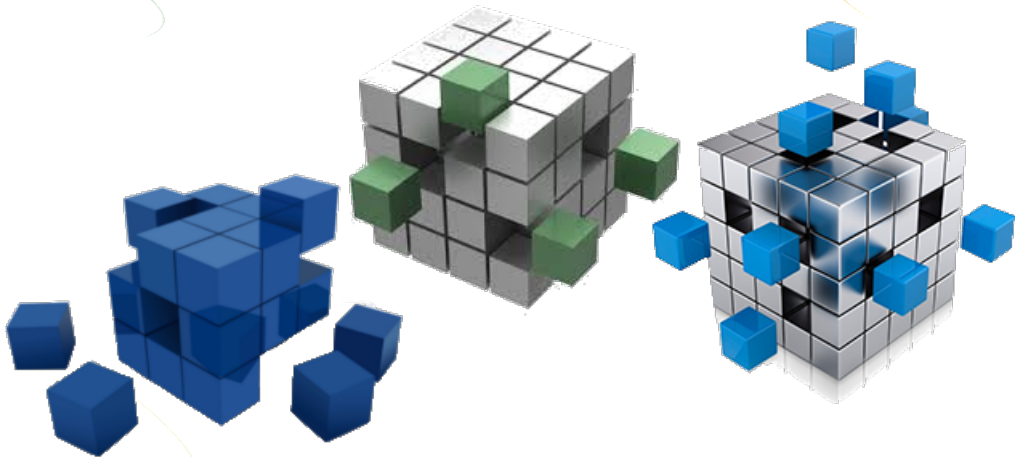


Fundamentos de Business Intelligence

Ing. Cabral, Juan B.



Universidad Nacional del Sur

SciPyCon Argentina 2014

10/2014 - Bahía Blanca - Argentina

About Me

Juan B Cabral

- Software engineer.
- Data scientist.

It's all about ME ME ME....

Unless it means being responsible for something. In which case, it's somebody else's fault, so don't look at ME.



somee cards
user card

Agenda

01. Historia y descripción del BI
02. Bases de datos transaccionales (OLTP) vs Analíticas (OLAP)
03. Data Marts y Data Warehouse
04. Facts y Dimensiones
05. Estructura de datos para análisis multidimensional (OLAP Cubes)
06. Implementaciones OLAP: ROLAP - MOLAP - HOLAP
07. Modelado relacional para RDBMS (ROLAP)
08. Diferentes alternativas de OLAP libres y gratuitas (Mondrian & Cubes)
09. Aplicaciones BI (Pentaho - Saiku - Cubes Viewer)
10. Consultas MDX (Multi Dimensional eXpressions)
11. ETL (Extract, Transform and Load)



Demo Time



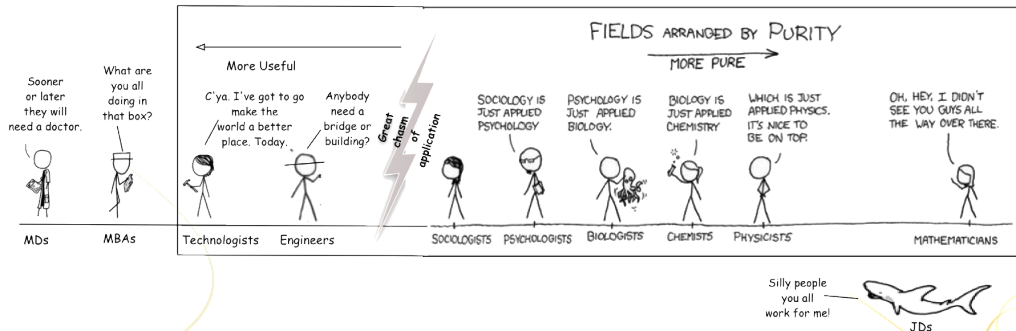
Veamos a que apuntamos con este tutorial

Historia y descripción del BI - Definición

El término inteligencias empresariales se refiere al uso de datos en una **empresa** para facilitar la toma de decisiones. Abarca la comprensión del funcionamiento actual de la **empresa**, bien como la anticipación de acontecimientos futuros, con el objetivo de ofrecer conocimientos para respaldar las decisiones **empresariales**. [WIKIPEDIA](#)

En 1989, Howard Dresner (más tarde, un analista de Gartner Group) propuso la "inteligencia de negocios" como un término general para describir "los conceptos y métodos para mejorar la toma de decisiones **empresariales** mediante el uso de sistemas basados en hechos de apoyo" [WIKIPEDIA](#)

Uno de los pocos casos que Nace en la industria migra a la Ciencia



Historia y descripción del BI - Características

- **Accesibilidad a la información.** El acceso a datos debe ser de forma independiente a su procedencia
- **Apoyo en la toma de decisiones.** La herramienta debe permitir la selección, análisis y manipulación selectiva de datos
- **Orientación al usuario final.** Se busca independencia entre los conocimientos técnicos de los usuarios y su capacidad para utilizar estas herramientas.



OLTP & OLAP - Versus otras Clasificaciones

Existen diferentes formas de clasificar bases de datos



- Segun la estructura que almacentan: **OO** (db4o), **Document-Oriented** (mongoDB, CouchDB), **RDBMS** (MySql, SQLite, PostgreSQL, Oracle, MicrosoftSQL Server, DB2), **Key-Value** (Redis, riak) o **Graph** (Neo4J)
- Segun si implementan o no SQL: **SQL** (MySql, SQLite, PostgreSQL, Oracle, MicrosoftSQL Server, DB2) o **NO-SQL** (Todas las demas)
- Segun su objetivo:
OLAP (Mondrian, Cubes, Cognos) y **OLTP** (Todas las demas)

OLTP & OLAP - OLAP vs OLTP

OLAP es el acrónimo en inglés de procesamiento analítico en línea (On-Line Analytical Processing). Es una solución utilizada en el campo de la llamada Inteligencia empresarial (o Business Intelligence) cuyo objetivo es agilizar la consulta de grandes cantidades de datos. ... contienen datos resumidos de grandes Bases de datos o Sistemas Transaccionales (OLTP). Se usa en informes de negocios de ventas, marketing, informes de dirección, minería de datos y áreas similares.

OLTP vs. OLAP

- | | |
|----------------------------|---------------------------------|
| • provides detailed audit | • provides big picture |
| • supports operations | • supports analysis |
| • needs detailed data | • needs aggregate data |
| • find one dataset quickly | • evaluate all datasets quickly |
| • relational model | • multidimensional model |

Q: "WHO lives in
A town?"

Q: "HOW MANY live in
A town?"

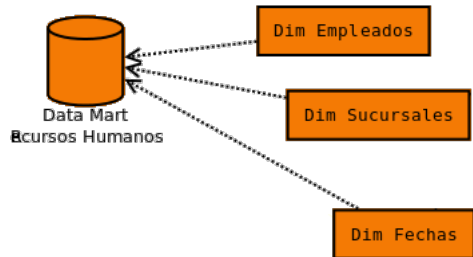
Data Marts

Un **Data mart** es una versión especial de almacén de datos. Son subconjuntos de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones.

Los Data marts son subconjuntos de datos de un almacén de datos para áreas específicas.

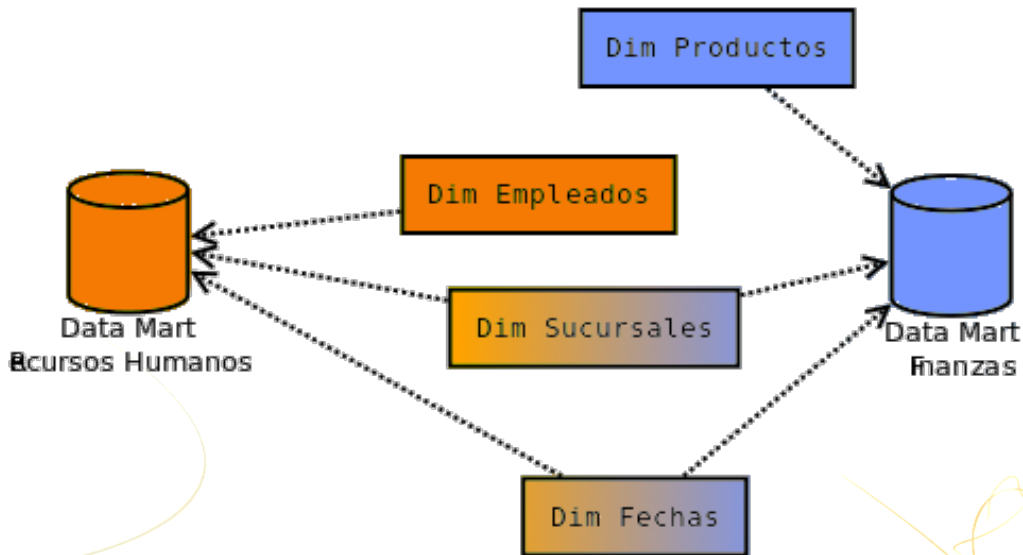
Entre las características de un data mart destacan:

- Usuarios limitados.
- Área específica.
- Tiene un propósito específico.
- Tiene una función de apoyo.



Data Warehouse

Según Ralph Kimball un almacén de datos o **Data Warehouse** es: "una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis"[cita requerida]. También fue Kimball quien determinó que un data warehouse no era más que: "la unión de todos los *Data marts* de una entidad". Defiende por tanto una metodología ascendente (bottom-up) a la hora de diseñar un almacén de datos.

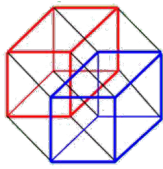


Hechos y Dimensiones - Definición

Hechos (o *Facts*)

Un hecho es un valor o una medida que representa un hecho (*sic) sobre una entidad o un sistema.

Es algo que efectivamente sucedió o existe y sobre los cuales queremos efectuar análisis.



Dimensiones

Es una estructura que categoriza a hechos y medidas para permitir responder preguntas del negocio.

El valor de una dimensión se llama **miembro**

Si una dimension se puede separar en "partes" suele llamarsele **Jerarquias** (Fecha = año + mes + día).

Las "partes" se llaman niveles.

Hechos y Dimensiones - Un ejemplo

Tito fue a comprar jabón en polvo gasto en total \$16 en la sucursal 7 el 16 de octubre del 2014

- **Hecho:** Sucedió 1 (métrica) venta que se gasto \$ 16 (métrica).
- **Dim. Cliente:** tito
- **Dim. Producto:** jabón en polvo.
- **Dim. Sucursal:** 7
- **Dim. Fecha:** 16 de octubre del 2014

Consultas multidimensionales basándonos en el ejemplo:

1. Promedio de gastos por cliente.
2. Quiero el promedio de de las ventas por producto y sucursal.
3. Quiero la suma de ingresos por producto.
4. Quiero conteo de ventas por día.



Hechos y Dimensiones - Ejemplo Científico

El telescopio X encontró una estrella tipo RR-Lyrae con una magnitud aparente Y en la posición Z en la fecha W.

- **Hecho:** Sucedió 1 (métrica) descubrimiento de una estrella de magnitud aparente Y (métrica).
- **Dim. Dispositivo:** Telescopio X
- **Dim. Tipo de Fuente:** RR-Lyrae
- **Dim. Zona:** rango R tal que R contiene a Z
- **Dim. Fecha:** W

Consultas multidimensionales basándonos en el ejemplo:

1. Cantidad de descubrimientos por posición.
2. Promedio de magnitud por tipo de fuente.



Dimensiones - Tipos

- **Regular:** cliente, articulo, tipo de fuente
- **Conformed:** Conectan mas de un datamart y tienen mismo significado semántico en todos los datamarts
- **Role Played:** Cambian de significado según el datamart
- **Junk:** Suelen tener banderas como [S|N] o Sexo
- **Dirty:** Son *role-playing* que no tienen significado en si mismos. Por ejemplo: una dimensión numero que en un datamart es un identificador de facturas y en otro es un DNI.



Slowly Change Dimension

- Se supone que una DW no cambia mucho en sus dimensiones.
- Si alguna cambia: **cambia lentamente**

Suponiendo que tengo alguna dimension con un miembro parecido a:

```
{sk: 1, bk: 001, nombre: "Plutón", cat: "Planeta"}
```

Ahora Plutón no es mas un planeta y existen **tres enfoques para cambiar los datos**

a. **SCD Tipo 1:** No Guardo Historia.

```
{sk: 1, bk: 001, nombre: "Plutón", cat: "Planeta Enano"}
```

b. **SCD Tipo 2:** Guardo Historia Versionando.

```
{sk: 1, bk: 001, nombre: "Plutón", cat: "Planeta", ver: 1}  
{sk: 1, bk: 001, nombre: "Plutón", cat: "Planeta Enano", ver: 2}
```

c. **SCD Tipo 3:** Guardo Historia Cambiando la Dimensión.

```
{sk: 1, bk: 001, nombre: "Plutón", cat0: "Planeta", cat1: "Planeta Enano"}
```

Cubos OLAP

Es una base de datos multidimensional, en la cual el almacenamiento físico de los datos se realiza en un vector multidimensional.

Pueden considerarse como una ampliación de las dos dimensiones de una hoja de cálculo.

¿Preguntas?

- Charla:
- **Contactos:**

- jbcabral.com
- Juan B Cabral <jbc.develop@gmail.com>

I hate it when
someone starts asking
me questions just
after waking up.



AR...EHHM...ARR U ~~ARR~~ KIDDIN...MEEE?

WIKIPEDIA(1, 2) http://es.wikipedia.org/wiki/Inteligencia_empresarial