CIFF
Centro Internacional
de Formación Financiera

CIFF Trustees:

Santander

Universidad de Alcalá

Business Intelligence
&
Data Mining

Profesor:
Pedro Pasquau

Octubre 2015

MASTER EN BUSINESS ANALYTICS & BIG DATA
2015

Presentación

Presentación

- **Profesor**:

  - Pedro Pasquau
  - Partner & Head of Business Intelligence And Data Mining
  - Especialización Open Source:
    - OLAP, DataWarehousing, ETL, Reporting
    - Data Mining, Machine Learning, IA
    - Analítica de datos

  pedro.pasquau@gmail.com

- **Alumnos**?

Planificación

Presentación

- **Sesión 1A**: **Introducción al Business Intelligence**
  - ❑ Business Intelligence
  - ❑ BI, BA y Big Data: Contexto
  - ❑ Pentaho BA
- **Práctica I: Modelado analítico**
- **Sesión 1B: ETL y Adquisición de Datos**
  - ❑ ETL
  - ❑ Pentaho Data Integration
  - ❑ Data Integration y Data Mining: Realimentación
- **Práctica II: Modelado BI**

Presentación

- **Sesión 2A: Data Mining – Machine Learning**
  - ❑ Algoritmos
  - ❑ Ejemplos
- **Sesión 2B: Pentaho Data Scientist (PDS)**
  - ❑ PDS (Weka)
  - ❑ DW→ETL→ PDS
  - ❑ PDS → ETL→DW
- **Práctica I: BA con Weka**
- **Práctica II: BA con R**

Introducción al Business Intelligence

1. ¿Qué es la inteligencia de negocio?
2. ¿Qué tipos de análisis puedo realizar?
3. ¿Dónde y cómo se almacenan los datos?
4. Convertir ‘datos’ en ‘información’
5. ¿Qué es un Data Warehouse?
6. Sistemas de Explotación

Introducción al Business Intelligence

- Llamamos **business intelligence** (BI) al conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización.

- **Características:**

  - **Accesibilidad a la información.** Los datos son la fuente principal de este concepto. Lo primero que deben garantizar este tipo de herramientas y técnicas será el acceso de los usuarios a los datos con independencia de la procedencia de estos.

  - **Apoyo en la toma de decisiones.** Se busca ir más allá en la presentación de la información, de manera que los usuarios tengan acceso a herramientas de análisis que les permitan seleccionar y manipulas sólo aquellos datos que les interesen.

  - **Orientación al usuario final.** Se busca independencia entre los conocimientos técnicos de los usuarios y su capacidad para utilizar estas herramientas.

Introducción al Business Intelligence

No confundir con Business Intelligence:

- **Proceso Analítico**: Es el proceso de recoger y analizar datos, tales como el comportamiento de los consumidores o sus patrones de compra; y utilizarla para desarrollar una serie de iniciativas como pueden ser acciones de promocion o lanzamiento de nuevos servicios. El análisis predictivo ayuda a las compañías a tomar decisiones en base a lo que es más probable que pueda suceder en el futuro, gracias al uso de modelos estadísticos y grandes bases de datos.

- **Business Intelligence:** Es un proceso que implica la utilización de herramientas, tales como software de aplicaciones y metodologías, y que ayudan a los usuarios, mediante su uso a tomar mejores decisiones. Dentro del Business Intelligence se incluye el reporting, query o consulta, análisis, cuadros de mando, scorecards, workflow, etc…

- **Centros de competencia:** Son grupos de personas dentro de una organización que desarrollan las mejores prácticas para un área o actividad determinada. En el caso del Business Intelligence, establecen standards para seleccionar las mejores herramientas, ponerlas a disposición de los usuarios y asesorarles en su correcto uso.

Introducción al Business Intelligence

- **Modelización:** Se trata de una técnica de análisis que mira los datos del pasado y posibilita a la empresa para predecir lo que ocurrirá en el futuro, en base a una serie de escenarios y condiciones.

- **OLAP (ONLINE ANALYTICAL PROCESSING):** Es un conjunto de tecnologías y aplicaciones de software que permite recoger los datos de la compañía, almacenarlos e indagar sobre ellos de forma rápida e intuitiva. Se trata de crear una 'capa de negocio' con lenguaje funcional por encima de estructuras complejas de Bases de Datos relacionales.

- **Reporting:** Es un elemento clave de cualquier solución Business Intelligence. Generar informes posibilita a los usuarios observar la marcha del negocio. Los informes deben incluir ratios financieros, datos de ventas, información sobre los clientes, cálculos estadísticos, etc…

Introducción al Business Intelligence

- **DataWarehouse:** Se trata de una gran Base de Datos centralizada que integra datos de varias fuentes dentro de una empresa. El DataWarehouse puede estar distribuido en diferentes plataformas, sistemas, incluso Base de datos. Los datos, generalmente se organizan por criterios tales como provincias, departamentos, fechas, etc…

- **Scorecards:** Permiten medir el funcionamiento de una compañía mediante la identificación de unas métricas clave (KPI´s, Key Performance Indicators). Los scorecards ayudan en determinar si una compañía esta consiguiendo unos determinados objetivos, si hace progresos o si hay aspectos claramente deficitarios que inciden directamente en el resultado global de la empresa.

- **Cuadros de Mando**: Los Cuadros de mando condensan grandes volúmenes de información en entornos visuales muy llamativos y prácticos. Mediante el uso de gráficos, mapas y otros recursos visuales se proporciona un entorno muy intuitivo. Consiguen hacer sencillo complejos modelos de datos, formulas y relaciones entre las variables.

Introducción al Business Intelligence

1. ¿Qué es la inteligencia de negocio?
2. ¿Qué tipos de análisis puedo realizar?
3. ¿Dónde y cómo se almacenan los datos?
4. Convertir ‘datos’ en ‘información’
5. ¿Qué es un Data Warehouse?
6. Sistemas de Explotación

Introducción al Business Intelligence

Para consultar la base de datos es necesario diseñar complejas consultas SQL. Así, **la obtención de respuestas a sencillas preguntas de negocio puede ser bastante complicado** y puede requerir de bases de datos temporales y complicadas consultas de varias páginas.

Los administradores deben conocer el negocio asociado a estas. Si esto no es así, el resultado que se obtenga puede ser erróneo. Además se debe elegir correctamente las herramientas que manejarán los usuarios.

A continuación se van a mostrar una primera categorización de las **herramientas de usuario final**:

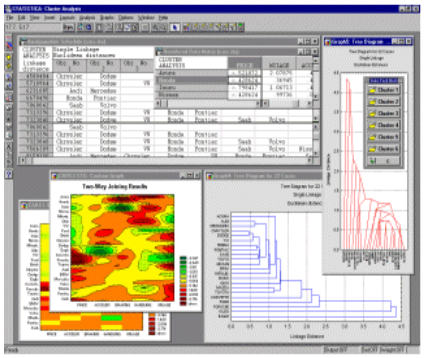
Introducción al Business Intelligence

- **Informes estáticos y dinámicos:** Los informes estáticos presentan la información de una manera predeterminada (p.e. ventas por mes y por región). Los informes dinámicos permiten a los usuarios interactuar con la información mediante “drill-down” para descender a niveles de más detalle. En cualquier caso, el dinamismo de estos informes también se encuentra predefinido.

- **OLAP:** El significado de estas siglas es “On-Line Analytical Processing”. Estas herramientas permiten al usuario tener acceso directo “on-line” al data warehouse, de manera que pueden interrogarlo con la consultas que deseen y puedan navegar libremente por la información.

Introducción al Business Intelligence

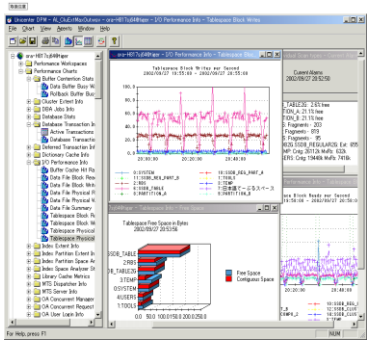
- **Data mining:** El objetivo del data mining es reconocer patrones y relaciones que no son evidentes si se emplean métodos de análisis más simples. El data mining es el corazón de la eficiencia de la bases de datos de clientes o usuarios.



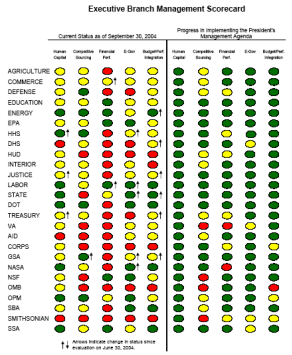
- **Excepciones y notificaciones:** Este tipo de herramientas también son conocidas como gestores y **generadores de alertas y alarmas**. Consisten en agentes software que cuando determinados indicadores clave de la gestión empresarial muestran valores que divergen de los objetivos perseguidos, se activan automáticamente para avisar al responsable adecuado, o para tomar las medidas adecuadas.

Estas nuevas herramientas permiten a los usuarios vincular eventos con la notificación correspondiente (por ejemplo: Si el tiempo no facturable del soporte técnico es superior a dos horas diarias, entonces los responsables del soporte recibirán un ***e-mail, un fax, o un SMS***).

Introducción al Business Intelligence



• **Presupuestos y Predicción:** Las tareas de elaboración de presupuestos es una de las labores más tediosas que tiene que abordar el departamento contable de una organización. Los contables tienen que cuadrar y distribuir las diferentes partidas presupuestarias entre los diferentes departamentos y oficinas de la organización.



• **Balanced Scorecards:** Los scorecards es un tipo de indicador cuyos padres son Robert Kaplan y David Norton de la Harvard Business School. Los scorecards permiten medir el rendimiento de individuos y grupos de individuos frente a los objetivos claves de la organización (tanto financieros, como no financieros).

Introducción al Business Intelligence

1. ¿Qué es la inteligencia de negocio?
2. ¿Qué tipos de análisis puedo realizar?
3. ¿Dónde y cómo se almacenan los datos?
4. Convertir ‘datos’ en ‘información’
5. ¿Qué es un Data Warehouse?
6. Sistemas de Explotación

Introducción al Business Intelligence

ENFOQUE RELACIONAL

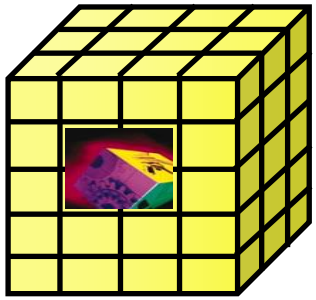
- Obtiene muy buen rendimiento en el tratamiento transaccional de los datos.
- No es eficaz de cara al análisis de la información con muchos datos involucrados.
- Los usuarios tienen dificultades en la comprensión del modelo y no les es posible "navegar" a través de él.
- No modeliza el negocio, sino las relaciones entre los datos.

ENFOQUE MULTIDIMENSIONAL

- Facilita la comprensión por parte de los usuarios: es intuitivo.
- Permite realizar análisis dimensional y "navegar" por el modelo para obtener la información deseada (herramientas OLAP).
- Acceso eficiente a la información, ya que se puede predecir el tipo de consultas que se efectuarán.
- Fácilmente extensible para dar respuesta a nuevos requerimientos.
- Existen soluciones estándar para modelizar situaciones habituales en el mundo de los negocios.

Relacional

Multidimensional



Introducción al Business Intelligence

Relacional

Los modelos de datos relacionales también son conocidos como modelos de datos **Entidad/Relación (ER).**

Estos modelos son los más adecuados y los comúnmente empleados para la implementación de sistemas operacionales.

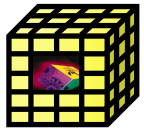
En estos modelos, las **entidades** representan "cosas" que son relevantes para el sistema a construir (usuarios, subvenciones, provincias, etc...). De acuerdo a la teoría Entidad/Relación, cada entidad se debe corresponder con una tabla en la base de datos.

Las entidades tienen **relaciones** (por ejemplo, los habitantes se encuentran relacionados con las provincias).

Los **atributos** son empleados para describir las entidades. Los atributos de un producto pueden incluir su nombre, su número de identificación, una descripción, el precio, etc... Los atributos se corresponden con las columnas de las tablas que representan la entidad.

Introducción al Business Intelligence

Multidimensional



OLAP es online analytical processing. Se trata de una forma de almacenar la información en una Base de Datos que permita realizar de forma más efectiva las queries. Es una definición abreviada, claro esta, la realidad es más compleja.

MOLAP, Multidimensional OLAP. Tanto los datos fuente como los datos agregados o precalculados residen en el mismo formato multidimensional. Optimiza las queries, pero requiere más espacio de disco y diferente software. El primer punto esta dejando ser un problema: el espacio de disco cada vez es más barato.

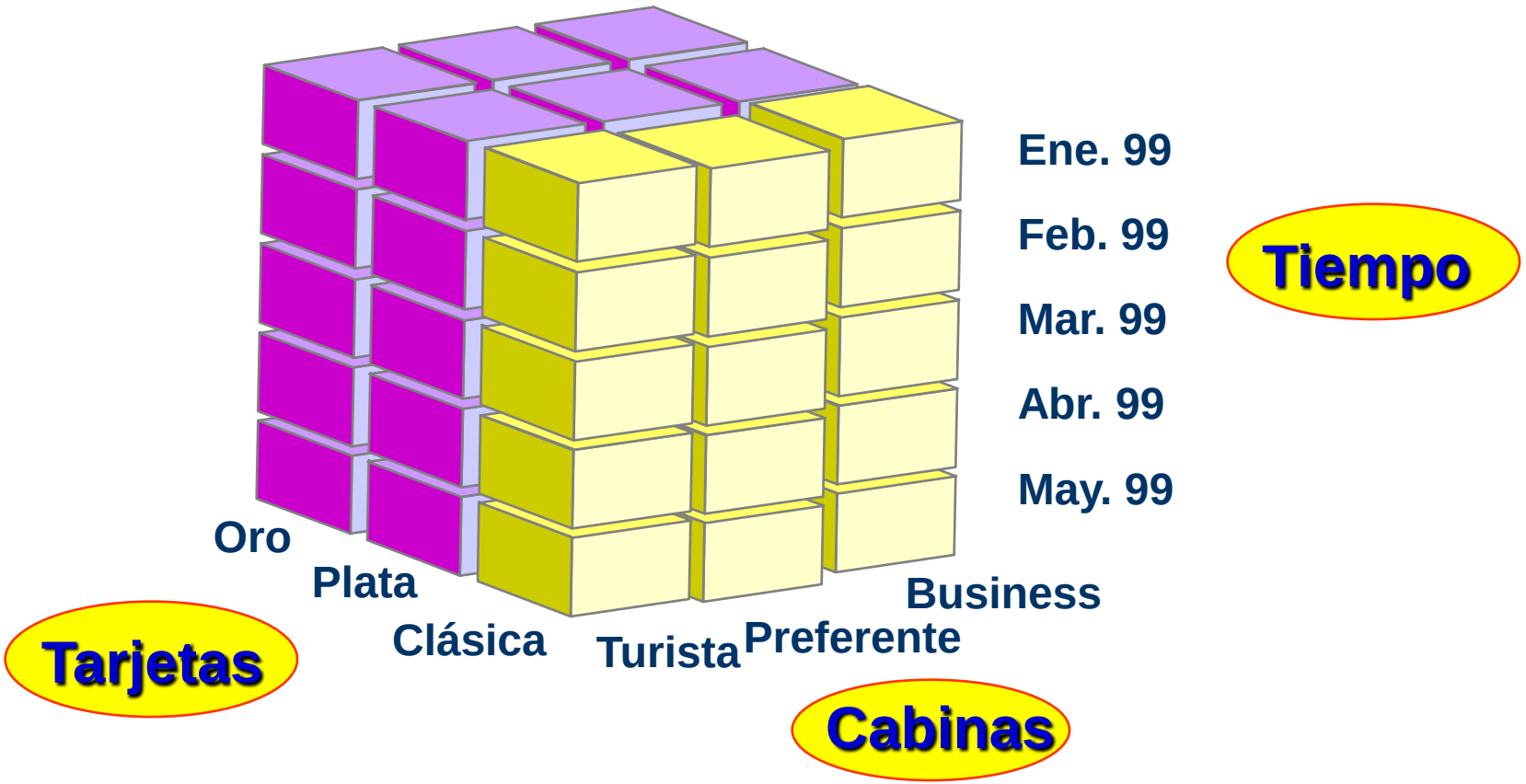
ROLAP, Relational OLAP. Tanto los datos precalculados y agregados como los datos fuente residen en la misma base de datos relacional. Si el DataWarehouse es muy grande o se necesita rapidez por parte de los usuarios puede ser un problema.

HOLAP, Hybrid OLAP: Es una combinación de los dos anteriores. Los datos agregados y precalculados se almacenan en estructuras multidimensionales y los de menor nivel de detalle en el relacional. Requiere un buen trabajo de análisis para identificar cada tipo de dato.

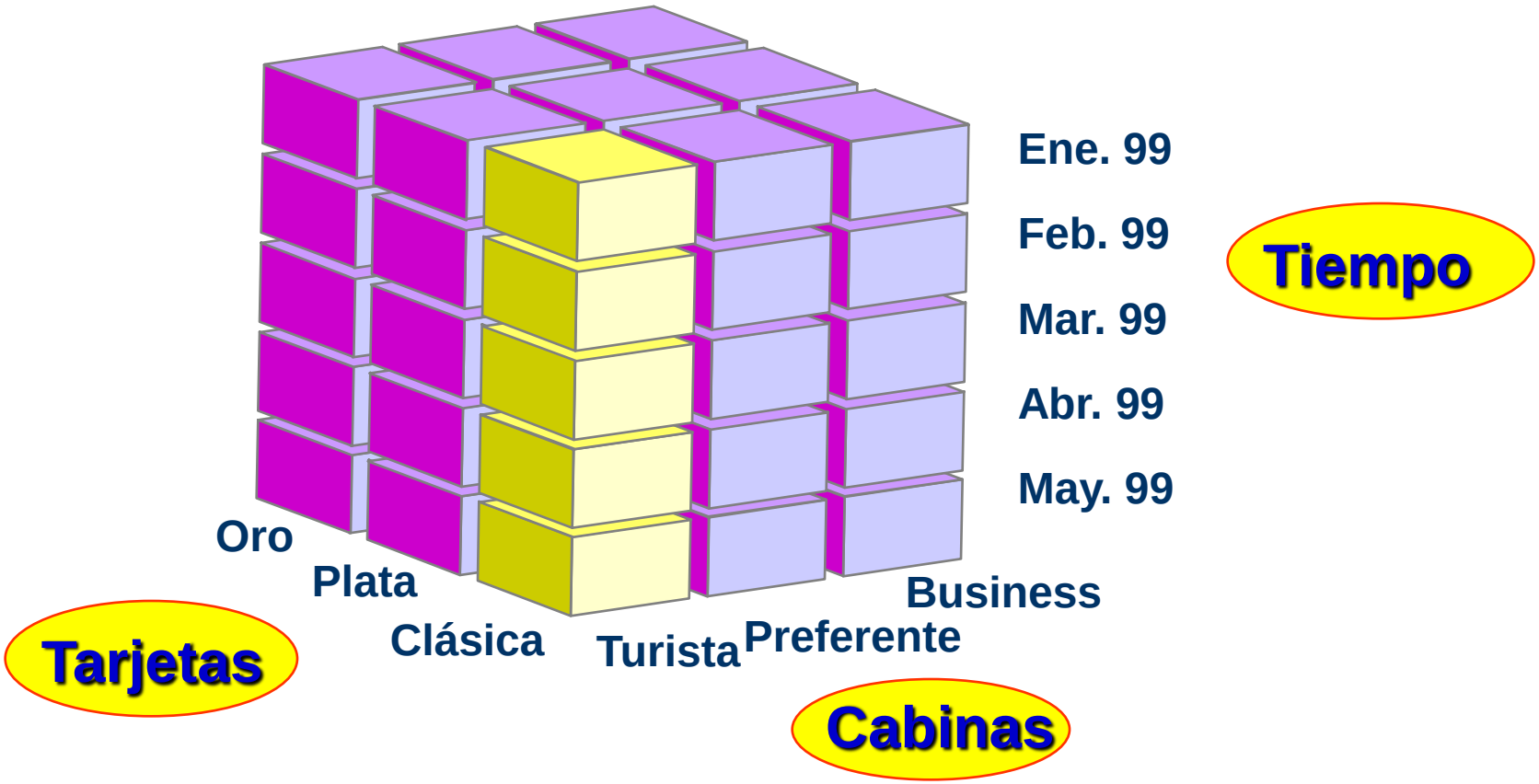
Los 6 elementos básicos OLAP:

- Dimensiones
- Valores
- Jerarquías
- Niveles
- Atributos
- Indicadores

Puntos Acumulados



Puntos Acumulados



Introducción al Business Intelligence

1. ¿Qué es la inteligencia de negocio?
2. ¿Qué tipos de análisis puedo realizar?
3. ¿Dónde y cómo se almacenan los datos?
4. Convertir ‘datos’ en ‘información’
5. ¿Qué es un Data Warehouse?
6. Sistemas de Explotación

El problema de la calidad de los datos

Durante el proceso de construcción de un data warehouse, uno de los principales y más complejos problemas a resolver es la calidad de la información procedente de las Fuentes de Datos.

CUST #	PRODUCTO	DIRECCION	TIPO
90328574	Digital Equipment	187 N. PARK St. Salem NH 01458	OEM
90328575	DEC	187 N. Pk. St. Saleem NH 01458	OEM
90238475	Digital	187 N. Park St Salem NH 01458	$#%
90233479	Digital Corp	187 N. Park Ave. Salem NH 01458	Comp
90233489	Digital Consulting	15 Main Street Andover MA 02341	Consult
90234889	Digital Info Service	PO Box 9 Boston MA 02210	Mail List
90345672	Digital Integration	Park Blvd. Boston MA 04106	SYS INT

No hay clave única

Anomalías

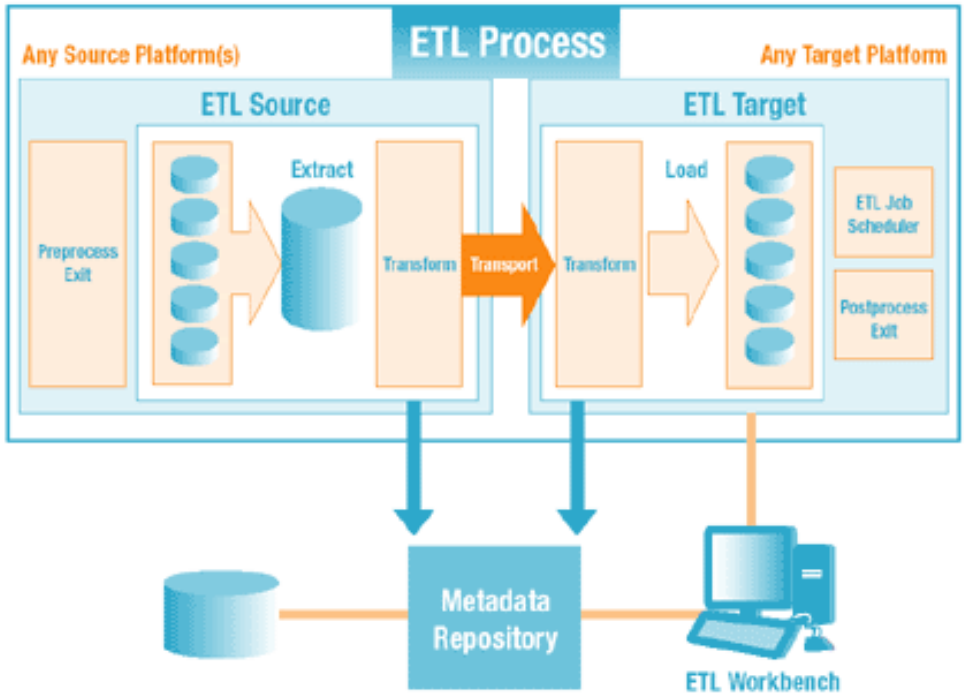
No hay standarización

Ortografía

Caracteres erróneos

Introducción al Business Intelligence

La fase de extraccion, transformación y carga de los datos (ETL)



La ejecución fallida de esta fase puede suponer el fracaso del proyecto.

Introducción al Business Intelligence

1. ¿Qué es la inteligencia de negocio?
2. ¿Qué tipos de análisis puedo realizar?
3. ¿Dónde y cómo se almacenan los datos?
4. Convertir ‘datos’ en ‘información’
5. ¿Qué es un Data Warehouse?
6. Sistemas de Explotación

Los Objetivos de un DW

• **Facilitar el acceso a la información corporativa:** Los contenidos del data warehouse deben ser entendibles, navegables y su acceso debe estar caracterizado por el alto rendimiento:

- Entendible significa correctamente etiquetado.
- Navegable significa que el destino deseado se encuentra localizable en la pantalla y que este se encuentre a un solo “clic de distancia”.
- Alto rendimiento en el acceso significa que el tiempo de espera es nulo.
- Dotar de consistencia a la información de la organización.

• Cuando los usuarios deseen realizar nuevas consultas contra el data warehouse, los datos existentes y las tecnologías empleadas no deben ser cambiadas o modificadas.

Introducción al Business Intelligence

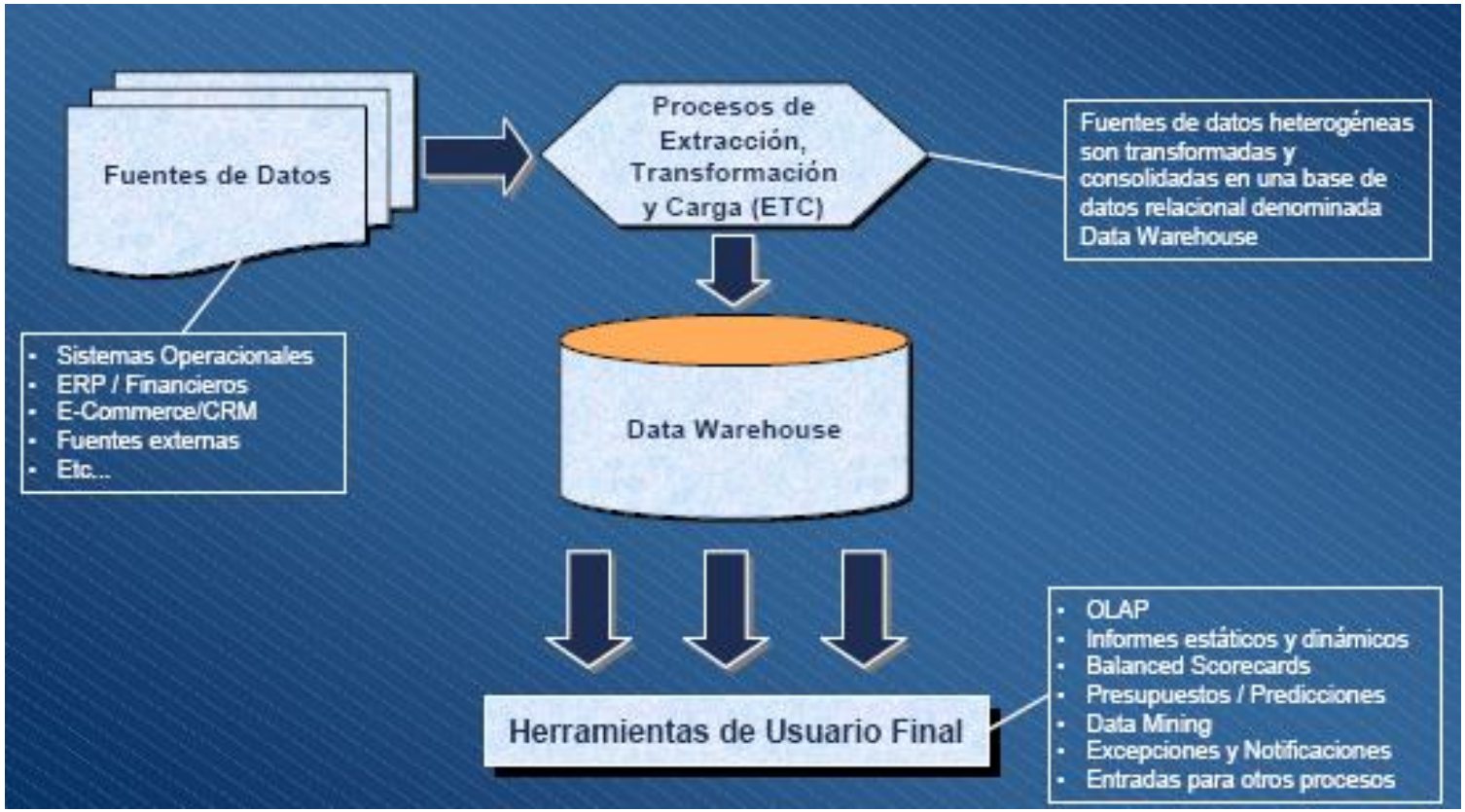
Lo mismo debe ocurrir cuando se introduce nueva información en el data warehouse. El diseño de diferentes data marts cuya suma formen el data warehouse debe ser llevada a cabo de manera cuidadosa e incremental.

• **Es un "seguro de vida" para proteger la información de la organización:** El data warehouse no sólo controla el acceso a los datos de manera efectiva, sino que suministra a los "dueños" de la información gran control y potentes medios acerca de quien usa y abusa de los datos.

• **Es el cimiento para la toma de decisiones:** El data warehouse tiene los datos adecuados para llevar a cabo el proceso de toma de decisión. El nombre original inicialmente empleado para hacer referencia a lo que hoy en día conocemos como data warehouse, sistemas de ayuda a la toma de decisión (decisión support system, DSS), muestra claramente cual es el objetivo de estos sistemas.

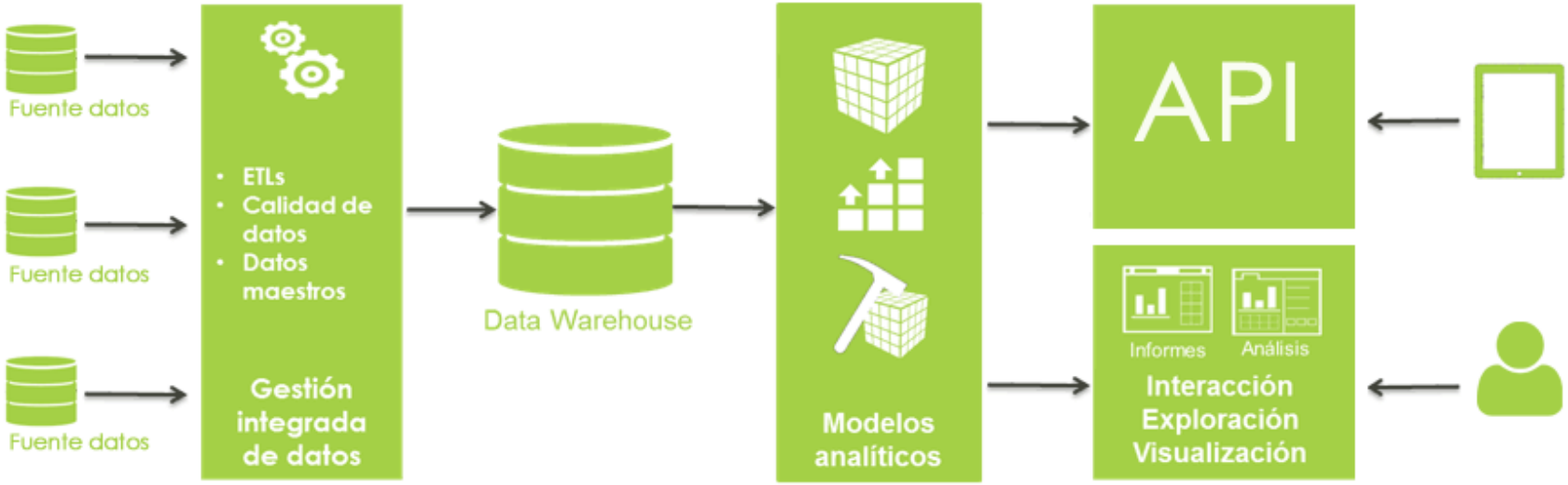
Introducción al Business Intelligence

Diagrama de un entorno DW



Introducción al Business Intelligence

Diagrama de un entorno DW



Tablas de Hechos

La tabla de hechos contiene todos los **datos que son relevantes** para el negocio de la entidad.

Cada registro de "hechos" está compuesto por un conjunto de claves foráneas (una clave por cada dimensión) y uno o varios valores numéricos.

Los valores vienen determinados por los objetivos de la organización y pueden incluir medidas como número de habitantes, importe de las subvenciones, número de inspecciones, edad media de los cargos electos, etc...

Generalmente, los valores son empleados para calcular otros valores. Por ejemplo, dividiendo el nº de inmigrantes por la poblacion total obtenemos el porcentaje de inmigrantes de cada poblacion.

De forma habitual, este tipo de cálculos se llevan a cabo durante el procesamiento de la consulta solicitada, por este motivo, este tipo de valores suelen denominarse "valores virtuales".

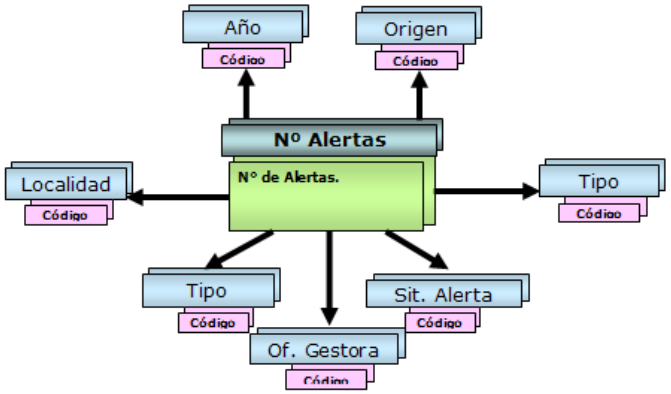
Introducción al Business Intelligence

Tablas de Dimensiones

Las dimensiones describen hechos, las tablas de dimensiones contienen numerosos atributos que permiten describir un hecho en mayor detalle.

La construcción de consultas multidimensionales sobre un modelo en estrella (multidimensional) es muy sencillo, debido a que el formato general de estas consultas sigue el siguiente formato:

Muéstrame el <hecho> de la <dimensión> de la <dimensión> .... de la <dimensión>.



Los principales componentes de un modelo multidimensional son:

- Facts o "hechos": variables numéricas que miden el negocio.
- Dimensiones: principales ejes de análisis de la información.
- Niveles o atributos: categorías dentro de una dimensión.
- Jerarquías: ordenaciones de los atributos en la dimensión.

Introducción al Business Intelligence

1. ¿Qué es la inteligencia de negocio?
2. ¿Qué tipos de análisis puedo realizar?
3. ¿Dónde y cómo se almacenan los datos?
4. Convertir ‘datos’ en ‘información’
5. ¿Qué es un Data Warehouse?
6. Sistemas de Explotación

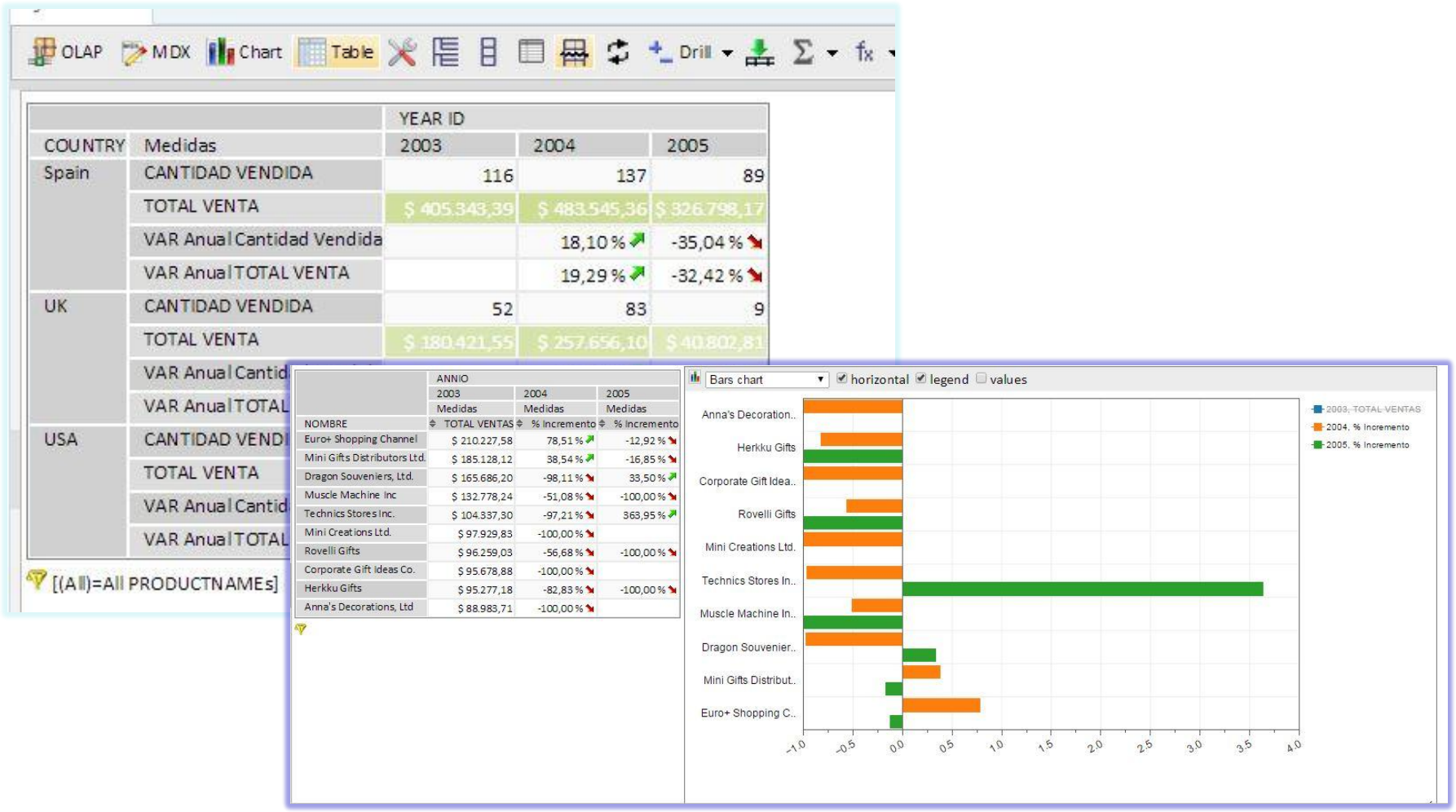
Introducción al Business Intelligence

Las herramientas de BI son usadas por usuarios finales para acceder, analizar y generar informes, CM, etc a partir de la información que frecuentemente reside en data warehouses, data marts o bases de datos operacionales.

- Cuadros de Mando
- Visores OLAP
- Reporting
- Scorecard
- Alertas
- ….

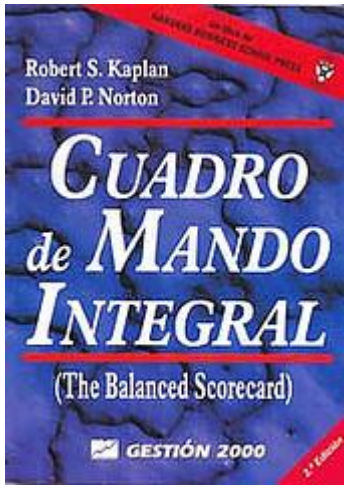
Introducción al Business Intelligence

Introducción al Business Intelligence



Introducción al Business Intelligence

¿Qué es un Scorecard?



La estrategia de una organización intenta describir como crea valor para sus accionistas, clientes y empleados. El Balance Scorecard es una herramienta poderosa que nos sirve para identificar los aspectos más críticos de una empresa.

Presentado en 1992 por **Robert S. Kaplan y David Norton**, el **Cuadro de Mando Integral o balance scorecard (BSC)** es un método para medir las actividades de una compañía en términos de su visión y estrategia. Proporciona a los administradores una mirada abarcativa del rendimiento del negocio.

Es una herramienta de management que muestra continuamente cuando una compañía y sus empleados alcanzan los resultados perseguidos por la estrategia.

También es una herramienta que ayuda a la compañía a expresar los objetivos e iniciativas necesarias para cumplir con la estrategia.

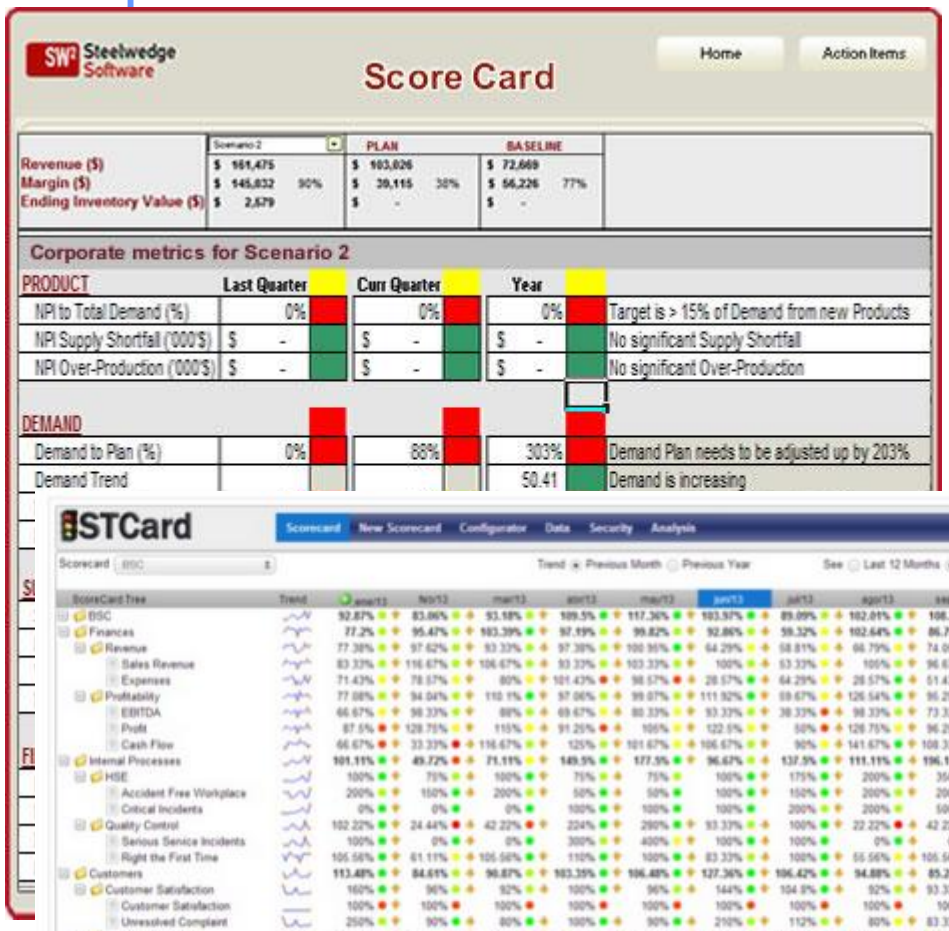
Definición de KPI´s o Indicadores de Negocio

Intentar construir un sistema de Inteligencia de negocio sin la ayuda de las métricas o **indicadores clave de negocio (KPIs)** es como intentar dirigir un barco sin timón, sextante, ni catalejo.

Sin las métricas, nos será imposible saber si el sistema que hemos creado se puede considerar satisfactorio o exitoso por parte de los usuarios finales.
No tendríamos ninguna idea sobre información tan importante como: tiempos de respuesta, utilización de la máquina, disponibilidad, satisfacción del usuario o calidad de los datos.

Introducción al Business Intelligence

Representacion visual de un Scorecard



-Indicadores por áreas.

- Codificación Semafórica.

- Explicación detallada del comportamiento del KPI.

- Indicar el área o responsable del mismo.

- Periodo de análisis para el que se analiza.

- Análisis o acciones a tomar.

BI, BA y Big Data

- **Business Intelligence es compatible con Big Data?**

- **Evolución BI to BA**

- **BI y Big Data: Metodología y Tecnología**

  - ***Business Intelligence*** **(tradicional)** se fundamenta en estructurar información empresarial útil y relevante para la toma de decisiones corporativas. Esta información suele ser conocida y focalizada en determinados ámbitos (ventas, producción, calidad…), inclusive nos permite generar sistemas proactivos de previsión con datamining. En cualquier caso siempre basándose en datos existentes y focalizados.

  - ***Big Data*** trata la información desde otra perspectiva, por la complejidad, cantidad, velocidad de ésta. Toda la información desestructurada disponible participa (o podría participar) en el sistema y esta se obtiene de nuestro sistema o de datos en la nube. Twitter registra 400 millones de tuits al día, Facebook 3000 millones de comentarios… compañías como estas hace años que utilizan estos sistemas de análisis de información. En cualquier caso, esos datos pueden ser o no ser útiles, el objetivo es analizarlos y localizar patrones, tendencias que nos permitan "destilar" esa ingente cantidad de información que hace unos años sería inimaginable.

- **Perfiles y Evolución**

BI, BA y Big Data

➢ **Pentaho Business Analytics**

-Descarga:
http://sourceforge.net/projects/pentaho/files/Business%20Intelligence%20Server/5.3/biserver-ce-5.3.0.0-213.zip/download

-Acceso: http://89.141.114.5:8181/pentaho