruera

PRINCIPALES ROLES 1/3

Chief Data Officer

- máximo responsable del programa de gobierno de datos ,
- liderar el equipo de gobierno de datos.
- Definir y/o colaborar en la definición (con poder de veto), sobre el alcance, objetivos, recursos, tiempos y prioridades de las iniciativas de gobierno de datos.
- Promover, negociar y justificar cambios en la estrategia de datos corporativa.

Arquitecto de datos

- Desarrollar la arquitectura de datos de la organización para atender requerimientos de negocio
- Desarrollar los estándares y procedimientos de diseño y modelado de datos a nivel corporativo
- Supervisar el diseño y modelado de datos para cada componente de la arquitectura
- Aprobar las características de desarrollo de aplicaciones e interfaces que impacten sobre la arquitectura de datos

PRINCIPALES ROLES 2/3

DATA OWNER (por dominio)

- Máxima autoridad de aprobación respecto a los issues/riesgos de gobierno dentro de su dominio
- Gestiona el ciclo de vida de datos, incluyendo los permisos de acceso a la información dentro de su dominio y respetando las políticas corporativas establecidas
- Gestiona la Calidad y Riesgo de los Datos dentro de su dominio
- Colabora en el gobierno corporativo
- Es el que conoce el significado de los datos

DATA STEWARD

- Apoyo a los Dueños de Datos
- Debe comprender tanto los procesos de negocio como los datos producidos por estos
- Es responsable de escribir e implementar las reglas de calidad de datos, medir, monitorear y remediar los problemas, y escalarlos en caso de ser necesario.
- Su rol forma parte del flujo de gobierno, al tener responsabilidades concretas (ej. Recibir alertas, realizar análisis) en la operación. Puede realizar acciones en nombre del Owner para liberar el flujo.

PRINCIPALES ROLES 3/3

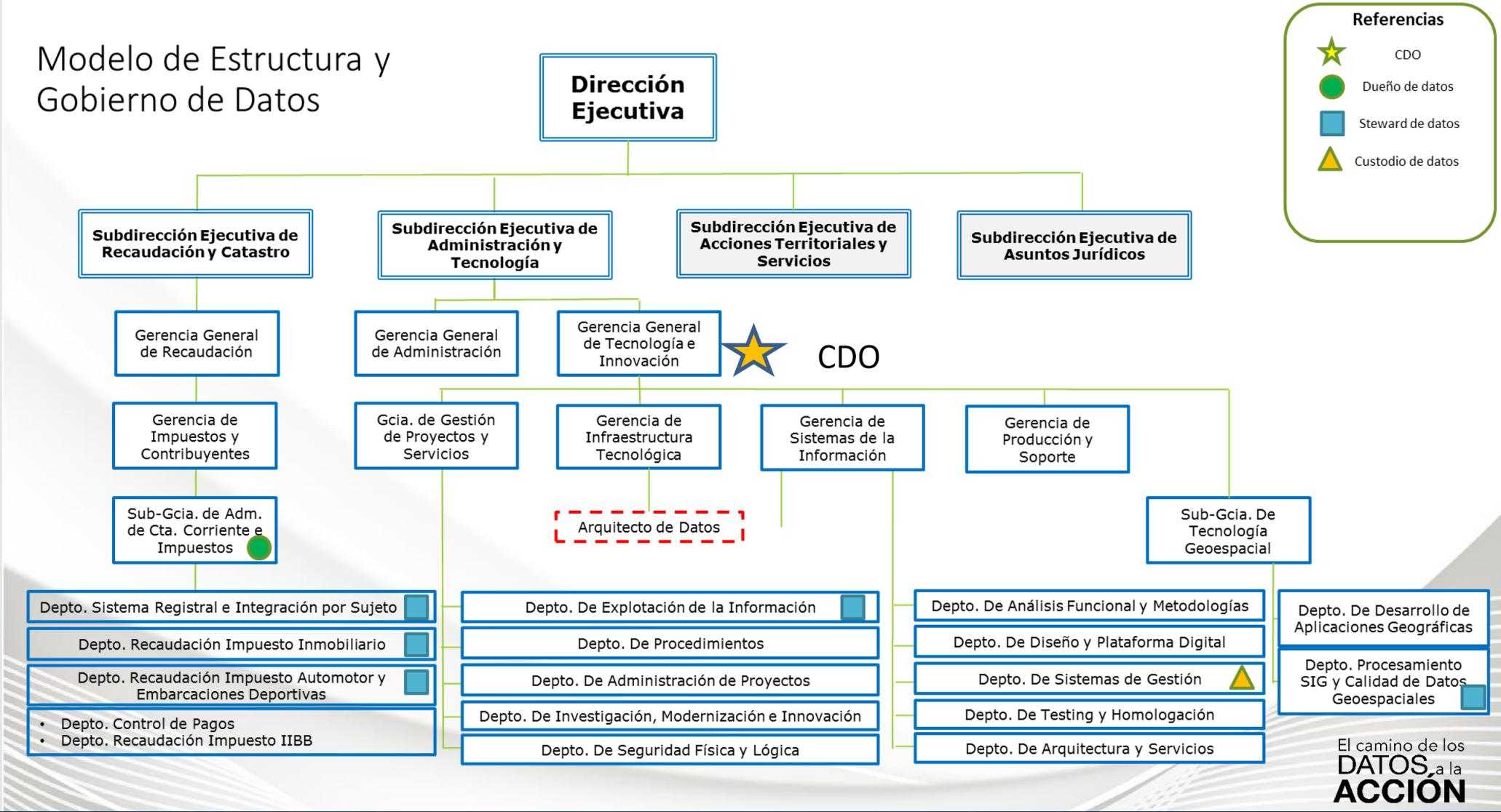
CUSTODIO DE DATOS

- Pertenece a las áreas de IT responsables por las plataformas, sistemas y aplicaciones en las que los datos residen.
- Son soporte a los Stewards y Owners para facilitar entendimiento de “bajo nivel” respecto a los atributos de datos (almacenamiento, estructuras, formatos, aplicaciones, etc.)
- Pueden tener responsabilidad operativa en el flujo de gobierno, como soporte al área de responsabilidad de los Stewards, bajo el entendimiento que típicamente tienen acceso a varios dominios de datos.
- Velan por la integridad y seguridad de los datos, y el cumplimiento de las políticas de gobierno, de acuerdo a su responsabilidad funcional sobre sistemas, aplicaciones y plataformas.

ENFOQUE

ARBA
AGENCIA DE RECAUDACIÓN
Buenos Aires Provincia

Modelo de Estructura y Gobierno de Datos



El camino de los DATOS, a la ACCIÓN

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.

AGENDA

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

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?
- Quien es el «dueño» de la tabla de cliente?
- Cuantas tablas de país tengo? Existe una equivalencia entre las mismas?

UNA TABLA CUALQUIERA….

CUC	TS_START	TS_END	ID_GRUP	FLAG_CP	FLAG_CF	FATTURATO	FLAG_FATT	
124589	30-lug-2004	1-gen-9999	92736	S	N	195000,00	N	
140904	15-mag-2001	15-giu-2005	35060	N	N	230600,00	N	
124589	5-mag-2001	30-lug-2004	92736	N	S	195000,00	S	
-452901	13-mag-2001	27-lug-2004	92770	S	N	392000,00	N	
129008	10-mag-2001	1-gen-9999	62010	N	S	247000,00	S	
-472900	10-mag-2001	1-gen-9999	62010	S	N	0,00	N	
130976	7-mag-2001	9-lug-2003	75680					

(Fuente : http://wp.sigmod.org/?p=871)

DATOS VS. INFORMACIÓN VS. CONOCIMIENTO

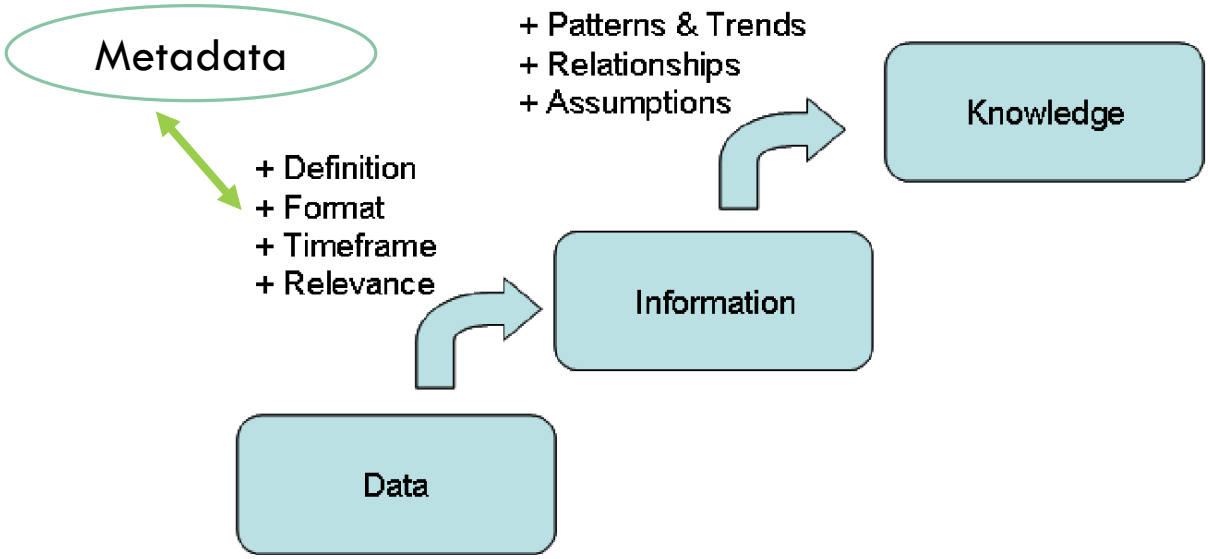


Figure 1.1 Data, Information, and Knowledge

Fuente: Dama Book

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.***

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) IRM: Data Administration VS. Database Administration, http://www.tdan.com/view-articles/4197

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 todos 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" Meta Data to Manage, http://www.tdan.com/view-articles/5069/

AGENDA

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

LA PRIVACIDAD ES UNA PREOCUPACIÓN CRECIENTE

- https://youtu.be/j-tFoYNHi1w
- 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

SITUACIÓN ARGENTINA

- Existe una “dirección nacional de protección de datos personales”
- https://www.argentina.gob.ar/aaip/datospersonales
- También una “agencia de acceso a la información publica”
- https://www.argentina.gob.ar/aaip

LEY DE HABEAS DATA

- Ley 25.326
- www.infoleg.gov.ar

BIBLIOGRAFIA

- The DAMA Guide to the Data Management Body of Knowledge , https://technicspub.com/dmbok/, la primera edición esta en la biblioteca
- Presentación “Gobierno de Datos” de Sandra D’Agostino

PREGUNTAS