

# Business intelligence

---



Base de datos II

UNIVERSIDAD  
**SIGLO 21**

MIEMBRO DE LA RED  
**ILUMNO**



# Business intelligence

## Concepto de Business Intelligence

“Business Intelligence (BI) es un término genérico que incluye las aplicaciones, infraestructura y herramientas y las mejores prácticas que permiten el acceso y análisis de información para mejorar y optimizar las decisiones y el desempeño”

(Gartner, <http://www.gartner.com/it-glossary/business-intelligence-bi>, s.f.)

### Introducción

El objetivo de Business Intelligence (BI), también conocida como Inteligencia de Negocios en español, consiste en “apoyar de forma sostenible y continuada a las organizaciones para mejorar su competitividad, facilitando la información necesaria para la toma de decisiones” (Cano, 2007, p. 22).

Howard Dresner, de Gartner, la reconocida empresa de consultoría en tecnologías de la información fue quién utilizó por primera vez el concepto de BI en el año 1989.

Casualmente, la definición más actualizada y apropiada del concepto la podemos ubicar en el actual Glosario de Términos del Sitio Web de Gartner (<http://goo.gl/9bNho4>, s.f.): “Business Intelligence (BI) es un término genérico que incluye las aplicaciones, infraestructura y herramientas y las

mejores prácticas que permiten el acceso y análisis de información para mejorar y optimizar las decisiones y el desempeño”.

En tanto, Olivia Parr Rud (2009) considera a BI como un conjunto de teorías, metodologías, arquitecturas y tecnologías que transforman datos en información útil y significativa para los objetivos de negocios. Así, BI puede manejar grandes cantidades de datos desestructurados para ayudar a identificar, desarrollar y crear nuevas oportunidades. BI, en definitiva, permite interpretar los datos voluminosos de manera amigable, facilitando el aprovechamiento de potenciales oportunidades, implementando una estrategia efectiva que le provea una ventaja competitiva a la empresa.

Las tecnologías de BI proporcionan vistas históricas y predictivas de las operaciones comerciales de las compañías. Entre las funciones más comunes de las tecnologías de BI podemos encontrar: reporting, OLAP, minería de datos, entre otras.

Boris Evelson, de la consultora especializada Forrester Research, sostiene que:

BI es un conjunto de metodologías, procesos, arquitecturas y tecnologías que transforman los datos en información útil y significativa tal que permite una toma de decisiones estratégica, táctica y operativa más efectiva, con datos en tiempo real que le posibilitan a una empresa superar de manera eficaz a sus demás competidores en el mercado.

(<http://goo.gl/oleCU>, 2008)

Ralph Kimball (2013) define que un sistema de BI debe satisfacer varios requerimientos, entre los cuales podemos citar los siguientes:

- Debe hacer que la información sea fácilmente accesible.
- Debe presentar la información de manera consistente.
- Debe adaptarse al cambio.
- Debe presentar la información de manera oportuna.
- Debe ser un resguardo seguro que proteja los activos de información.
- Debe servir como el fundamento autorizado y confiable para la toma de decisiones.

- La comunidad de negocios debe aceptar al sistema de BI para que sea exitoso.

Por otro lado, Ralph Kimball (2013) sostiene que estos dos últimos requerimientos son los más críticos y frecuentemente los más desatendidos por las organizaciones a la hora de implementar un sistema de BI.

## Evolución de los SSD

### Historia

Si bien el término BI fue popularizado a partir de 1989, fue en el año 1958 cuando Hans Peter Luhn, investigador de IBM, definió a la inteligencia de negocios haciendo referencia a “la capacidad de comprender las interrelaciones de hechos presentados de una manera tal que guía la acción hacia una meta deseada” (BCW, <http://goo.gl/LEiSV0>, 2012).

Incluso las primeras empresas de software especializadas en lo que hoy denominamos BI surgen en la década de 1970, como por ejemplo: Information Builders (1975) o SAS (1976).

Sin embargo, el término BI, tal como lo conocemos hoy, ha evolucionado significativamente desde los originales Sistemas de Soporte de Decisión (DSS, *Decision Support Systems* en inglés).

William Inmon (2005) sostiene que en la década de 1980 aparecieron los Sistemas de Información Gerencial (MIS, *Management Information Systems* en inglés); posteriormente, ya más conocidos como DSS, los primeros MIS comenzaron a usarse como software para tomar decisiones. Anteriormente, tanto los datos como las tecnologías existentes se usaban exclusivamente para tomar decisiones operativas detalladas, y hasta entonces ningún motor de bases de datos podía llegar a servir para propósitos de procesamiento operacional/transaccional y analítico en forma simultánea.

Los software MIS, que generalmente contaban con una interfaz de usuario poco amigable, tampoco tenían grandes capacidades de modelado y estaban más bien orientados a la integración de información para la toma de decisiones.

Los Sistemas de Gestión de Bases de Datos (DBMS, *Database Management Systems* en inglés) que surgieron en la década de 1970, verdaderamente iniciaron una nueva era informática con el advenimiento del denominado **procesamiento transaccional en línea** (OLTP, *Online Transaction Processing*).

William Inmon (2005) alega que los primeros **Almacenes de Datos** (más conocidos como *Data Warehouses* en inglés) nacen en 1983.

Tras la implementación de masivos sistemas operacionales/transaccionales (OLTP, *OnLine Transaction Processing* en inglés), hacia el año 1985 surgieron los primeros sistemas de extracción de datos (ETL, *Extract, Transform and Load* en inglés), que posibilitaban transferir datos hacia otra base de datos o archivo. Estos programas de extracción comenzaron a ser muy populares en su época ya que permitían que los análisis de información para la toma de decisiones se ejecutaran sobre una base de datos distinta, diferente de la que usaban los sistemas transaccionales u operacionales. Naturalmente estos OLTP son fundamentales para la operación exitosa de cualquier negocio.

Los DSS nacieron también en la década de 1980. Estos sistemas fueron creados para asistir al proceso de toma decisiones y planificación; con mayores capacidades de modelado y mejor interfaz de usuario que los tradicionales MIS, pronto tuvieron un gran éxito y en pocos años evolucionaron en sistemas multi-usuario, es decir, permitieron la toma de decisiones en grupo.

Posteriormente, a comienzos de la década de 1990, nacen los Sistemas de Información Ejecutivos (EIS, *Executive Information Systems* en Inglés), software con capacidades adicionales para navegar sobre información más detallada, extraordinarias interfaces y, sobre todo, intensivos en el acceso a bases de datos integradas y unificadas, pero a la vez carentes, en muchos casos, de herramientas avanzadas de modelado. Information Builders en 1991 anuncia su primer software EIS, pero es en 1992 cuando la empresa MicroStrategy lanza al mercado el primer producto de software de BI completo.

En los '90 también surgieron las herramientas de Minería de Datos (*Data Mining*, en inglés) orientadas a la búsqueda de patrones y relaciones ocultas en los datos. Gracias a la implementación de enormes Almacenes de Datos (más conocidos como *Data Warehouses* en inglés) se podía obtener conocimiento de grandes volúmenes de datos.

William Inmon (2005) menciona que a partir de 1994 surgieron otros conceptos ligados a BI tales como las bases de datos multidimensionales

(OLAP), *Exploration Warehouses*, ODS, entre otros (conceptos que veremos más adelante en esta misma unidad).

En definitiva, si bien el concepto actual de BI fue propuesto en 1989 por Howard Dresner, fue recién a fines de la década de 1990 en que el software BI logra gran inserción en las organizaciones hasta la actualidad.

Además, cabe destacar la impresionante convergencia de tecnologías surgidas en las últimas dos décadas, tales como:

- La aparición de **Internet**, que posibilitó que el software BI estuviera disponible a través de la visualización de información en un navegador o browser.
- El concepto de **Software como Servicio** (*SaaS, Software as a Service* en Inglés) y **Computación en la Nube** (*Cloud Computing* en Inglés) que permitió disminuir drásticamente los costos de implementación de grandes sistemas de BI, al contratar por usuario un abono mensual.
- **Bases de Datos In-Memory y Técnicas de Lógica Asociativa** que le proveen a los usuarios de BI la posibilidad de analizar grandes cantidades de datos sin necesidad de generar costosas estructuras analíticas por separado (como herramientas OLAP u otras);
- El boom de la tecnología **Mobile** a través de Smartphones y Tablets, que logró que los usuarios BI pudieran trabajar muchas veces en forma remota accediendo a información crítica del negocio para la toma de decisiones en forma descentralizada;
- **Big Data** y la evolución hacia grandes volúmenes de datos desde algunos terabytes hasta varios petabytes en único repositorio;
- **Aplicaciones Analíticas**: dada la gran proliferación de herramientas de BI, algunas de ellas se focalizaron en especializarse en la gestión de métricas o Indicadores Claves de Negocio (*KPI, Key Performance Indicator* en Inglés) para determinadas industrias verticales; es decir, BI para el sector finanzas, BI para el sector salud, BI para el sector manufacturero, entre otras.

Evidentemente, esta convergencia tecnológica está generando una constante evolución del software BI para adaptarse a nuevos requerimientos de los usuarios, a nuevas tecnologías, a nuevas metodologías y a nuevas formas de acceder a los datos.

## Necesidades

Ralph Kimball (2013) considera que uno de los activos más importantes de cualquier organización es su información. Este activo es casi siempre usado para dos propósitos: la preservación de los registros operativos/transaccionales y la toma de decisiones analítica.

Así, mientras los sistemas operacionales/transaccionales se ocupan de incorporar datos a las bases de datos, los sistemas de BI, en cambio, permiten explotar o visualizar esos datos convirtiéndolos en información para la toma de decisiones.

Los usuarios de los sistemas operacionales/transaccionales son los que realmente mueven las ruedas de la organización (generando solicitudes, dando de alta clientes y monitoreando el estado de las actividades, entre otros). De esta manera, estos sistemas están optimizados para procesar las transacciones rápidamente.

Casi siempre los sistemas transaccionales trabajan con un registro o una transacción por vez en un momento dado. En definitiva, automatizan los procesos de negocio de la empresa. Pero dado su enfoque operativo, generalmente no guardan datos históricos con gran precisión, sino que, por el contrario, sus estructuras de datos frecuentemente solo guardan el estado actual.

En cambio, los usuarios de un sistema de BI se encargan de vigilar y controlar el movimiento de las ruedas de la organización, con el objeto de evaluar su desempeño. Así, cuentan la cantidad de solicitudes generadas y las comparan con las del mes o año anterior, etc. Se preocupan por que los procesos operativos estén trabajando correctamente.

Aunque necesitan datos detallados para soportar sus requerimientos, los usuarios BI casi nunca se enfocan en una transacción en un momento dado.

Los sistemas BI generalmente están optimizados para consultas de alto desempeño, teniendo en cuenta que las consultas de los usuarios frecuentemente requieren la agregación o sumarización de cientos, miles o incluso millones de registros diferentes. Por lo tanto, el contexto histórico (y principalmente el período o tiempo asociado) es vital para la evaluación del desempeño que hacen.

Ahora, ¿quién necesita de BI? En este sentido, Josep Lluís Cano (2007, p. 30) sostiene que “la información que podemos generar a partir de Business

“Intelligence es útil para todos los departamentos de nuestra organización”.  
Esto incluye a los responsables o gerentes de departamentos tales como:

- Compras
- Ventas / Comercial
- Finanzas
- Marketing
- Recursos Humanos
- Operaciones
- entre otros.

Es decir, para todas aquellas personas que deban tomar decisiones.

## Problemática

Ralph Kimball (2013) sostiene que los problemas típicos de los usuarios pueden sintetizarse en algunas de las demandas que mencionamos a continuación:

- Recolectamos toneladas de datos, pero no podemos acceder a ellos.
- Los directivos necesitan obtener los datos fácilmente.
- Sólo muéstrame lo que es importante.
- Desperdiciamos reuniones completas discutiendo quién tiene los números correctos, en lugar de tomar decisiones.
- Queremos que la gente use la información para mejorar su proceso de toma de decisiones.

Por otro lado, William Inmon (2005) expresa que la inexistencia de una solución BI lleva a los siguientes problemas:

- Falta de credibilidad en los datos.
- Problemas con la productividad .
- La Incapacidad de convertir datos en Información.

## Falta de credibilidad en los datos

Cuando no existe un único origen de información, puede suceder que en una misma organización encontramos dos departamentos, por ejemplo, uno de los cuales experimentó un incremento en las ventas del 10% y el otro un decremento del 5%. Conciliar la información de ambos puede ser difícil o incluso imposible. Cuando la gerencia recibe esta información contradictoria, no puede tomar decisiones razonablemente dada la falta de credibilidad en las fuentes. Las diferencias pueden surgir por varios factores:

- No existe una misma base (tiempo) de cálculo
- Algoritmos diferentes
- Niveles de extracción disímiles
- Datos externos tenidos (o no) en cuenta
- Distinta base de datos, fuente de la información.

## Problemas con la productividad

Del mismo modo que en el caso anterior, si no existe una base de datos única e integrada para el análisis de la información se produce una fuerte pérdida de productividad al tener que, por cada análisis requerido, revisar distintos archivos, bases de datos, etc., compilando información muchas veces contradictoria y sin una regla explícita de integración que puede depender de lo que hace cada analista en un momento dado. Hay una evidente sobrecarga de trabajo al tener que producir un reporte o informe manualmente.

## La incapacidad de convertir datos en información

Lo primero que descubrieron los analistas de los sistemas DSS a la hora de satisfacer la solicitud de información, es que ir a los sistemas transaccionales para obtener los datos necesarios era el peor escenario, puesto que estos sistemas y sus bases de datos subyacentes se construyeron sin tener en cuenta una futura integración de datos entre sí.

Otro obstáculo, es que no hay suficientes datos históricos almacenados en las aplicaciones. Es decir, son sistemas que fueron diseñados para responder rápidamente, con altos niveles de performance, y muchas veces esto se consigue guardando sólo los últimos doce meses de trabajo en la base de datos, sin grandes registros históricos.

## Beneficios

### Beneficios Tangibles, Intangibles y Estratégicos

De acuerdo con Josep Lluís Cano (2007, p. 32), “uno de los objetivos básicos de los sistemas de información es que nos ayuden a la toma de decisiones. Cuando un responsable tiene que tomar una decisión pide o busca información, que le servirá para reducir la incertidumbre”.

Por lo tanto, un sistema de información de BI es clave para transformar los datos en información y la información en conocimiento.

Actualmente, las empresas, ante una dinámica de cambios casi permanente en los mercados, están sometidas a fuertes presiones competitivas y sus directivos requieren del conocimiento que puede brindar un sistema de BI para poder soportar un adecuado proceso de toma de decisiones.

Los beneficios que se pueden obtener a través del uso de BI pueden ser de distintos tipos:

- **Beneficios tangibles**, por ejemplo: reducción de costos, generación de ingresos y reducción de tiempos para las distintas actividades del negocio.

- **Beneficios intangibles:** el hecho de que tengamos disponible la información para la toma de decisiones hará que más usuarios utilicen dicha información para tomar decisiones y mejorar la nuestra posición competitiva.
- **Beneficios estratégicos:** Todos aquellos que nos facilitan la formulación de la estrategia, es decir, a qué clientes, mercados o con qué productos dirigirnos.

(Cano, 2007, pp. 32-33)

Este mismo autor también sostiene que (2007, p. 37) “la principal razón de un proyecto de Business Intelligence es el análisis de un problema o problemas interrelacionados”.

## Cálculo del ROI en Proyectos de BI: Porqué es importante calcular el ROI.

De acuerdo con Josep Lluís Cano (2007), al plantear cualquier proyecto de sistema de información es imprescindible calcular cuál es la rentabilidad esperada del mismo.

Por lo tanto es necesario:

- Definir el **valor esperado**, es decir, tratar de estimar cuál es el beneficio o valor agregado total que aportará el proyecto a la empresa.
- Determinar la **inversión total** requerida del proyecto con el objetivo de asegurar los fondos necesarios para su concreción, identificando los distintos retornos que se podrán conseguir.
- Implementar el proyecto y **evaluar si** se ha logrado alcanzar el valor y retorno esperados.
- **Medir** los resultados del proyecto e implementar un plan de acción correctivo alternativo.

La medida comúnmente utilizada en el entorno empresarial para comprobar la rentabilidad de un proyecto es el **retorno de la**

**Inversión (ROI).** El ROI pone en relación el valor aportado al negocio con las inversiones necesarias para obtenerlo. Una forma simplificada del cálculo del ROI es:

$$ROI = \frac{\text{Valor para el Negocio}}{\text{Costo del Proyecto}}$$

(Cano, 2007, p. 45)

## Metodología para el Cálculo del ROI

Josep Lluís Cano, en base a un artículo de Bill Whittemore, propone la siguiente **metodología paso a paso para el cálculo del ROI de un proyecto de BI**:

- 1) Definir cuál es el problema u oportunidad de negocio y los objetivos del mismo. Los objetivos deben ser específicos, medibles, alcanzables, adecuados y referidos a un periodo de tiempo.
- 2) Recoger los requerimientos de negocio.
- 3) Construir el proyecto de Business Intelligence.
- 4) Identificar y cuantificar los beneficios (tangibles, estratégicos e intangibles).
- 5) Establecer el punto de partida de medida, tanto de los costos como de los ingresos.
- 6) Calcular el costo total de propiedad (TCO): incluye el hardware, software, los servicios de consultoría, los costos de los recursos internos (costos de personal) y los costos de lanzamiento, mantenimiento y formación.
- 7) Calcular el ROI. Para ello se aplica la siguiente fórmula:

$$ROI = \frac{NPV}{\text{Inversión inicial}} \times 100$$

En donde **NPV** es el **Valor Neto Actual**, es decir, la **suma actualizada de los beneficios esperados del proyecto**.

8) Una vez aprobado e implementado el proyecto, deberemos hacer un seguimiento tanto de la inversión como de los costos y de los beneficios que realmente se han conseguido, para poder tomar las medidas correctivas que sean necesarias.

(2007, pp. 47-49)

Además, de acuerdo con Josep Lluís Cano (2007, p. 52), "los proyectos de Business Intelligence tienen un ROI elevado, su comportamiento es mucho mejor que en el resto de proyectos de Sistemas de Información".

## Proceso de toma de decisiones

Podemos denominar "**Proceso de toma de decisiones**" al proceso de tener que elegir entre distintos cursos de acción alternativos con el propósito de alcanzar un objetivo en particular.

De acuerdo con Kennet Laudon (2004), este proceso consta de cuatro etapas:

- Inteligencia
- Diseño
- Selección
- Implementación

A continuación veremos cada una de estas fases:

### Fase de Inteligencia

La etapa de **Inteligencia** consiste en:

Identificar y entender los problemas que se presentan en la organización: el por qué ocurre un problema, dónde y cuáles son sus efectos.

Los sistemas MIS tradicionales que suministran una gran variedad de información detallada pueden ayudar a identificar problemas, especialmente si los sistemas reportan excepciones.

(Kennet Laudon, 2004, p. 88)

### Fase de Diseño

La etapa de **Diseño**, según Kennet Laudon es aquella durante la cual “el individuo genera posibles soluciones para los problemas. Los DSS más pequeños son ideales en esta etapa de la toma de decisiones, porque trabajan sobre modelos sencillos, es posible desarrollarlos rápidamente y pueden operar con datos limitados” (2004, p. 88).

### Fase de Selección

La etapa de **Selección**:

Consiste en elegir entre alternativas de solución. Aquí el encargado de la toma de decisiones podría necesitar un sistema DSS más grande para obtener datos más extensos a partir de varias alternativas y modelos complejos, o bien, herramientas de análisis de datos para recabar todos los costos, consecuencias y oportunidades.

(Kennet Laudon, 2004, p. 88)

## Fase de Implementación

La etapa de **Implementación** se lleva a cabo:

Cuando la decisión se pone en práctica. En ella, los gerentes pueden utilizar un sistema que elabore informes de rutina sobre el progreso de una solución específica.

Los sistemas de apoyo pueden abarcar desde sistemas MIS con características avanzadas hasta sistemas mucho más pequeños, así como software de planeación de proyectos que se ejecute en computadoras personales.

(Kennet Laudon, 2004, p. 88)

Tal como destaca Mónica López Gutiérrez (<http://goo.gl/Xy2REI>, 2004), debemos tener en cuenta que un DSS no es más que un sistema de información gerencial que permite “resolver problemas semi-estructurados y no estructurados, involucrando al usuario a través de una interfaz amigable”.

Mónica López Gutiérrez también sostiene que un DSS permite “mejorar el Proceso de Toma de Decisiones a lo largo de las etapas del mismo: inteligencia, diseño, selección e implementación... Los DSS principalmente se utilizan para decisiones estratégicas y tácticas en la gestión a nivel superior” (<http://goo.gl/Xy2REI>, 2004).



## Bibliografía de referencias

- **Harris, H.** (2014, 30 de Abril). *The History of Business Intelligence* [post de la web]. Recuperado de: <http://www.businesscomputingworld.co.uk/the-history-of-business-intelligence-infographic/>
- **Cano J. L.** (2007). *Business Intelligence: Competir con Información*. España: Banesto Fundación Cultural y ESADE.
- **Evelson, B.** (2008). *Topic Overview: Business Intelligence*. Recuperado el 30 de Abril de 2014: <http://www.forrester.com/Topic+Overview+Business+Intelligence-/E-RES39218?objectid=RES39218>
- **Gartner (s.f.)**. *IT Glossary - Business Intelligence*. Recuperado el 30 de Abril de 2014: <http://www.gartner.com/it-glossary/business-intelligence-bi>
- **Inmon William** (2005). *Building the Data Warehouse* (4º Edición). Estados Unidos: Wiley Publishing.
- **Kimball Ralph, R. M.** (2013). *The Data Warehouse Toolkit* (3º Edición). Estados Unidos: Wiley Publishing.
- **Laudon Kenneth, L. J. P.** (2004). *Sistemas de Información Gerencial: administración de la empresa digital* (8º Edición). México: Pearson Educación.
- **López Gutiérrez, M.** (2004, 30 de Abril). El lugar de los DSS en el proceso de toma de decisión. *GestioPolis* [post de la web]. Recuperado de 2014: <http://goo.gl/Xy2REI>
- **Parr Rud, O.** (2009). *Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy*. Hoboken, New Jersey: John Wiley & Sons.

# CIF

---



Base de datos II

UNIVERSIDAD  
**SIGLO 21**

MIEMBRO DE LA RED  
**ILUMNO**

# » CIF

## Niveles de Datos

De acuerdo con William Inmon (2005), podemos diferenciar cuatro niveles de datos:

- Nivel Operacional
- Nivel Atómico (o *Data Warehouse*)
- Nivel Departamental (o *Data Mart* u OLAP o Multidimensional)
- Nivel Individual

Estos niveles son la base del concepto del **Corporate Information Factory (CIF)**. A continuación, describiremos las principales características de cada uno de estos niveles de datos:

- **Nivel Operacional**

El **nivel operacional** de datos mantiene solamente datos primitivos (es decir, operacionales, transaccionales) orientados a la aplicación, y sirve específicamente a la comunidad de usuarios de procesamiento OLTP, de alta performance requerida.

En este nivel generalmente no se mantienen registros históricos ya que los datos se actualizan y se pierden los valores anteriores.

- **Nivel del Data Warehouse**

El **nivel de datos atómico** o de *Data Warehouse* mantiene datos primitivos, pero históricos e integrados que no pueden actualizarse, además de algunos datos derivados/analíticos.

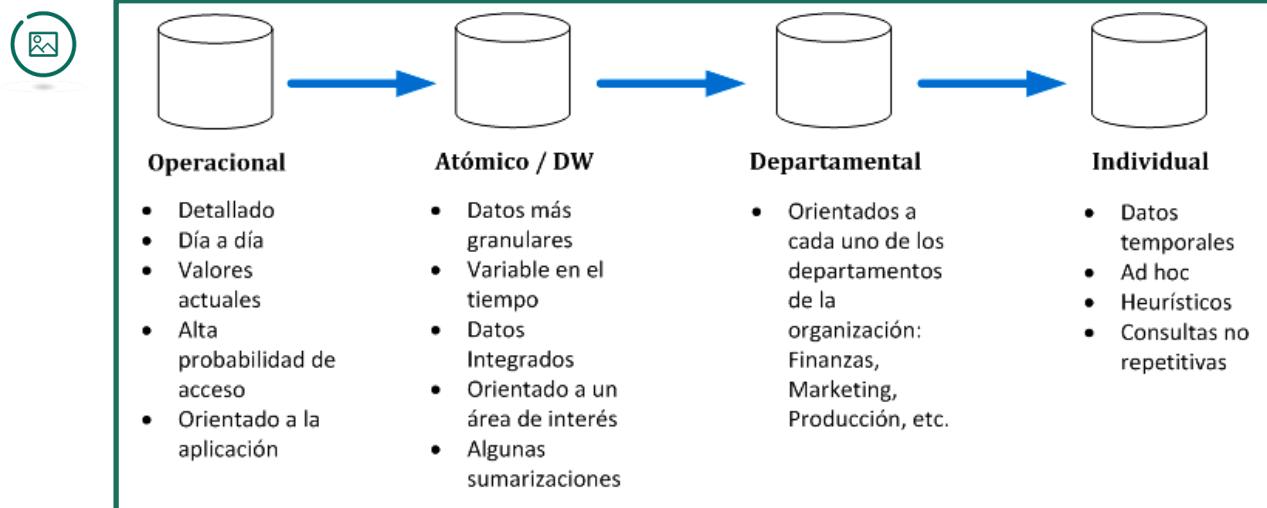
- **Nivel Departamental**

El **nivel de datos Departamental** contiene datos derivados que son formateados especialmente en función de los requerimientos de análisis de datos de los usuarios, ajustados a las necesidades de un departamento de la organización en particular.

- **Nivel Individual**

El **nivel de datos Individual** es donde mayor análisis heurístico se realiza. Generalmente, aquí hay datos temporales. Como regla general, típicamente los Sistemas de Información Ejecutivos (EIS) procesan datos a este nivel y corren sobre una PC.

Imagen 1: Niveles de Datos



Fuente: Elaboración propia en base a William Inmon (2005, p. 16)

## Componentes del CIF

Resulta evidente que los tipos de consultas que se pueden efectuar en cada nivel de datos son distintos. Así, los diferentes niveles de datos requieren un conjunto de entidades arquitectónicas diferenciadas. Estas entidades constituyen el **Corporate Information Factory (CIF)**.

El CIF es un concepto que hace referencia al conjunto de todas las estructuras de datos que posee una determinada organización (bases de datos transaccionales y analíticas, entre otras), conjuntamente con todos aquellos procesos y herramientas utilizados en las distintas etapas para poder generar información y lograr que la misma se encuentre disponible en tiempo y forma para todos los niveles de la organización.

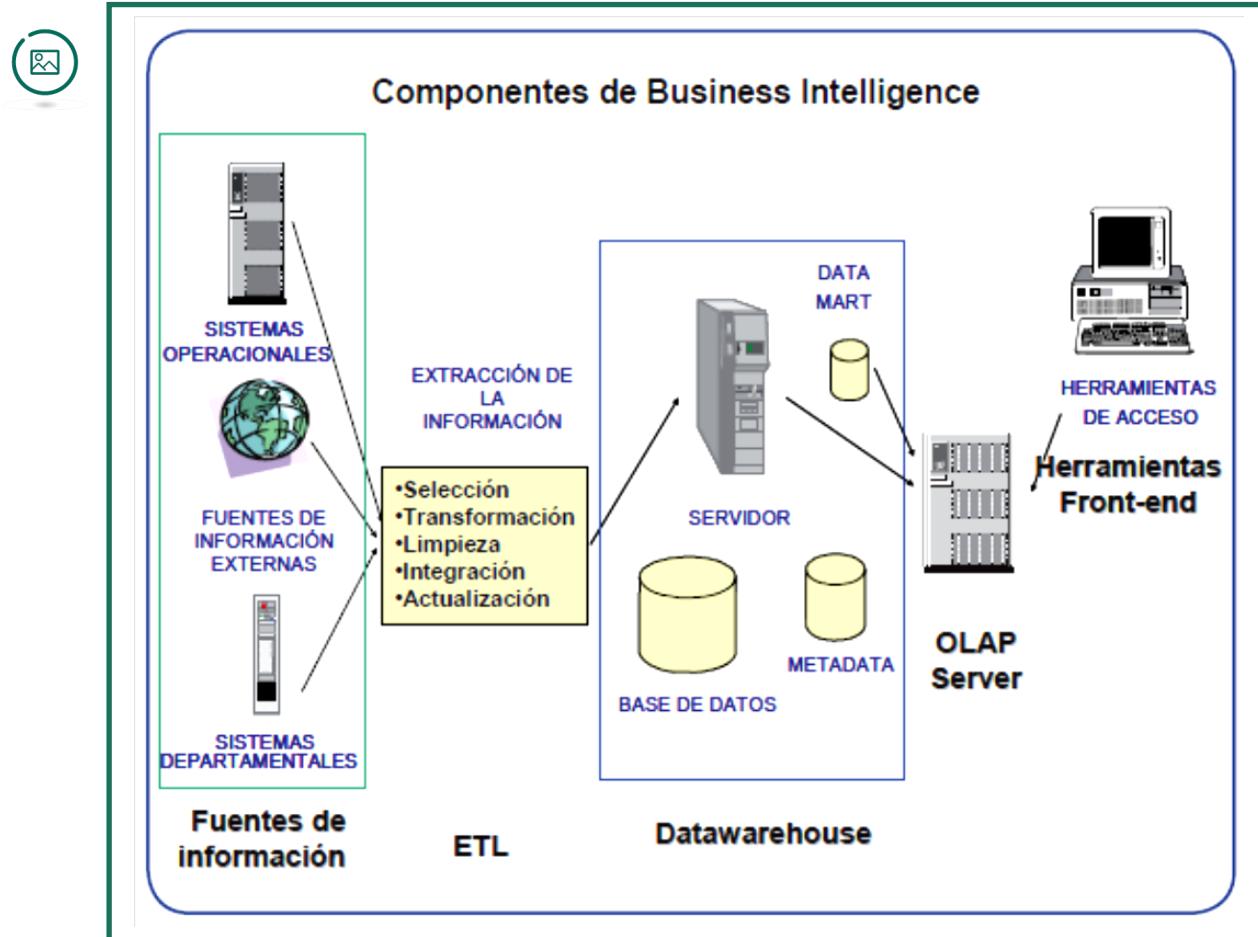
Como señalan Josep Lluís Cano (2007), William Inmon (2005) y Ralph Kimball (2013), los distintos componentes del CIF y, por ende, de una solución de BI, son (aunque no siempre encontraremos todos ellos):

- Fuentes de Información
- Procesos ETL de Extracción, Transformación y Carga de datos en el *Data Warehouse*
- *Data Warehouse*
- *Data Mart*

- Repositorio de Metadatos
- *Operational Data Store (ODS)*
- *Exploration Warehouse*
- *Data Mining Warehouse*
- Herramientas de *Business Intelligence* para la visualización y explotación de datos.

En el siguiente gráfico podemos ver estos componentes del CIF y las relaciones entre ellos:

**Imagen 2: Componentes del CIF**



Fuente: Josep Lluís Cano (2007, p. 93)

## Fuentes de Información

Según Josep Lluís Cano (2007), las fuentes de información hacen referencia a los lugares (bases de datos, etc.) de donde se extraerán los datos para alimentar el *Data Warehouse*.

### Principales Fuentes de Datos

Las principales fuentes de información son:

- **Bases de Datos de cada uno de los Sistemas Operacionales/Transaccionales**, que pueden incluir tanto aplicaciones desarrolladas a la medida de la organización, como así también productos “enlatados” o grandes sistemas corporativos, como por ejemplo: los sistemas de Gestión de las Relaciones con los Clientes (**CRM, Customer Relationship Management**), sistemas de Gestión Financiera y Planificación de Recursos Empresariales (**ERP, Enterprise Resource Planning**), sistemas de Gestión de la Cadena de Abastecimiento (**Supply Chain Management**) y sistemas de Gestión de Recursos Humanos o del Capital Humano (**HCM, Human Capital Management**), entre otros.
- **Sistemas de Información Departamentales**, muchas veces basados en planillas de cálculo Excel, etc.
- **Fuentes de información correspondientes a Bases de Datos externas a la empresa**, que en algunos casos pueden haber sido compradas a terceras organizaciones, como por ejemplo: consultoras de investigación de mercado. Las fuentes de información externas son fundamentales para enriquecer la información que tenemos de los clientes. En algunos casos es interesante incorporar información referente, por ejemplo, a población y número de habitantes, entre otros. Podemos acceder a información de este tipo en la correspondiente web del Instituto Nacional de Estadísticas.

Pero también existen otras fuentes de información de las cuales se pueden llegar a tomar datos según las necesidades, tales como:

- **Otras fuentes de Internet:** Muchas veces es necesario poder realizar comparaciones en los indicadores con otras compañías u organizaciones; es lo que se denomina una actividad de *Benchmarking*. Esto es fundamental dado que para saber si determinado valor es bueno o malo - en realidad se trata de un valor relativo a cómo les va a otras empresas- pueden obtenerse datos públicos de Internet e incorporarlos al *Data Warehouse*.
- **Datos aportados por analistas expertos**, especializados en alguna temática en particular.
- Otros.

## Factores a considerar

Según Josep Lluís Cano (2007), hay distintos factores que contribuyen a la complejidad de la carga de información; entre ellos el autor destaca la cantidad de fuentes diferentes de información, teniendo en cuenta que en las grandes corporaciones es natural hablar de una media de 8 bases de datos hasta llegar incluso a 50.

A medida que se requiere acceder a un número creciente de fuentes de información, la complejidad de todo proyecto de creación de un *Data Warehouse* se incrementa notablemente dado que probablemente algunas bases de datos están en SQL Server, otras en Oracle, otras en IBM DB2, etc.; además de la propia complicación inherente al modelo de datos subyacente de cada aplicación.

Otro problema es el que deviene de la falta de documentación de estas bases de datos, correspondientes en muchas ocasiones a aplicaciones que han sido modificadas a lo largo de los años por distintos programadores sin seguir ningún tipo de estándares y con criterios sumamente heterogéneos.

La información que cargamos en un *Data Warehouse* normalmente es estructurada, es decir, aquella que se puede almacenar en tablas; en la mayoría de los casos es información numérica. Cada vez más, la tecnología nos permite trabajar con información no estructurada, y se espera que este tipo de información sea cada vez más importante. Dentro de la

información no estructurada tenemos: correos electrónicos, cartas, informes, videos, etc.

(Cano, 2007, pp. 96-97)

## Calidad de Datos

De acuerdo con Josep Lluís Cano (2007), una vez que ya se han establecido cuáles serán las fuentes de información debe procederse a verificar la calidad de los datos del *Data Warehouse*, lo cual es un aspecto esencial.

Consecuentemente, es necesario asegurar que la calidad de los datos es máxima. Si en el Data Warehouse hay errores, éstos se propagarán a lo largo de toda la organización y son muy difíciles de localizar. Además, pueden ocasionar que se tomen decisiones erróneas que afecten a los resultados de la organización. Los costes derivados de que la calidad de los datos no sea la correcta pueden llegar a ser muy elevados.

(Cano, 2007, p. 98)

Por otro lado, el autor también expresa que:

Asumir que la calidad de los datos es buena puede ser un error fatal en los proyectos de Business Intelligence. Normalmente, cuando se construye un Data Warehouse la mayoría de las organizaciones se focalizan en identificar los datos que necesitan analizar, los extraen y los cargan en el Data Warehouse. Generalmente no se piensa en la calidad de los datos, permitiendo que los errores sean cargados al Data Warehouse. Debería por tanto establecerse un control o conjunto de controles en el proyecto que localizara los errores en los datos y no permitiera la carga de los mismos. Las comprobaciones se deberán llevar a cabo, de forma manual o automatizada, teniendo en cuenta distintos niveles de detalle y variando los períodos de tiempo, comprobando que los datos cargados coinciden con los de las fuentes de datos origen.

En algunos casos se detectan errores que se originan por fallos en los sistemas transaccionales, lo que debería provocar proyectos

de mejora en los mismos. Muchos de estos casos se deben a que los usuarios pueden introducir datos sin ningún tipo de control. Siempre que se pueda, es recomendable que los usuarios elijan entre distintos valores, en lugar de introducirlos libremente ellos. No es una buena opción corregirlos en el proceso ETL y no modificar las aplicaciones origen. Esta alternativa es mucho más rápida inicialmente, pero mucho más costosa a largo plazo.

Los errores también se pueden producir, por ejemplo, en el proceso de ETL o al integrarlos en el Data Warehouse.

(Cano, 2007, p. 99)

Respecto de la responsabilidad de la calidad de los datos, Josep Lluís Cano determina que la misma:

No pertenece sólo a los departamentos de tecnología: Debe asumirse la parte correspondiente en cada uno de los propietarios de los procesos y de las aplicaciones que los soportan. Desde el proyecto debemos velar por la calidad de los datos, puesto que si la calidad no es la adecuada nunca podremos obtener los beneficios esperados del proyecto. Debemos entender que la problemática de la calidad de datos no es un problema de los departamentos de tecnología, sino un problema estratégico al que debemos asignar objetivos, recursos y planificación.

(2007, p. 100)

Finalmente, las principales características que deben satisfacer los datos, con el objeto de que alcancen un nivel de calidad elevado son:

- 1) Precisión:** ¿Representan los datos con precisión una realidad o una fuente de datos que se pueda verificar?
- 2) Integridad:** ¿Se mantienen constantemente la estructura de los datos y las relaciones a través de las entidades y los atributos?
- 3) Coherencia:** ¿Son los elementos de datos constantemente definidos y comprendidos?
- 4) Totalidad:** ¿Están todos los datos necesarios?

**5) Validez:** ¿Son los valores aceptables en los rangos definidos por el negocio?

**6) Disponibilidad:** ¿Están los datos disponibles cuando se necesitan?

**7) Accesibilidad:** ¿Se puede acceder a los datos fácil y comprensiblemente?

(Cano, 2007, p. 102)

## Procesos de Extracción, Transformación y Carga

### Concepto de ETL

El ETL se trata del proceso correspondiente a:

La extracción, transformación y carga de los datos en el Data Warehouse. Antes de almacenar los datos en un Data Warehouse, éstos deben ser transformados, limpiados, filtrados y redefinidos. Normalmente, la información que tenemos en los sistemas transaccionales no está preparada para la toma de decisiones.

(Cano, 2007, pp. 93-94)

Es decir, es el proceso que consiste en extraer información de las fuentes de datos, transformarla, re-codificarla, limpiarla, explicitar reglas de negocio ocultas, formatearla y organizarla de manera de poder incorporarla en el Data Warehouse. A estos procesos se los conoce con la sigla ETL (*Extract, Transform and Load*, en inglés).

De acuerdo con William Inmon (2005), el proceso ETL se encarga de transferir datos desde el entorno operacional al entorno del Data Warehouse, integrando las distintas fuentes de información.

## Proceso ETL en el Proyecto de BI

El diseño de los procesos ETL insume gran parte del tiempo de un proyecto BI. Se trata de una tarea compleja, pero para poder aprovechar los beneficios del *Data Warehouse* es fundamental que se logre la integración de datos desde los distintos orígenes.

Josep Lluís Cano señala que “el proceso de ETL consume entre el 60% y el 80% del tiempo de un proyecto de Business Intelligence, por lo que es un proceso clave en la vida de todo proyecto” (2007, p. 103).

### Pasos del Proceso ETL

El proceso ETL se divide en 5 subprocesos:

- 1) Extracción:** Este proceso recupera los datos físicamente de las distintas fuentes de información. En este momento disponemos de los datos en bruto.
- 2) Limpieza:** Este proceso recupera los datos en bruto y comprueba su calidad, elimina los duplicados y, cuando es posible, corrige los valores erróneos y completa los valores vacíos, es decir se transforman los datos -siempre que sea posible- para reducir los errores de carga. En este momento disponemos de datos limpios y de alta calidad.
- 3) Transformación:** Este proceso recupera los datos limpios y de alta calidad y los estructura y sumariza en los distintos modelos de análisis. El resultado de este proceso es la obtención de datos limpios, consistentes, sumarizados y útiles.
- 4) Integración:** Este proceso valida que los datos que cargamos en el Data Warehouse son consistentes con las definiciones y formatos del Data Warehouse; los integra en los distintos modelos de las distintas áreas de negocio que hemos definido en el mismo. Estos procesos pueden ser complejos.
- 5) Actualización:** Este proceso es el que nos permite añadir los nuevos datos al Data Warehouse.

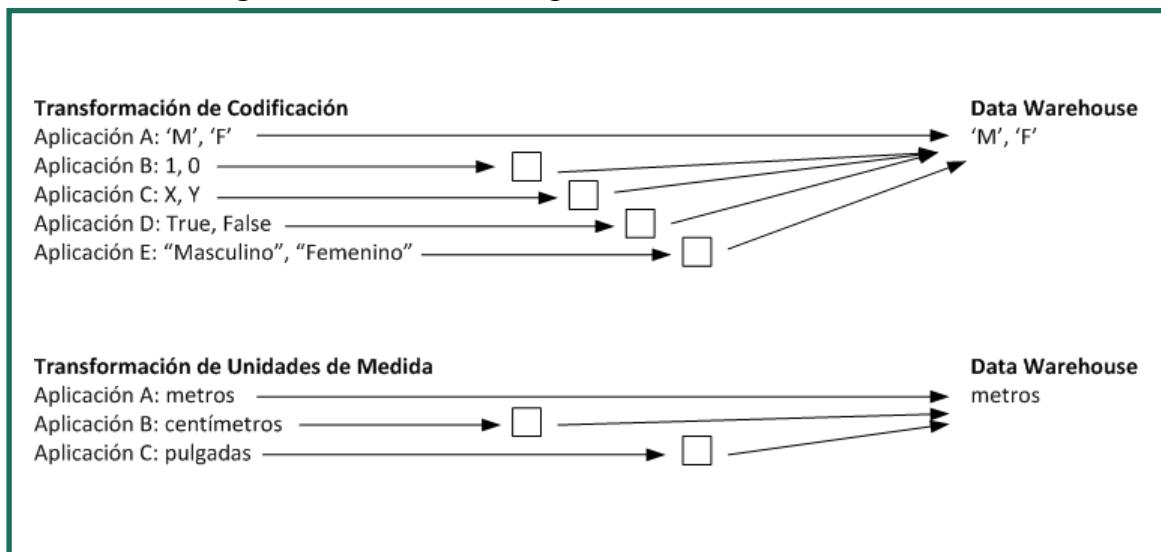
(Cano, 2007, pp. 104-105)

## Problema de Integración de Datos

De acuerdo con William Inmon (2005), se extraen datos desde distintas fuentes/sistemas/aplicaciones, datos que no están integrados. Incluirlos en el *Data Warehouse* sin integrarlos sería un gran error.

Ocurre que cuando fueron diseñados esos sistemas nadie pensó en futuras integraciones. Cada una tuvo sus propios requerimientos. Por lo tanto, extraer datos de distintos lugares e integrarlos en una base de datos única es un problema complejo.

**Imagen 3:** Problema de Integración de Datos



Fuente: Elaboración propia en base a William Inmon (2005, p. 73)

Algunos ejemplos de falta de integración que comúnmente podemos encontrar son:

- Valores que se encuentran codificados de manera distinta
- Valores con distintas unidades de medida
- Campos que tienen distintos nombres pero representan lo mismo
- Formatos diferentes.

## Eficiencia en el acceso a los Sistemas Transaccionales

William Inmon (2005) destaca, sin embargo, que la integración no es la única dificultad en la transformación de datos desde los sistemas transaccionales; otro problema es la eficiencia en el acceso a los datos de los sistemas existentes. No tiene sentido cargar todos los datos de los sistemas operacionales en el *Data Warehouse* cada vez que haya una extracción.

Podemos identificar entonces tres tipos de carga en el *Data Warehouse*:

- Datos antiguos, archivados (carga única, de una sola vez).
- Datos actualmente contenidos en el entorno operacional.
- Cambios continuos al *Data Warehouse* a partir de cambios (actualizaciones) que ocurrieron en los sistemas operacionales desde la última actualización. (Esto presenta el mayor desafío para un arquitecto de datos puesto que no es fácil identificar dichos cambios).

Según William Inmon (2005), existen cinco técnicas que comúnmente se usan para limitar la cantidad de datos operacionales extraídos:

- Extraer los datos que tienen marca temporal en las bases de datos operacionales. Sólo se traen los datos desde la última fecha y hora de actualización.
- A través de un archivo delta que contenga solo los cambios hechos en una aplicación como resultado de las transacciones ejecutadas sobre las bases de datos operacionales. El proceso es muy eficiente, pero muy pocas aplicaciones cuentan con un archivo delta.
- Escanear un archivo log (similar al archivo delta, pero con mayor información).
- Modificar el código de la aplicación.
- Tomar una imagen o snapshot de la base de datos operacional, antes y después de cada extracción; luego se comparan ambas versiones para identificar las diferencias.

## Desafíos en los Procesos ETL

De acuerdo con William Inmon (2005), los procesos ETL encaran varios desafíos complejos:

- La extracción de datos desde los sistemas transaccionales hacia el *Data Warehouse* generalmente requiere un cambio de tecnología (por ejemplo: leer desde una base de datos Microsoft SQL Server y cargar un *Data Warehouse* en Oracle).
- La selección de datos a extraer puede ser muy compleja, tal como se describió anteriormente.
- Los campos claves en las bases de datos operacionales frecuentemente deben ser reestructurados y convertidos antes de escribirlos en el *Data Warehouse*.
- Reformateo de campos no claves como, por ejemplo, el formato de las fechas.
- Limpieza de datos (formato, verificación de clave foránea, etc.), a medida que se introduce en el *Data Warehouse*.
- Consolidación de datos de distintas fuentes.
- Provisión de valores por default.
- Sumarización de datos.
- Renombre de datos.
- Conversiones en los formatos de datos.
- Registro de cantidad de datos extraídos, transformados y cargados en el *Data Warehouse* (tema clave de auditoría/metadatos cuando se cargan grandes volúmenes de datos).
- Otros.

## Enterprise Data Warehouse

El *Data Warehouse* es el núcleo del *Corporate Information Factory*; contiene datos históricos, integrados, de toda la organización, generalmente con alto nivel de detalle (granularidad), actualizado a partir de los procesos ETL.

Además, es una colección de información corporativa que alimenta a otras estructuras de datos como *Data Marts*, *Exploration Warehouses*, *Data Mining Warehouses*, *Operational Data Stores (ODS)* y, en definitiva, construidos con el objetivo de ser explotados por diferentes Sistemas de Soporte de Decisión (DSS).

De acuerdo con William Inmon (2005), el *Data Warehouse* está en el corazón del CIF y es el fundamento de todo DSS. El *Data Warehouse* contiene datos corporativos con alta granularidad (alto nivel de detalle).

Por diversas razones (implicancias en performance, integración y transformación de datos, entre otras), es conveniente que los sistemas DSS realicen las consultas analíticas sobre un *Data Warehouse* aislado de las bases de datos operacionales.

La aparición de los *Data Warehouse* o Almacenes de Datos son la respuesta a las necesidades de los usuarios que necesitan información consistente, integrada, histórica y preparada para ser analizada para poder tomar decisiones.

Al recuperar la información de los distintos sistemas, tanto transaccionales como departamentales o externos, y almacenándolos en un entorno integrado de información diseñado por los usuarios, el *Data Warehouse* nos permitirá analizar la información contextualmente y relacionada dentro de la organización.

(Cano, 2007, p. 113)

Según William Inmon (2005), los aspectos más importantes de un *Data Warehouse* son:

- **Orientado a una materia o área de interés:** Los sistemas transaccionales se organizan alrededor de las aplicaciones funcionales de la organización; por ejemplo, para una compañía de seguros las aplicaciones pueden ser para el procesamiento de

seguros de autos, vida, casa, salud, etc. Las principales áreas de interés podrían ser: cliente, póliza, etc. Cada tipo de empresa tiene sus propias áreas de interés. A su vez, cada área de interés está físicamente implementada como una serie de tablas relacionadas en el *Data Warehouse*. Por ejemplo, para el área “Clientes” podría haber una tabla con todos los clientes de la empresa y varias tablas con la actividad de estos clientes, algunas con los eventos detallados y otras con sumarizaciones o agregaciones (como cantidad de transacciones realizadas por mes, promedio de accidentes por año, etc.); todas las tablas estarán relacionadas por el ID del Cliente.

- **Integrado:** De todos los aspectos del *Data Warehouse*, este es el más importante. Los datos provienen de múltiples orígenes. Antes de insertarse, se convierten, reformatan, e incluso pueden sumarizarse. Dado que cada aplicación en su momento fue desarrollada en forma aislada, sin tener en cuenta futuras integraciones, generalmente se encuentran inconsistencias de datos entre los distintos orígenes o bases de datos, ya sea a nivel de nomenclatura, codificación, atributos físicos y unidad de medida, entre otros.
- **No volátil:** los datos operacionales son generalmente accedidos y manipulados de a un registro por vez. Los datos operacionales son actualizados regularmente, no así en el *Data Warehouse* donde los datos se cargan (usualmente en forma masiva, en un formato estático o “snapshot”) y se consultan, pero no se actualizan. Cuando hay un cambio en los sistemas transaccionales, en el *Data Warehouse* se graba un nuevo registro o *snapshot*. Esto permite tener un registro histórico de datos.
- **Variable con el tiempo:** Esto implica que cada unidad de datos en el *Data Warehouse* es precisa y en un momento dado. Generalmente, cada registro está asociado a una fecha de transacción en particular o a una marca de tiempo. Habitualmente, el *Data Warehouse* tiene un horizonte de tiempo almacenado de 5 a 10 años, aunque a veces puede extenderse mucho más. En cambio, las bases de datos transaccionales contienen datos de valor actual, es decir, datos cuya precisión solo es válida en el momento en que se consulta o se accede. Por ejemplo: un Banco sabe exactamente cuánto dinero tiene un cliente en su cuenta hoy; este valor se actualiza cada vez que hay un depósito o una extracción.

En el *Data Warehouse* la serie de *snapshots* provee una secuencia histórica de actividades y eventos; la estructura clave de los datos operacionales puede o no contener un elemento de tiempo como

año, mes o día, pero el *Data Warehouse* siempre contiene algún elemento de tiempo.

## Data Mart

De acuerdo con William Inmon (2005), un *Data Mart* es una estructura de datos que está dedicada a servir las necesidades analíticas de un grupo de personas, como el departamento de finanzas, por ejemplo.

El trabajo de construir un Data Warehouse corporativo puede generar inflexibilidades, o ser costoso y requerir plazos de tiempo que las organizaciones no están dispuestos a aceptar. En parte, estas razones originaron la aparición de los Data Mart. Los Data Mart están dirigidos a una comunidad de usuarios dentro de la organización, que puede estar formada por los miembros de un departamento, o por los usuarios de un determinado nivel organizativo, o por un grupo de trabajo multidisciplinar con objetivos comunes.

Los Data Mart almacenan información de un número limitado de áreas; por ejemplo, pueden ser de marketing y ventas o de producción. Normalmente se definen para responder a usos muy concretos.

Normalmente, los Data Mart son más pequeños que los Data Warehouses. Tienen menos cantidad de información, menos modelos de negocio y son utilizados por un número inferior de usuarios.

Los Data Mart pueden ser independientes o dependientes. Los primeros son alimentados directamente de los orígenes de información, mientras que los segundos se alimentan desde el Data Warehouse corporativo. Los Data Mart independientes pueden perpetuar el problema de los “silos de información” y en su evolución pueden llegar a generar inconsistencias con otros Data Mart.

(Cano, 2007, pp. 117-118)

## Repository de Metadatos

De acuerdo con William Inmon (2005), un componente central del *Data Warehouse* es el *Repository de Metadatos* ya que le permite al usuario final de un DSS navegar entre distintas posibilidades. Cuando el usuario se acerca a un *Data Warehouse* sin metadatos no sabe dónde iniciar su análisis o cómo encontrar los datos correctos, ni cómo interpretar los datos encontrados. Con la ayuda de metadatos el usuario puede ir rápidamente a los datos principales. Así, los metadatos actúan como un índice de los contenidos del *Data Warehouse*.

Generalmente, almacenan el seguimiento de:

- Estructuras de datos importantes para un desarrollador
- Estructuras de datos importantes para un analista de BI
- Fuentes de datos que alimentan el *Data Warehouse*
- Transformación de datos en el *Data Warehouse*
- Modelo de datos
- Relaciones entre el modelo de datos y el *Data Warehouse*
- Historia de extracciones (ejecuciones de los procesos ETL).

Josep Lluís Cano expresa que:

Un componente crítico de un Data Warehouse es el Metadata. El Metadata es el repositorio central de información de la información. Nos da el significado de cada uno de los componentes y sus atributos que residen en el Data Warehouse (o Data Mart). La información que contiene el Metadata es útil para los departamentos de tecnología y los propios usuarios. Puede incluir definiciones de negocio, descripciones detalladas de los tipos de datos, formatos y otras características.

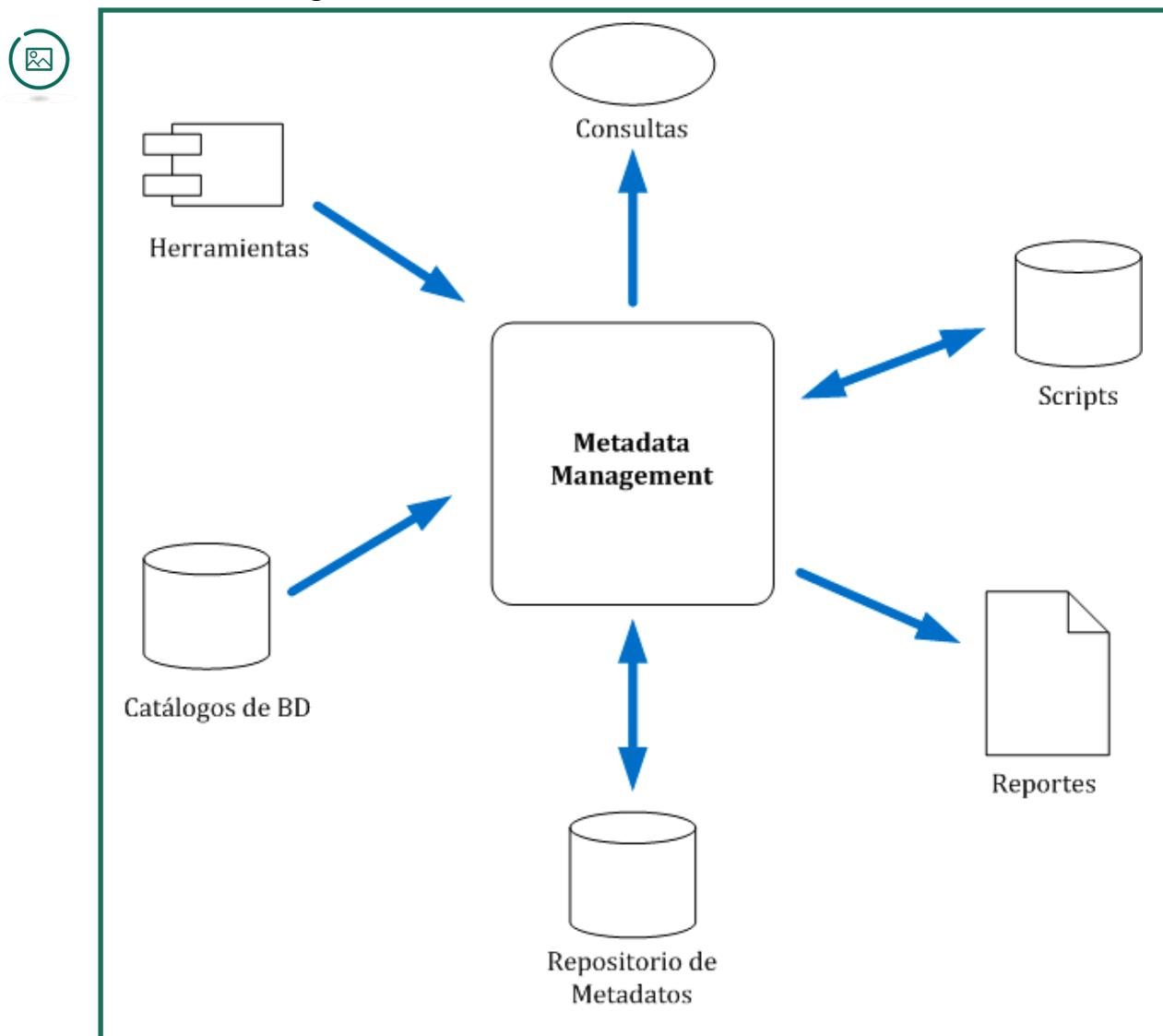
La construcción del Metadata supone que se defina el significado de cada una de las tablas y cada uno de los atributos que se cargan en el Data Warehouse. Este es un punto complejo de todo proyecto, ya que obliga a que se definan los conceptos de negocio y se homogeneicen entre los distintos departamentos, filiales, etc. Obliga a que todos los componentes de la organización hablen

utilizando la misma terminología y con el mismo significado, lo cual no siempre es sencillo. Cuando alguien hable de "margen bruto" o "margen de contribución" deberá estar absolutamente definido para la organización. Evidentemente, organizaciones distintas tendrán normalmente definiciones distintas.

(Cano, 2007, pp. 120-121)

En el siguiente gráfico, podemos ver distintos aspectos involucrados al Repositorio de Metadatos:

**Imagen 4: Metadatos**



Fuente: Elaboración propia

## Operational Data Store (ODS)

El ODS, como estructura de datos, da soporte a la toma de decisiones operativas, rutinarias y diarias de los niveles más bajos de la organización.

De acuerdo con William Inmon (2005), con los ODS fue posible hacer procesamiento en tiempo real contra datos integrados.

Existe un componente tecnológico, los Operational Data Store (ODS) que a veces se confunden con los Data Warehouses. Los ODS son una extensión de la tecnología de los Data Warehouses.

Los ODS consolidan datos de múltiples fuentes provenientes de distintos sistemas de información no integrados y facilitan un acceso online integrado sobre esa información. Su objetivo es proporcionar información integrada, con el fin de facilitar la toma de decisiones en entornos operacionales. Algunas veces se utilizan para evitar integraciones o implementaciones de soluciones ERP. La información que reside en los ODS es volátil y normalmente tiene, como máximo, una antigüedad de dos o tres meses. La principal diferencia con los Data Warehouses es que los datos de los ODS son volátiles y se actualizan en tiempo real. Los ODS habitualmente se convierten en una fuente de datos para el Data Warehouse.

(Cano, 2007, pp. 121-122)

## Clases de ODS

Siguiendo a William Inmon (2005), enunciamos cuatro clases de ODS cuya clasificación depende de la rapidez con que llegan los datos a la estructura del ODS:

- Clase I, en donde las actualizaciones de datos desde los sistemas transaccionales hacia el ODS son síncronas. En general, podemos hablar de que pasan milisegundos entre una actualización en el entorno operacional y en el ODS; el usuario, por lo tanto, no se da cuenta de la diferencia entre un esquema y el otro. Esta clase de ODS es cara y tecnológicamente desafiante; casi nunca se justifica

económicamente. Mayormente se opta simplemente por pasar datos del entorno operacional al ODS sin integrarlos.

- **Clase II**, en donde las actualizaciones ocurren dentro de un marco temporal de 2 a 3 horas. Esta clase de ODS es común. El mayor tiempo de sincronización permite la integración de los datos, es decir, que se ejecuten procesos ETL con mayor nivel de procesamiento y transformación de datos.
- **Clase III**, en donde la sincronización de las actualizaciones se produce cada 24 horas, generalmente de noche. Es similar a la clase II, aunque con una implementación marcadamente más económica.
- **Clase IV**, en donde las actualizaciones del ODS son frecuentemente a partir del *Data Warehouse* pero no están calendarizadas; es decir, existe un largo período de tiempo (incluso meses o años) entre la coordinación del ODS y su fuente. Generalmente la fuente aquí es el *Data Warehouse*, aunque puede provenir de otras fuentes.

De acuerdo con William Inmon (2005), un *Data Warehouse* nunca puede accederse en milisegundos. Debido a la naturaleza de sus datos, no está preparado para soportar procesos de tipo OLTP. Sin embargo, poder obtener tiempos de respuestas tan óptimos es algo muy valioso.

Cuando se requieren tiempos de respuesta óptimos y a su vez debe accederse a datos integrados, consolidados en una plataforma BI, debe emplearse el ODS, que es el repositorio para procesamiento de alta performance.

A diferencia del *Data Warehouse*, el ODS es opcional. Así, ambas estructuras son complementarias, ya que las dos residen fuera del entorno operacional, soportan procesamiento DSS o BI y usan datos integrados. Además, los flujos de datos entre ambos son bidireccionales. En algunas situaciones el ODS alimenta el *Data Warehouse*; en otras, el *Data Warehouse* alimenta el ODS. Pero a diferencia del *Data Warehouse*, el ODS está diseñado para procesamiento en tiempo real.

Las actualizaciones de datos en el ODS son normales, a diferencia del *Data Warehouse* donde siempre queda la foto histórica. Es decir, mientras en el *Data Warehouse* se almacenan valores históricos, en el ODS se almacenan valores actuales; esto implica que las aplicaciones BI que necesitan visualizar datos actuales, corren sobre el ODS y no sobre el *Data Warehouse*. Generalmente, un ODS no contiene más de un mes de histórico.

El diseño del ODS sigue un modelo híbrido, puede hacerse siguiendo modelo relacional en una parte y multidimensional en otra parte; dependerá de los requerimientos en cuanto a flexibilidad y performance.

El ODS suele contar con un volumen de datos muy inferior al *Data Warehouse*.

## Exploration Warehouse

De acuerdo con William Inmon (2005), el *Exploration Warehouse* es una forma especial de *Data Warehouse*. El objetivo de esta estructura consiste en proporcionar un fundamento para el análisis estadístico “pesado”. Una vez construido, se pueden ejecutar sobre él análisis estadísticos sin problemas, ya que estará en una máquina separada de donde ocurre el procesamiento regular del *Data Warehouse*. Es decir, el *Exploration Warehouse* permite satisfacer requerimientos de procesamiento temporal y desestructurado con altas velocidad de respuesta.

Otra explicación para la creación de un *Exploration Warehouse* es que la tecnología de análisis estadístico es tan diferente de otros estilos de análisis que tiene sentido ejecutarla sobre entornos separados. Además, otra razón es el diseño de bases de datos: el *Exploration Warehouse* casi nunca es una copia directa de los datos encontrados en el *Data Warehouse*. En lugar de ello, el *Exploration Warehouse* inicia con un subconjunto de los datos del *Data Warehouse*, incluso pueden agregarse algunos campos pre-calculados. Debido a la naturaleza de los requerimientos de explotación, sus datos se encuentran también altamente indexados.

Los *Exploration Warehouses* generalmente están enfocados en proyectos; es decir, una vez terminado un proyecto, desaparecen. Por tanto, tienen una existencia transitoria, lo cual contrasta con un *Data Warehouse*, ya que éste es de vida permanente. Además, a diferencia de un *Data Warehouse*, en ocasiones se puede “congelar” la información, sin actualizarla con los datos más nuevos. Dado a que se prueban distintos algoritmos, conviene hacerlos sobre los mismos datos, sin actualizarlos con tanta frecuencia.

En algunas situaciones, el *Exploration Warehouse* puede estar asociado con el *Data Mining Warehouse*.

## Data Mining Warehouse

De acuerdo con William Inmon (2005), un *Data Mining Warehouse* es similar en distintos aspectos a un *Exploration Warehouse*, pero con algunas diferencias:

- El objetivo primario de un *Exploration Warehouse* es la creación de afirmaciones, hipótesis y observaciones. Un *Data Mining Warehouse* tiene el objetivo de probar dichas hipótesis.
- Un *Exploration Warehouse* está optimizado en amplitud de información, mientras que el *Data Mining Warehouse* está optimizado en profundidad de información.
- Sin embargo, dado que las diferencias son sutiles, salvo en grandes corporaciones, pueden estar bajo la misma estructura.

Hay que tener en cuenta que se trata de un almacén de datos creado específicamente para contener las entradas y las salidas de los procesos de *Data Mining*, de manera que no interfieran ni con los sistemas operacionales, ni con los DSS.

## Herramientas de Business Intelligence

Las principales herramientas de Business Intelligence son:

- **Generadores de Reportes / Informes:** Utilizados por desarrolladores profesionales para crear informes estándar para grupos, departamentos o la organización.
- **Herramientas de Usuario Final de Consultas e Informes (Query & Reporting):** Empleadas por usuarios finales para crear informes para ellos mismos o para otros; no requieren programación.
- **Herramientas OLAP:** Permiten a los usuarios finales tratar la información de forma multidimensional para explorarla desde distintas perspectivas y períodos de tiempo.

- **Herramientas de Dashboard y Scorecard:** Permiten a los usuarios finales ver información crítica para el rendimiento con un simple vistazo, utilizando iconos gráficos y con la posibilidad de ver más detalle para analizar información detallada e informes si así lo desean.
- **Herramientas de Planificación, Modelización y Consolidación:** Permite a los analistas y a los usuarios finales crear planes de negocio y simulaciones con la información de Business Intelligence. Se utilizan para elaborar la planificación, los presupuestos, las previsiones. Estas herramientas proveen a los Dashboards y los Scorecards con los objetivos y los umbrales de las métricas.
- **Herramientas de Data Mining:** Permiten a estadísticos o analistas de negocio crear modelos estadísticos de las actividades de los negocios. Data Mining es el proceso para descubrir e interpretar patrones desconocidos en la información mediante los cuales resolver problemas de negocio. Los usos más habituales del Data Mining son: segmentación, venta cruzada, sendas de consumo, clasificación, previsiones, optimizaciones, entre otros.

(Cano, 2007, p. 132-133)

## Herramientas OLAP

Josep Lluís Cano expresa que “existen distintas tecnologías que nos permiten analizar la información que reside en un Data Warehouse, pero la más extendida es el OLAP” (2007, p. 125).

El autor también sostiene que:

Los usuarios necesitan analizar información a distintos niveles de agregación y sobre múltiples dimensiones: Por ejemplo, ventas de productos por zona de ventas, por tiempo, por clientes o tipo de cliente y por región geográfica. Los usuarios pueden hacer este análisis al máximo nivel de agregación o al máximo nivel de detalle. OLAP provee de estas funcionalidades y algunas más, con la flexibilidad necesaria para descubrir las relaciones y las

tendencias que otras herramientas menos flexibles no pueden aportar.

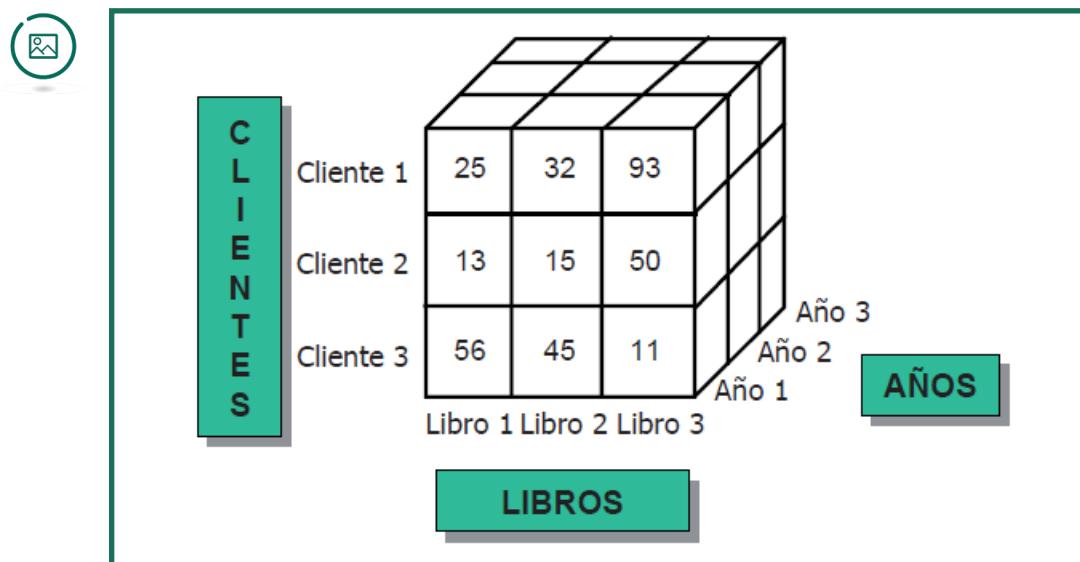
A estos tipos de análisis les llamamos **multidimensionales**, porque nos facilitan el análisis de un hecho desde distintas perspectivas o **dimensiones**. Esta es la forma natural que se aplica para analizar la información por parte de los tomadores de decisiones, ya que los **modelos de negocio normalmente son multidimensionales**.

(Cano, 2007, p. 126)

Generalmente, al hablar de análisis multidimensional u **OLAP (Online Analytical Processing)**, en Inglés) pensamos en un cubo, puesto que es la representación gráfica habitual del mismo.

Como puede verse en el siguiente ejemplo de Josep Lluís Cano, tenemos un cubo con la cantidad de unidades vendidas de cada uno de los libros (libro 1, libro 2 y libro 3), para cada uno de los clientes (cliente 1, cliente 2, cliente 3) y en cada uno de los distintos años (año 1, año 2, año 3):

**Imagen 5:** Cubo OLAP



Fuente: Josep Lluís Cano (2007, p. 127)

En el ejemplo anterior, la cantidad de unidades vendidas de cada uno de los libros por cliente y por año es lo que se denomina un **Hecho**. A su vez, las dimensiones son:

- Clientes
- Libros
- Años (Tiempo)

Incluso, una dimensión dada puede tener una jerarquía específica. Por ejemplo, generalmente la dimensión Tiempo tiene una jerarquía del tipo:

- Año
- Mes
- Día

Aunque, dependiendo el modelo de negocio, puede requerirse el análisis por semestre, trimestre, semana, intervalos horarios, día de la semana (lunes, martes, etc.).

Por otra parte, las Herramientas OLAP permiten realizar diferentes operaciones sobre los cubos:

- *Slice*
- *Dice*
- *Roll-Up*
- *Drill-Down*
- *Roll-Across*
- *Drill-Across*
- *Pivot*

## Herramientas de Visualización por Lógica Asociativa

De acuerdo con Josep Lluís Cano, “una alternativa al OLAP son las herramientas que utilizan consultas de lógica asociativa” (2007, p. 131). En este tipo de herramientas, cuando se realiza una consulta se accede directamente a los datos (ya sea de la base de datos transaccional o al *Data Warehouse* preferentemente) pero sin diseñar un cubo, sin dimensiones ni jerarquías predefinidas y sin restricciones en cuanto al volumen de información.

El modelo de almacenamiento interno proporciona una visión “vertical” (basada en columnas) de los datos así como una visión “horizontal ampliada” (basado en filas) que va más allá de la tecnología de bases de datos relacionales. Cada columna almacena cada valor diferente de forma separada con la frecuencia y usos de cada valor. Las consultas son altamente eficientes por el nuevo modelo de almacenamiento y el conjunto de operaciones utilizados para resolver las consultas.

Se indexan automáticamente el 100% de los datos y se eliminan automáticamente los datos redundantes y los valores nulos, lo que significa un menor uso de espacio de disco y menores tiempos de escritura y lectura.

(Cano, 2007, p. 132)

Este tipo de herramientas ha tenido un gran auge en los últimos años, en detrimento de las soluciones OLAP, puesto que las primeras reducen drásticamente los tiempos de desarrollo e implementación de una solución BI.



## Bibliografía de referencias

- **Harris, H.** (2014, 30 de Abril). *The History of Business Intelligence* [post de la web]. Recuperado de: <http://www.businesscomputingworld.co.uk/the-history-of-business-intelligence-infographic/>
- **Cano J. L.** (2007). *Business Intelligence: Competir con Información*. España: Banesto Fundación Cultural y ESADE.
- **Evelson, B.** (2008). *Topic Overview: Business Intelligence*. Recuperado el 30 de Abril de 2014: <http://www.forrester.com/Topic+Overview+Business+Intelligence-/E-RES39218?objectid=RES39218>
- **Gartner (s.f.)**. *IT Glossary - Business Intelligence*. Recuperado el 30 de Abril de 2014: <http://www.gartner.com/it-glossary/business-intelligence-bi>
- **Inmon William** (2005). *Building the Data Warehouse* (4º Edición). Estados Unidos: Wiley Publishing.
- **Kimball Ralph, R. M.** (2013). *The Data Warehouse Toolkit* (3º Edición). Estados Unidos: Wiley Publishing.
- **Laudon Kenneth, L. J. P.** (2004). *Sistemas de Información Gerencial: administración de la empresa digital* (8º Edición). México: Pearson Educación.
- **López Gutiérrez, M.** (2004, 30 de Abril). El lugar de los DSS en el proceso de toma de decisión. *GestioPolis* [post de la web]. Recuperado de 2014: <http://goo.gl/Xy2REI>
- **Parr Rud, O.** (2009). *Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy*. Hoboken, New Jersey: John Wiley & Sons.

# Data warehouse

---



Base de datos II

UNIVERSIDAD  
**SIGLO 21**

MIEMBRO DE LA RED  
**ILUMNO**

# » Data warehouse

“Un Data Warehouse es una colección de información creada para soportar las aplicaciones de toma de decisiones”

(Cano, 2007, p. 114)

## Data Warehouse: Definición

En el Módulo 1 definimos el concepto de *Data Warehouse* o Almacén de Datos como el componente central dentro del Corporate Information Factory (CIF) y, en general de toda solución de Business Intelligence (BI).

A continuación, analizaremos dos aspectos fundamentales en la construcción de un *Data Warehouse*:

- Granularidad
- Particionamiento

### Granularidad

De acuerdo con William Inmon (2005), el aspecto más importante en el diseño de un *Data Warehouse* es la **granularidad**, la cual se refiere al nivel de detalle o sumarización de las unidades de datos dentro del *Data Warehouse*.

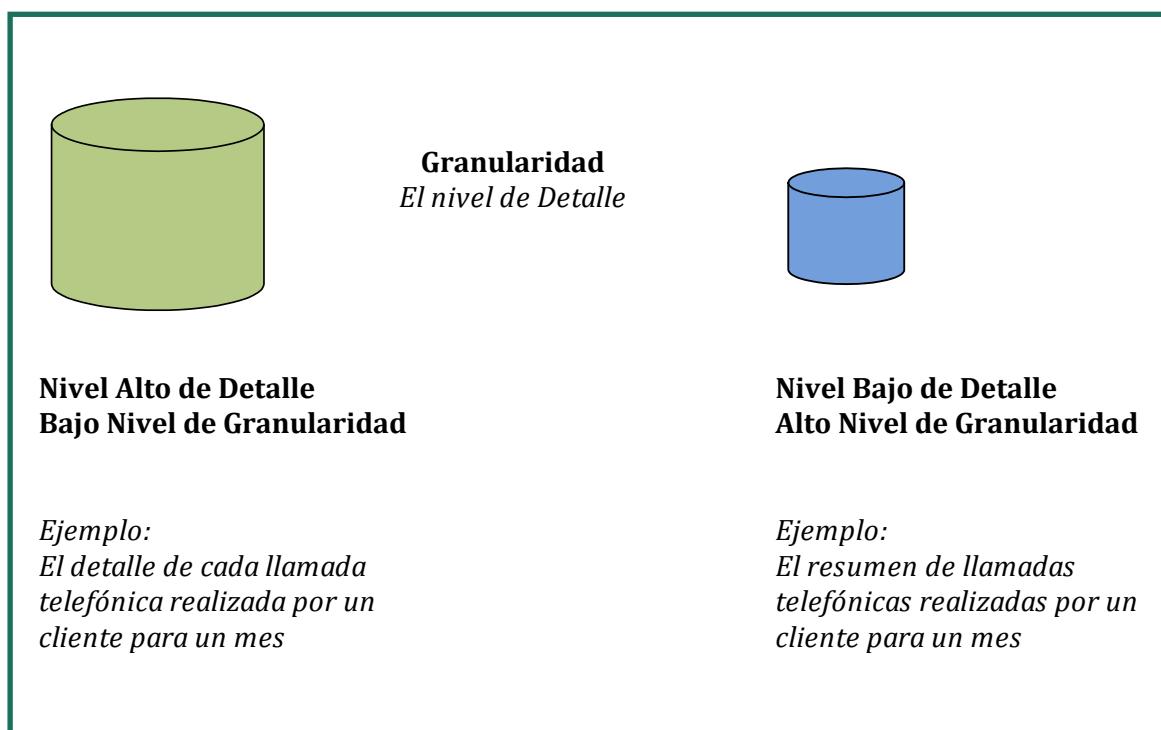
Mientras más detalle tenga el *Data Warehouse*, más bajo será el nivel de granularidad. A menor detalle, mayor nivel de granularidad. Por ejemplo,

una única transacción estaría en el nivel más bajo de granularidad, mientras que un resumen de todas las transacciones realizadas durante el mes en curso estaría en un nivel más alto de granularidad.

En los sistemas operacionales/transaccionales siempre se almacena en las bases de datos a un bajo nivel de granularidad, es decir, máximo detalle. Sin embargo, en el *Data Warehouse* (si bien es recomendable) no siempre es así.

Se dice que la granularidad es el principal aspecto de diseño de un *Data Warehouse* debido a que afecta profundamente el volumen de datos general que reside en el *Data Warehouse* y el tipo de consultas que se pueden ejecutar. Mientras más bajo sea el nivel de granularidad, más versátil será el *Data Warehouse* a la hora de poder responder distintos tipos de consultas.

**Imagen 1:** Granularidad



Fuente: Elaboración propia en base a William Inmon (2005, p. 42)

Según William Inmon (2005), los beneficios de un bajo nivel de granularidad son los siguientes:

- Los datos granulares del *Data Warehouse* son la clave para la responsabilidad, dado que pueden ser usados por distintas personas y de diversas maneras (un usuario quizás quiera ver los datos sumarizados por mes, otro por año, otro por región geográfica, entre otras dimensiones posibles).
- La capacidad para conciliar datos: si hay discrepancias entre dos o más departamentos sobre un determinado valor de una métrica o cálculo, existe un sólido y único fundamento detallado sobre el cual indagar.
- La flexibilidad para formatear los datos según las necesidades de los usuarios de distintos departamentos.
- Contiene una historia detallada de las actividades y eventos de la corporación.
- Pero el principal beneficio es que permite acomodarse a los requerimientos futuros. Con un *Data Warehouse* de bajo nivel de granularidad se logra mejorar sustancialmente la adaptación al cambio frente a nuevos requerimientos.
- Permite, además, responder con precisión a determinadas consultas o requerimientos, lo cual no podría ser posible si los datos estuvieran sumarizados.
- Permite la construcción de ***Data Marts*** o tablas sumarizadas a partir de los datos detallados.
- Permite la optimización de los procesos de Visualización/Exploración y *Data Mining*, que requieren de datos detallados e históricos para descubrir patrones ocultos en los datos.

## Niveles Diales de Granularidad y Tablas Sumarizadas

Sin embargo, William Inmon (2005) destaca que cuando una organización tiene gran necesidad de eficiencia en el almacenamiento y acceso a los datos, y a la vez la necesidad de contar con la capacidad de analizar datos en detalle sobre un *Data Warehouse* de gran volumen de datos, debería considerarse la posibilidad de contar con dos o más niveles de granularidad en simultáneo.

De esta manera, puede tenerse una tabla detallada y una o más tablas **sumarizadas** que favorecen determinadas consultas al estar ya pre-calculadas; es decir, además de contar con una estructura de tablas con todos los registros históricos detallados, contar también con tablas que sumaricen esos registros para facilitar las consultas y mejorar la performance de las aplicaciones DSS que usen el *Data Warehouse* como fuente.

Estas tablas summarizadas tienen, notoriamente, menos registros. Según William Inmon (2005), hay que tener en cuenta que aproximadamente el 95% del procesamiento de un DSS corre sobre datos summarizados y sólo un 5% sobre datos detallados (esto se debe a que, por lo general, los directivos de una organización están interesados en conocer los grandes números -cómo van las ventas, por ejemplo, y no el detalle completo de cada transacción en particular).

Por lo tanto, contar con niveles duales de granularidad puede permitir procesar la mayoría de los requerimientos de manera eficiente, accediendo generalmente a la tabla summarizada y solo en casos puntuales a los casos detallados.

Según William Inmon (2005), y en función de lo anterior, contar con niveles duales de granularidad es la mejor opción arquitectónica para un *Data Warehouse*.

## Particionamiento

De acuerdo con William Inmon (2005), después de la granularidad, un segundo aspecto de importancia en el diseño de un *Data Warehouse* es el **particionamiento**, el cual se refiere a la división o partición de los datos en tablas o unidades físicas separadas que pueden ser manejadas independientemente.

Un particionamiento apropiado permite mejorar el *Data Warehouse* en cuanto a la carga, acceso, archivado, eliminación, monitoreo y almacenamiento de los datos.

Las pequeñas unidades o tablas pueden ser:

- Reestructuradas
- Indexadas
- Secuencialmente escaneadas
- Reorganizadas
- Recuperadas
- Monitoreadas.

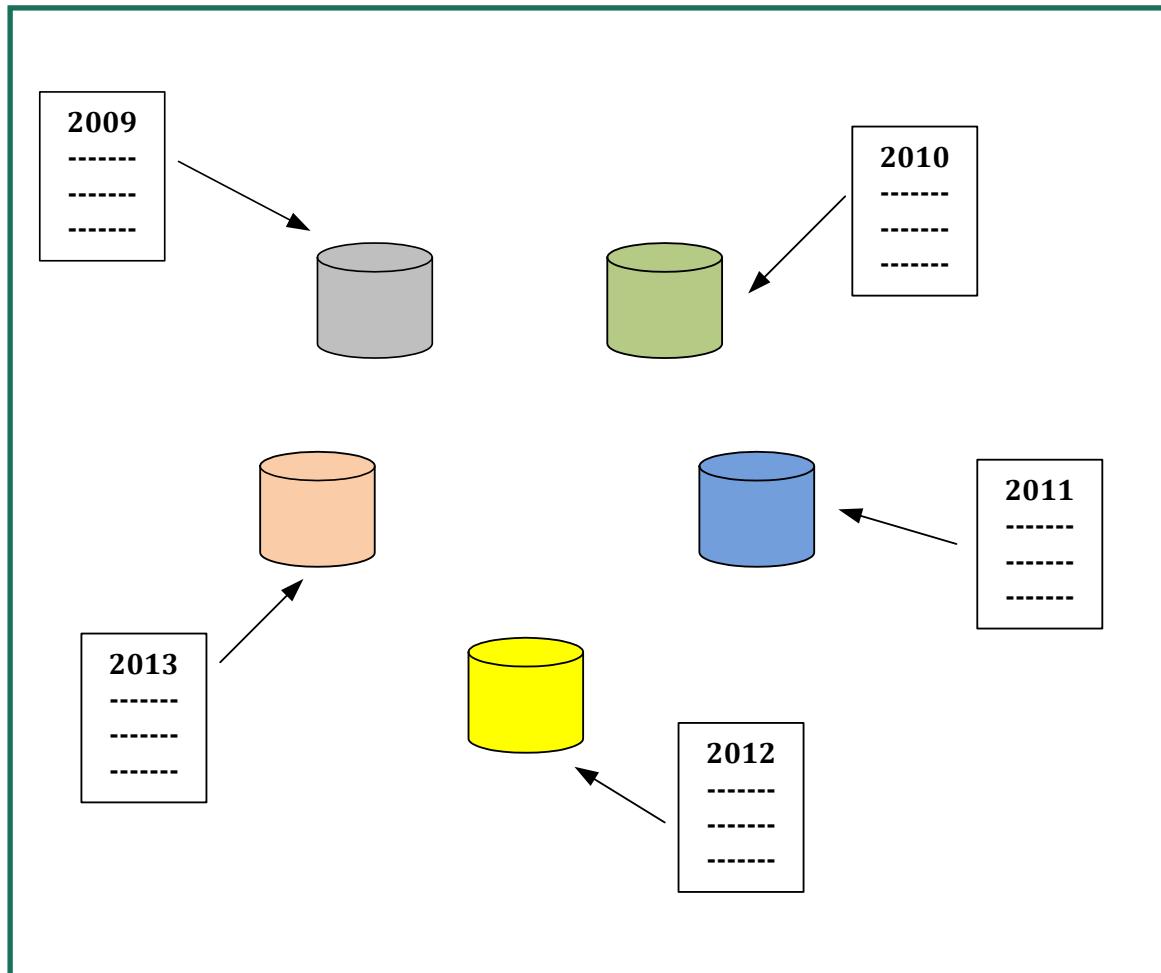
El particionamiento puede hacerse a nivel del motor de Base de Datos (DBMS, *DataBase Management System* en inglés) o incluso del sistema operativo y a nivel de aplicación. Cada enfoque tiene sus ventajas y sus desventajas. Cuando se hace a nivel de aplicación, presenta un mayor grado de flexibilidad.

Los datos pueden dividirse por muchos criterios, tales como:

- Por fecha
- Por unidad de negocios
- Por región geográfica
- Por unidad organizacional
- Por todo lo anterior.

Aunque definitivamente la fecha es un criterio mandatorio, casi siempre es empleado. Incluso, en ocasiones puede ser necesario hacerlo ya que la definición o estructura de los datos puede variar de un año a otro. Por ejemplo, en lugar de tener una tabla con gran cantidad de registros, podríamos tener varias tablas con las actividades de los años 2009, 2010, 2011, 2012 y 2013, por separado.

**Imagen 2:** Particionamiento



**Fuente:** Elaboración propia en base a William Inmon (2005, p. 42)

## Datawarehouse: Comparación con OLTP

Según William Inmon (2005), es importante hacer la distinción entre datos primitivos y datos derivados. Los datos primitivos son los que se almacenan en las bases de datos por los propios sistemas transaccionales (OLTP), mientras que los datos derivados son en realidad una transformación de los mismos para poder ser usados por los DSS con el objeto de tomar decisiones.

Las diferencias son las siguientes:

**Tabla 1:** Comparación con OLTP



| Datos Primitivos / Operacionales (OLTP)                                 | Datos Derivados / Decisionales (DSS)                          |
|---|---|
| Están orientados y organizados según las necesidades de cada aplicación | Están orientados y organizados por área de interés            |
| Detallado   | Detallado y sumarizado  |
| Preciso al momento del acceso (valores actuales)                        | Representa valores en el tiempo, históricos, <i>snapshots</i> |
| Sirve a la comunidad operativa de la organización                       | Sirve a la comunidad directiva de la organización             |
| Pueden actualizarse   | No se actualizan  |
| Las aplicaciones corren repetitivamente sobre ellos                     | Las aplicaciones corren heurísticamente sobre ellos           |
| Requerimientos de procesamiento comprendidos a priori                   | Requerimientos de procesamiento NO comprendidos a priori      |
| Compatibles con el ciclo de vida de desarrollo de <i>software</i>       | Ciclo de vida completamente diferente                         |
| Sensible a la <i>performance</i>  | Relajado en cuanto a <i>performance</i>                       |
| Accedidos una unidad por vez  | Se accede a un conjunto por vez (valores sumarizados)         |

| <b>Datos Primitivos / Operacionales (OLTP)</b>                                | <b>Datos Derivados / Decisionales (DSS)</b>                                     |
|---|---|
| Gestionado por la transacción   | Gestionado por el análisis de información                                       |
| Control de las actualizaciones en términos de propiedad o dueño del dato      | El control de las actualizaciones no interesa                                   |
| Alta disponibilidad   | Disponibilidad más relajada   |
| Gestionado en su completitud  | Gestionado por subconjuntos   |
| Sin redundancias  | Se permiten las redundancias  |
| Estructura estática, contenido variable                                       | Estructura flexible   |
| Pequeñas cantidades de datos usados en un proceso                             | Grandes cantidades de datos usados en un proceso                                |
| Soporta las operaciones diarias   | Soporta las necesidades gerenciales de la organización                          |
| Alta probabilidad de acceso a un dato en particular                           | Baja probabilidad de acceso a un dato en particular                             |
| Naturaleza dinámica de los datos  | Naturaleza estática hasta el próximo <i>refresh</i> de datos                    |
| Se usan para el procesamiento repetitivo de los datos, altamente estructurado | Se usan para el procesamiento analítico de los datos, altamente desestructurado |
| Estructura con muchas tablas altamente normalizadas                           | Estructura con pocas tablas de-normalizadas                                     |

Fuente: Elaboración propia en base a William Inmon (2005, p. 15)

En definitiva, según William Inmon (2005) las principales diferencias entre ambos esquemas son:

- Los datos primitivos son datos detallados, usados para ejecutar las operaciones diarias de la organización. Los datos derivados han sido sumarizados o pre-calculados para satisfacer las necesidades de la gerencia.
- Los datos primitivos pueden actualizarse. Los datos derivados no pueden re-calcularse porque no pueden ser actualizados directamente.
- Los datos primitivos son datos de valor actual. Los datos derivados son datos históricos.
- Los datos primitivos son operados por procedimientos repetitivos. Los datos derivados son operados por programas heurísticos, no repetitivos.
- Los datos de los sistemas transaccionales OLTP son datos primitivos. Los datos de los sistemas DSS son datos derivados.
- Los datos primitivos apoyan las tareas operativas. Los datos derivados soportan la toma de decisiones gerenciales.

## Tipos de implementación

Existen dos modelos básicos para el diseño de la base de datos de un *Data Warehouse*:

- El **Modelo Relacional** (propuesto y defendido por William Inmon)
- El **Modelo Multidimensional** (propuesto y defendido por Ralph Kimball).

De acuerdo con William Inmon (2005):

- En el **Modelo Relacional**, los datos están altamente normalizados (tercera forma normal). Esta normalización implica que el diseño de la base de datos genere un muy bajo nivel de granularidad. El valor del Modelo Relacional para el *Data Warehouse* es que hay disciplina en la forma en que se construye el diseño de la base de datos, claridad del significado y uso del nivel detallado de datos

normalizados bajo la tercera forma normal. Así, el Modelo Relacional produce un diseño muy flexible. Flexibilidad en términos de que el diseño puede adaptarse a distintas vistas. Esta es la mayor fortaleza junto con la **versatilidad**, dado que los datos detallados pueden combinarse: pueden soportarse muchas vistas distintas.

- En cambio, el **Modelo Multidimensional**, también conocido como esquema Estrella (o *Star Join*) tiene una tabla *Fact* (tabla de Hechos) en el centro del modelo, que contiene gran cantidad de registros, de-normalizados, y alrededor de ella una serie de tablas de Dimensiones que describen cada uno de los campos de la tabla *Fact*. Generalmente, las Dimensiones tienen pocos registros (sobre todo en comparación con la *Fact*) y contienen información relevante separada (ubicación de sucursales, calendario, etc.). La tabla *Fact* y las tablas de dimensiones están asociadas por un campo referencial en común. La gran ventaja del modelo Multidimensional es su eficiencia de acceso. Su estructura se basa en los requerimientos del usuario.

En la Unidad 3 (Módulo 3) dedicaremos más tiempo a trabajar con el modelado multidimensional.

Las principales diferencias entre ambos modelos, según William Inmon (2005), son:

- El Modelo Relacional es altamente flexible, pero no está optimizado en términos de performance para ningún usuario en particular. El modelo Multidimensional es altamente eficiente al servir las necesidades de una comunidad de usuarios en particular, pero no es bueno en cuanto a su flexibilidad.
- El Modelo Multidimensional tiene un alcance más limitado en el sentido de que el diseño se optimiza para un conjunto de requerimientos de usuarios, se ajusta mejor a los requerimientos de un departamento. En cambio, el Modelo Relacional está orientado a un alcance mayor (modelo empresarial).
- El Modelo Relacional está basado en un modelo de datos corporativo o empresarial, mientras que el Modelo Multidimensional lo está en función de un modelo basado en los requerimientos de procesamiento para satisfacer las demandas del usuario.
- Si bien el Modelo Relacional no es óptimo en performance para el acceso directo a los datos, debido a su diseño flexible pueden generarse estructuras especiales (tablas sumarizadas) para un acceso indirecto a los datos más óptimo. En estos términos, el

Modelo Multidimensional ofrece una performance similar con un acceso directo a los datos.

## Arquitectura

### La Arquitectura de William Inmon

William Inmon (2005) sentó las bases de la primera definición arquitectónica de BI en base al *Corporate Information Factory*, tal como ya vimos en la Unidad 1 (Módulo 1).

Además, sostiene que el Modelo Relacional es mucho mejor para el diseño de un *Data Warehouse* dado que se necesita soportar el acceso de muchos usuarios distintos con diferentes requerimientos. Es decir, considera que el *Data Warehouse* no debe estar optimizado para el acceso de un usuario en particular.

Para William Inmon (2005), la clave está en el “reshaping” del Modelo Relacional. Debido al nivel de granularidad que provee este modelo, es relativamente fácil crear tablas “sumarizadas” que se elaboran a partir de necesidades específicas de un conjunto único de usuarios.

De esta manera, esta tabla sumarizada se encuentra lista para su acceso directo, altamente eficiente en términos de performance. Se pueden crear tantas tablas sumarizadas como sean necesarias. La sencillez de su creación se deriva de que:

- Los datos están almacenados al nivel más granular, más normalizado.
- Las relaciones entre tablas relacionales ya están identificadas.
- Nuevas tablas pueden contener nuevas sumarizaciones, nuevos criterios de selección y nuevas agregaciones.

En el caso del Modelo Multidimensional, dado que está optimizado para un grupo único de usuarios, cualquier otro usuario debe pagar el precio de una performance que no es la óptima. Según William Inmon (2005), este modelo no garantiza la optimización para todos los usuarios para todos sus requerimientos.

La ventaja del Modelo Relacional apunta a su flexibilidad al cambio, hacia nuevos requerimientos o nuevos grupos de usuarios que tienen necesidades distintas.

Las tablas sumarizadas proveen agregaciones de los datos a nivel granular, es decir, cualquier combinación de átomos. Además, otra ventaja se deriva de que si para otro usuario se requiere otro cálculo, no es necesario tocar el modelo base, ya que simplemente se agrega una nueva tabla sumarizada. Se reduce y aísla el impacto al cambio. En el Modelo Multidimensional, en cambio, el impacto puede ser mucho más profundo, al tener que cambiar la misma estructura.



*Debido a estas causas William Inmon (2005) sostiene que el modelo Relacional es ideal para el diseño de un Data Warehouse, mientras que el modelo Multidimensional es ideal para el diseño de un Data Mart.*

William Inmon (2005) alega que los *Data Warehouses* se diseñan a partir de los requerimientos de información corporativos de toda la organización, y no desde los requerimientos departamentales (como un *Data Mart*). Por lo tanto, crear un *Data Warehouse* con un modelo Multidimensional sería un error ya que el resultado será un *Data Warehouse* optimizado para una comunidad de usuarios a expensas de otros usuarios.

## La Arquitectura de Ralph Kimball

Ralph Kimball (2013), en cambio, sostiene que el *Data Warehouse* debería basarse sobre un Modelo Multidimensional, también con datos detallados.

Según su enfoque, el paso final de los procesos ETL consiste en la carga de los datos en la estructura física bajo un Modelo Multidimensional, tablas que luego serán la base del área de presentación de BI (las aplicaciones DSS). Estos subsistemas ETL son críticos, muchos de los cuales se enfocan en el procesamiento de las tablas de dimensión, mientras que las tablas de hechos si bien tienen muchos registros, no suelen demandar gran preparación. Una vez actualizadas, indexadas y con la calidad de datos asegurada, estas tablas son publicadas para los usuarios.

Existe un debate respecto de si los datos debieran cargarse primero en una estructura normalizada previo a cargarlos en un modelo dimensional para la consulta y *reporting*. Si bien esto es aceptable, la creación de estructuras normalizadas para el ETL y estructuras dimensionales para el área de

presentación implican doble proceso de ETL, lo cual requiere más tiempo e inversión para el desarrollo, mayor tiempo de actualización de los datos y más capacidad para almacenar las múltiples copias de datos. Aunque la consistencia de datos a nivel empresarial es un objetivo fundamental de un *Data Warehouse*, puede ser más efectivo y menos costoso no incluir en el ETL estas estructuras normalizadas.

El Área de Presentación BI es aquella donde los datos se organizan, almacenan y están disponibles para la consulta de los usuarios y aplicaciones analíticas. Dado que las herramientas ETL están fuera de sus límites, esta área es todo lo que ven los usuarios.



*Ralph Kimball (2013) insiste en que los datos sean presentados, almacenados y accedidos en esquemas multidimensionales.*

Además, insiste en que el área de presentación debe tener datos detallados, atómicos, los cuales se requieren para responder a consultas de usuarios *ad hoc*, impredecibles. Aunque contenga datos sumarizados con alta performance, no es suficiente con entregar esta información sin contar con los datos granulares en una forma dimensional. Es decir, es inaceptable guardar datos sumarizados en modelos dimensionales teniéndolos en estructuras normalizadas.

El Área de Presentación BI debería ser estructurada alrededor de los eventos de medición del proceso de negocio. Este enfoque se alinea con los sistemas operacionales. Los modelos dimensionales deberían corresponderse a los eventos de captura de datos físicos, no deben diseñarse para entregar el reporte del día. Los procesos de negocio cruzan los límites de los departamentos de la organización. Debe construirse una sola tabla de hechos con los datos atómicos de las ventas en lugar de generar estructuras separadas similares.

Todas las estructuras dimensionales deben diseñarse usando dimensiones comunes. Esta es la base del *Enterprise Data Warehouse Bus Architecture*. Sin dimensiones compartidas, un modelo dimensional se vuelve una aplicación *standalone*. Los datos aislados generan vistas incompatibles de la empresa. Con dimensiones compartidas, los modelos dimensionales pueden ser compartidos. En una gran empresa, podemos encontrar una docena de modelos dimensionales combinados con tablas de dimensión compartidas.

Aunque la arquitectura de Kimball habilita normalización opcional para soportar el procesamiento ETL, el *Enterprise Data Warehouse* normalizado es un requisito obligatorio en el Corporate Information Factory de Inmon.

Al igual que el enfoque de Kimball, el Corporate Information Factory subraya la coordinación e integración de datos empresariales. El Corporate Information Factory dice que el *Enterprise Data Warehouse* normalizado satisface este rol, mientras que la arquitectura de Kimball recalca la importancia de un *Enterprise Data Warehouse Bus Architecture* con dimensiones compartidas.

## La Arquitectura Híbrida Inmon – Kimball

Según Ralph Kimball (2013), se puede plantear una arquitectura híbrida que integre los conceptos tanto de uno como de otro enfoque.

Esta arquitectura carga un *Enterprise Data Warehouse* al estilo del Corporate Information Factory, que está completamente fuera de alcance para los usuarios para *reporting* y análisis. Es solo la fuente para cargar un área de presentación tipo Kimball en la cual los datos son multidimensionales, atómicos/muy detallados (complementados por sumarizaciones), centrado en procesos, y dentro del *Enterprise Data Warehouse Bus Architecture*.

Este enfoque trae lo mejor de los dos mundos, pero este esquema híbrido sólo puede ser apropiado si la empresa ya invirtió en un *Enterprise Data Warehouse normalizado*, porque iniciar desde cero implica más costos y más tiempo, tanto en desarrollo como operación continua, dados los múltiples movimientos de datos y almacenamiento redundante de datos detallados.

## Desarrollo iterativo del Data Warehouse

De acuerdo con William Inmon (2005), es conveniente construir un Data Warehouse de manera iterativa ya que el usuario es incapaz de articular muchos requerimientos hasta que la primera iteración finalice, además de que la gerencia generalmente no se compromete por completo hasta ver algunos resultados tangibles.

Ralph Kimball (2013) sostiene desde su enfoque, que usando el *Enterprise Data Warehouse Bus Architecture* puede desarrollarse un *Data Warehouse* de manera iterativa, gradual, ágil, descentralizada.



## Bibliografía de referencias

- **Cano, J. L.** (2007). *Business Intelligence: Competir con Información*. España: Banesto Fundación Cultural y ESADE.
- **Inmon, W.** (2005). *Building the Data Warehouse* (4º Edición). Estados Unidos: Wiley Publishing.
- **Kimball Ralph, R. M.** (2013). *The Data Warehouse Toolkit* (3º Edición). Estados Unidos: Wiley Publishing.
- **Niven, P.** (2002). *El Cuadro de Mando Integral Paso a Paso*. España: Ediciones Gestión 2000.
- **Two Crows Corporation** (1999). *Introduction to Data Mining and Knowledge Discovery* (3º Edición). Recuperado el 5 de Mayo de 2014:  
<http://www.stanford.edu/class/stats315b/Readings/DataMining.pdf>

# Datamart, Datamining y Balanced Scorecard

---



Base de Datos II

UNIVERSIDAD  
**SIGLO 21**

MIEMBRO DE LA RED  
**ILUMNO**

# » Datamart, Dataminig y Balanced Scorecard

## Datamart

### Arquitecturas Disponibles

En cuanto al diseño de los *Data Marts*, también podemos destacar distintos tipos de arquitecturas disponibles.

- *Data Marts* aislados/independientes
- *Data Marts* interconectados/dependientes

### Data Marts Independientes

De acuerdo con William Inmon (2005), la arquitectura de *Data Marts* Independiente consiste en diseñarlos directamente desde las aplicaciones fuente (sistemas transaccionales). Un *Data Mart* Independiente puede ser creado por un departamento, sin consideración alguna de otros departamentos ni la participación del área centralizada de Sistemas.

Un *Data Mart* independiente representa el conjunto de requerimientos de BI de un sector en particular, por ello es relativamente fácil, rápido y económico de construir, a la vez que le permite a ese sector o departamento manejar por sí mismo su política de BI. Es por esto que son tan populares.

Con este enfoque, los datos analíticos se despliegan en una base departamental sin consideraciones de compartir e integrar información a lo largo de la empresa. Generalmente, un único departamento identifica los requerimientos de datos desde un sistema operacional. Así, se construye una base de datos que satisface sus necesidades departamentales, pero este *Data Mart* departamental refleja sólo los requerimientos analíticos del departamento al trabajar en forma aislada.

Mientras tanto, otro departamento se interesa en los mismos datos fuente, pero dado que no tiene acceso al *Data Mart* inicialmente construido por el otro, procede de manera similar construyendo una solución departamental aislada. De esta manera, cuando los usuarios se reúnen a discutir sobre una misma métrica, encontramos dos visiones distintas porque el cálculo no

necesariamente dio el mismo resultado al usar distintas fórmulas o reglas de negocio.

Algunos de los defensores del modelo Multidimensional favorecen esta estrategia. Por su parte, William Inmon (2005) sostiene que se trata una estrategia de corto plazo, de enfoque limitado, ya que no provee una base firme para la consolidación de la información corporativa.

En su análisis, los *Data Marts* Independientes presentan varias desventajas:

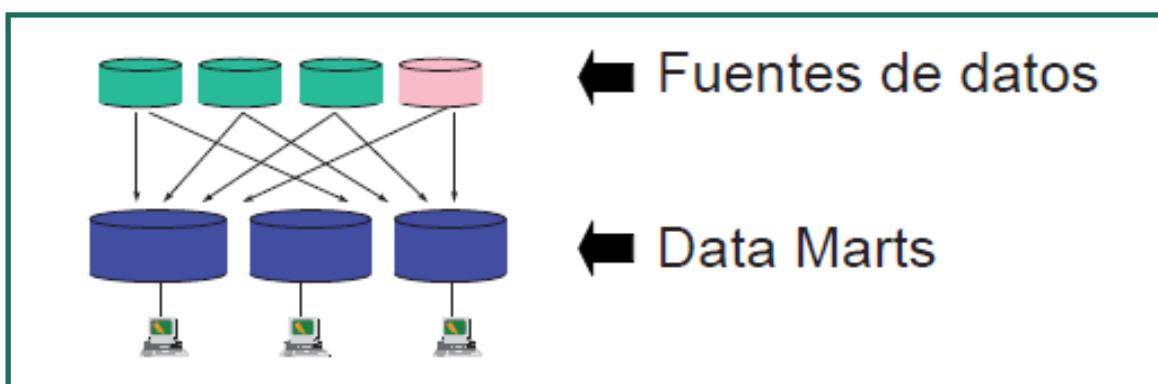
- No proporcionan una plataforma para la reusabilidad.
- No proporcionan una base para la reconciliación de datos.
- No proporcionan un único conjunto de programas de interfaz (ETL) con las aplicaciones transaccionales.
- Requieren que cada *Data Mart* independiente construya su propio “pool” de datos detallados, lo cual es redundante con otros *Data Marts* independientes de la misma empresa.

Ralph Kimball (2013) también sostiene que es una estrategia de desarrollo rápida, de bajo costo en el corto plazo; pero múltiples extracciones de datos no coordinadas y el almacenamiento redundante de datos analíticos son ineficientes y de alto costo en el largo plazo.

Sin una perspectiva empresarial, este enfoque termina en una gran cantidad de aplicaciones aisladas y vistas incompatibles del desempeño organizacional. Kimball también desaconseja este enfoque.



**Imagen 3: *Data Marts* Independientes**



Fuente: Josep Lluís Cano (2007, p. 118)

## Data Marts Dependientes

Según William Inmon (2005), la arquitectura de *Data Marts* Dependientes se basa en un *Data Warehouse* centralizado; los datos de los *Data Marts* provienen de allí (tal cual se plantea en el Corporate Information Factory).

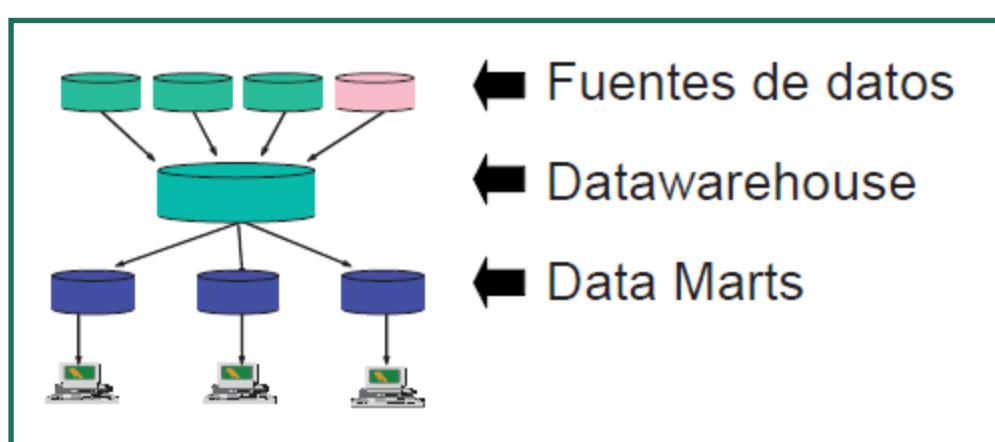
Es decir, el *Data Mart* dependiente no se somete de los datos operacionales, sino del *Data Warehouse*. Esta estrategia requiere, desde luego, de una mayor inversión, planificación, perspectiva a largo plazo y cooperación y coordinación en la definición de los requerimientos entre los diferentes departamentos de la organización.

De acuerdo con William Inmon (2005), un aspecto clave es cómo transferir los datos desde el *Data Warehouse* hacia el *Data Mart*. Los datos en un *Data Warehouse* están muy detallados, con bajo nivel de granularidad. Los datos en un *Data Mart* suelen estar más compactos y sumarizados. Periódicamente deben enviarse los datos de una estructura a la otra. Este movimiento es análogo a un proceso ETL. Además, los *Data Marts* se diseñan, generalmente, siguiendo un modelo Multidimensional.

Asimismo, la estructura de datos multidimensional en cada uno de los *Data Marts* está especialmente diseñada para cada uno de los requerimientos departamentales. Así, cada departamento requerirá una estructura de *Data Mart* distinta. Todas estas estructuras se alimentarán desde el mismo *Data Warehouse* y pueden estar en una base de datos relacional (esquema multidimensional conocido como "Star Join") o en una base de datos multidimensional (cubos OLAP).

De acuerdo con William Inmon (2005), diferentes *Data Marts* requieren de un distinto nivel de granularidad. Ahora, para poder alimentar a los distintos *Data Marts* se requiere que el *Data Warehouse* tenga el nivel más bajo de granularidad que cualquiera de los *Data Marts* que alimentará.

 **Imagen 4:** Data Marts Dependientes



Fuente: Josep Lluís Cano (2007, p. 118)

## Modos de implementación

Existen dos modos de implementación de un *Data Warehouse* y los respectivos *Data Marts*:

- **Top Down**, estrategia propuesta por William Inmon (2005). Como afirma Josep Lluís Cano, “propone definir un Data Warehouse corporativo y a partir de él ir construyendo los modelos de análisis para los distintos niveles y departamentos de la organización; es decir, una estrategia de arriba abajo, desde la estrategia a lo más operativo” (2007, p. 118).
- **Bottom Up**, estrategia inicialmente propuesta por Ralph Kimball (2013). Como afirma Josep Lluís Cano, propone “construir distintos Data Marts que cubran las distintas necesidades de la organización, sin la necesidad de construir un Data Warehouse” (2007, p. 119).

Es evidente que ambas alternativas pueden ser viables en función de las características de cada proyecto. Cada una de ellas presenta sus propias ventajas y desventajas.

Como afirma el profesor Hugh J. Watson, cuando se desarrollan correctamente las dos estrategias son válidas.

Con la estrategia de definir un Data Warehouse corporativo, el Data Warehouse es desarrollado en fases y cada una de las mismas debe ser diseñada para generar valor para el negocio. Se construye un Data Warehouse corporativo, del que se cuelga un Data Mart dependiente con una parte de la información del Data Warehouse. En fases posteriores se van desarrollando Data Marts usando subconjuntos del Data Warehouse. Igual que los proyectos complejos, es caro, necesita mucho tiempo y es propenso al fracaso. Cuando tenemos éxito conseguimos un Data Warehouse integrado y escalable.

Si optamos por la estrategia más común, la de construir distintos Data Marts, el proyecto comienza con un Data Mart único al que posteriormente se irán añadiendo otros Data Marts que cubrirán otras áreas de negocio. Normalmente no requiere de grandes inversiones y es fácil de implementar, aunque conlleva algunos riesgos; de entre ellos, cabe destacar fundamentalmente dos: puede perpetuar la existencia del problema de “silos de

información” y posponer la toma de decisiones que conciernen a la definición de criterios y modelos de negocio. Si seguimos esta estrategia, debemos tener claro el plan de acción, es decir, qué áreas cubriremos y la integración de los distintos modelos. Esta estrategia se utiliza a veces como un paso previo al desarrollo de un Data Warehouse corporativo.

Las dos aproximaciones abogan por construir una arquitectura robusta que se adapte fácilmente a los cambios de las necesidades de negocio y que nos proporcione una sola versión de la verdad.

(Cano, 2007, p. 119)

## Data Mining

### Definición

Según un informe de Two Crows Corporation (1999), la Minería de Datos (más conocida por el término *Data Mining*, en Inglés) consiste en la búsqueda de patrones y relaciones ocultas en los datos de las organizaciones.

*Data Mining*, no es más que la aplicación de algoritmos específicos al *Data Warehouse* para obtener resultados útiles. En otras palabras, consiste en la aplicación de técnicas matemáticas, estadísticas y principalmente de Inteligencia Artificial para descubrir relaciones, patrones y tendencias en los datos almacenados por una organización.

Actualmente, existen enormes bases de datos en donde se encuentra oculta información estratégica de gran relevancia, a la que no se puede acceder a través de las técnicas tradicionales de recuperación de información. Para revelar esta información, es necesario hacer uso de la minería de datos que se trata de un proceso que utiliza una amplia variedad de herramientas de análisis de datos para descubrir patrones y relaciones en los datos, los cuales son utilizados posteriormente para realizar predicciones válidas.

La minería de datos es parte de un proceso iterativo mayor llamado *descubrimiento de conocimiento en base de datos* (KDD, Knowledge Discovery in Databases en Inglés).

Es importante destacar que la minería de datos no reemplaza a las técnicas de estadísticas tradicionales, sino que, por el contrario, muchas de las técnicas más utilizadas en la minería de datos tienen sus raíces en las aplicaciones estadísticas, como los modelos utilizados para *clustering* y segmentación.

Generalmente, no se distingue la diferencia entre OLAP (On-Line Analytical Processing) y *Data Mining*. El análisis OLAP es esencialmente un proceso deductivo: el analista OLAP genera una serie de patrones y relaciones hipotéticas y utiliza consultas contra la base de datos para verificarlas o refutarlas. La Minería de Datos difiere del análisis OLAP, puesto que en lugar de verificar patrones hipotéticos utiliza los datos para descubrir patrones y es esencialmente un proceso inductivo. La minería de datos y OLAP se pueden complementar entre sí ya que OLAP puede ser usado antes de la minería de datos para ayudar a explorar los datos, focalizar la atención sobre las variables importantes, identificar excepciones o encontrar interacciones, contribuyendo de esta manera a un mayor entendimiento de los datos.

## Aplicaciones de Minería de Datos

Entre las principales aplicaciones de Minería de Datos, podemos detallar las siguientes:

- **Venta Cruzada (Cross Selling):** las técnicas de minería se utilizan para identificar productos que se venden bien juntos; por ejemplo, los clientes masculinos que compran pañales, viven en el centro y tienen entre 20 y 27 años, generalmente también compran cerveza. De esta forma, los vendedores pueden diagramar la distribución en góndolas (poner estos productos más cerca o más lejos) y planificar su publicidad, entre otras estrategias basadas en este conocimiento.
- **Control de calidad:** aquí la aplicación se basa en procesar la información de seguimiento de los procesos productivos en busca de desviaciones anormales de los mismos, tendencias y demás indicadores de un proceso fuera de control.
- **Retención de clientes:** esta área se avoca a la tarea de conservar los clientes y, si es posible, hacerlos leales. En dicho proceso, las

técnicas de minería nos ayudan a identificar los aspectos característicos de nuestros servicios que han hecho leales a algunos de nuestros clientes para tratar de replicarlos con los demás clientes. Además, nos puede informar sobre las características particulares de nuestros clientes leales para obtener un perfil de los mismos y luego focalizar las campañas de incorporación de nuevos clientes en personas con el mismo perfil, es decir, con una alta probabilidad de tener una relación a largo término con la empresa.

- **Adquisición de nuevos clientes:** Se complementa con la anterior. Es la tarea de conseguir nuevos clientes, es decir, de utilizar minería para determinar estrategias dirigidas a ganar nuevos clientes. La minería nos ayudará a definir la mejor estrategia de adquisición, analizando campañas anteriores y el mejor grupo de personas que será el blanco de la misma.
- **Análisis de la lealtad del cliente:** consiste en minar los datos de los consumidores para extraer modelos de lealtad que muestren el grado de lealtad a un determinado producto o marca. De esta manera, la compañía puede determinar niveles de lealtad y diseñar estrategias diferenciadas para cada nivel.
- **Marketing dirigido:** para realizar marketing dirigido es necesario poder identificar subconjuntos de la población que responderán más probablemente a una campaña publicitaria: en el caso de la adquisición de nuevos clientes, será la población en general; si se trata de lanzamiento de nuevos productos, estarán incluidos también los clientes actuales. Mientras mejor seleccionados estén los miembros del subconjunto elegido, más efectiva será la campaña y esto se traducirá en mejores beneficios y menores costos debido a la focalización de los recursos.
- **Prevención de fraude:** uso de la minería de datos para detectar patrones de fraude.
- **Administración del riesgo:** las aseguradoras deben poder calcular en forma precisa los riesgos asociados a cada una de las pólizas que otorgan y verificar que los valores de las mismas se correspondan con los riesgos asociados; es decir que no se debe sobrevalorar una póliza que implica un riesgo pequeño y, de la misma forma, tampoco se debe infravalorar una póliza que tiene un riesgo muy alto. Muchos de los factores que afectan el riesgo asociado a un evento son claros, pero pueden existir algunas relaciones más sutiles, no intuitivas, entre las variables que son difíciles de discernir si no se les aplican herramientas de análisis más perfeccionadas. Las técnicas modernas de minería ofrecen un modelo más preciso y

más eficiente que las tecnologías anteriores. Los algoritmos de minería permiten echar luz sobre tendencias y relaciones que no son evidentes en los grandes cúmulos de datos, y las nuevas técnicas gráficas permiten visualizar complejos modelos de información que facilitan el análisis y la comparación. Las técnicas de minería ayudan a los aseguradores a segmentar sus clientes para poder tratarlos de forma diferenciada y dividirlos en subconjuntos con un riesgo asociado a cada grupo.

## Operaciones de Minería de Datos

Entre las principales operaciones de Minería de Datos, podemos mencionar las siguientes:

- Modelado Predictivo
- Análisis de Relaciones
- Segmentación de Bases de Datos
- Detección de Desviaciones

## Técnicas de Minería de Datos

Antes de poder diseñar buenos modelos predictivos, es necesario en primer lugar llegar a una comprensión (entiéndase por el conocimiento) de los datos.

En primera instancia, el proceso se inicia recuperando una variedad de resúmenes numéricos (incluyendo estadísticas descriptivas tales como: promedios, desviaciones estándar y otras), como así también observando cuidadosamente la correspondiente distribución de los datos.

En algunos casos, incluso, es deseable producir tabulaciones cruzadas (o *pivot tables*) para datos de carácter multidimensional.

Los datos pueden desde ya, ser continuos, caso en el cual pueden asumir cualquier valor numérico (por ejemplo, la cantidad vendida); o discretos, incluidos dentro de clases discretas (como azul, verde, rojo, etc.). Los datos discretos pueden ser, además, definidos o categorizados en ordinales, los

cuales tienen un orden significativo (por ejemplo: alto, medio, bajo); o nominales, los que están desordenados (por ejemplo, los códigos postales).

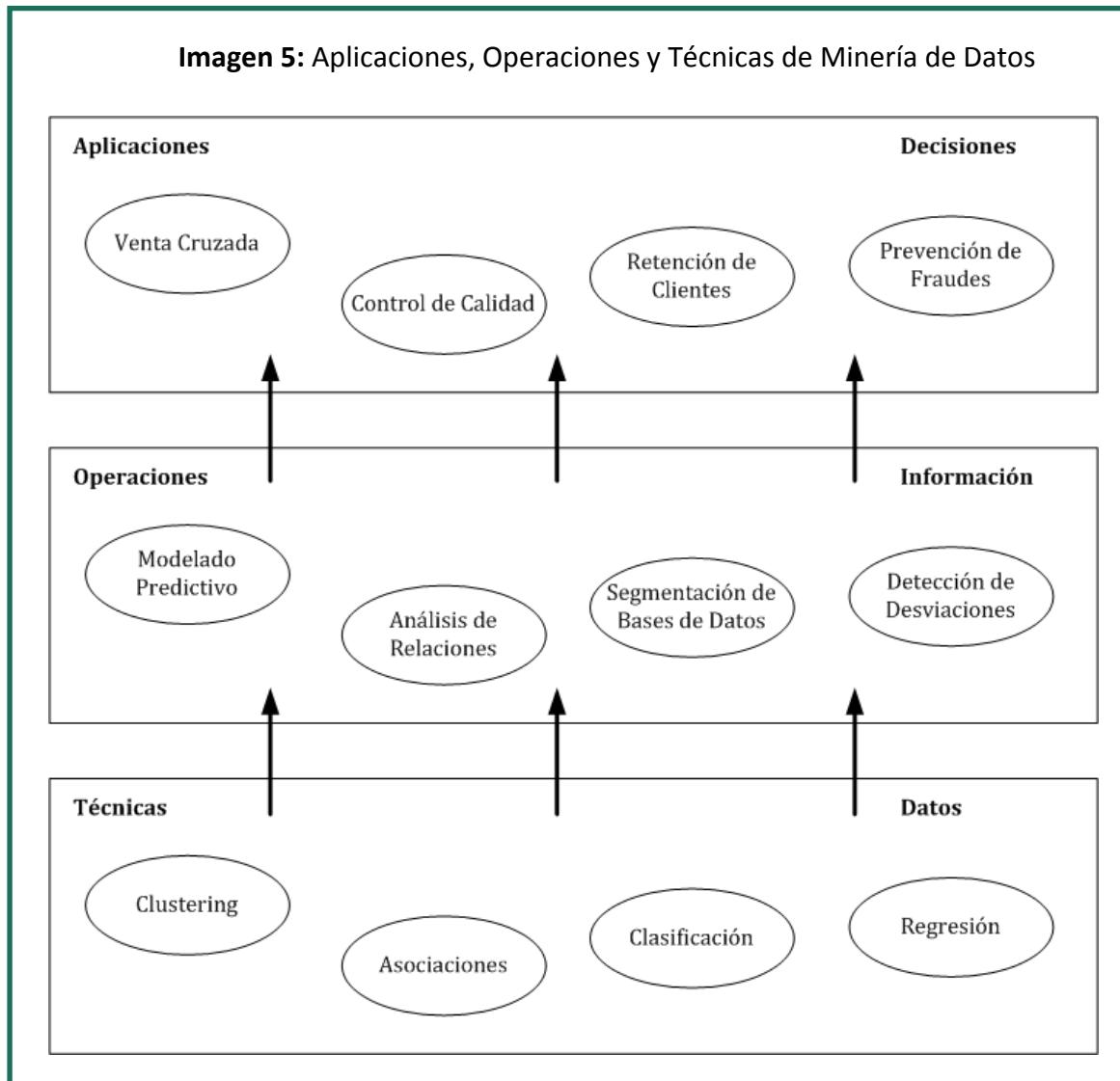
Entre las principales técnicas encontramos:

- **Clustering:** divide o segmenta una base datos en diferentes grupos. El objetivo del *Clustering* es encontrar grupos que son muy diferentes uno del otro y cuyos miembros son muy similares entre sí. A diferencia de la clasificación (Minería de Datos Predictiva), en el Clustering uno no es capaz de conocer los clusters a priori, o bien por cuáles atributos los datos serán segmentados. Consecuentemente, alguien que tenga un gran conocimiento del negocio es quien debe interpretar los clusters.
- **Asociaciones:** se trata de una aproximación descriptiva para la exploración de datos que puede ayudar en la identificación de relaciones entre valores en una base de datos. Las dos aproximaciones más comunes para el análisis de relaciones son: el descubrimiento de asociación y el descubrimiento de secuencia. El primero encuentra reglas sobre ítems que aparecen conjuntamente en un evento como una transacción de compra. El análisis de la canasta de mercado (ó simplemente conocido como *análisis del 'changuito'*) es un ejemplo bastante conocido de descubrimiento de asociación. Por su parte, el descubrimiento de secuencia es muy similar al anterior, con la diferencia de que se analiza una secuencia que es una asociación relacionada a lo largo del tiempo.
- **Clasificación:** los problemas de clasificación ayudan a identificar las características que indican el grupo al cual pertenece cada caso específico. Este patrón, puede ser usado tanto para comprender los datos existentes como para predecir el comportamiento que tendrán las nuevas instancias. Por ejemplo, puede ser deseable predecir si los individuos serán clasificados como aquellos que probablemente responderán a un envío de correo o se someterán a un procedimiento quirúrgico, entre otros. La minería de datos crea modelos de clasificación examinando los datos ya clasificados (llamados *casos*), e inductivamente encuentra un patrón de predicción.
- **Regresión:** usa los valores existentes para pronosticar qué otros valores habrá. En el caso más simple, la regresión emplea técnicas estadísticas tales como la regresión lineal.
- **Series de tiempo:** los pronósticos de series de tiempo predicen valores desconocidos en el futuro, basados sobre una serie variante en el tiempo de predictores. Al igual que la Regresión (explicada

anteriormente), usa resultados conocidos para guiar sus predicciones.



**Imagen 5:** Aplicaciones, Operaciones y Técnicas de Minería de Datos



Fuente: Elaboración Propia

# Balanced Scorecard

“Cuando una persona puede medir aquello sobre lo que está hablando y expresarlo con números, sabe alguna cosa sobre la cuestión; pero cuando no puede medirlo, cuando no puede expresarlo con números, lo que sabe es escaso e insatisfactorio”

Lord Kelvin (en Paul Niven, 2002, p. 23)

## La insuficiencia de los Indicadores Financieros

Paul Niven sostiene en su descripción del Balanced Scorecard que “desde que existen las organizaciones empresariales, el método tradicional para medir los resultados ha sido fijarse en los aspectos financieros” (2002, p. 29). Sin embargo, los indicadores financieros no pueden ser la única forma de medir el desempeño de una empresa. En los últimos años han aparecido numerosas críticas al uso excesivo de indicadores pura y estrictamente financieros:

- Los indicadores financieros muchas veces no son del todo compatibles con la realidad empresarial actual en el sentido de que muchas actividades de valor agregado no se encuentran reflejadas en los activos fijos y tangibles (valores monetarios) de una empresa.
- Los indicadores financieros proporcionan una excelente revisión histórica de los resultados pasados de la organización, pero no tienen poder de predicción para el futuro (por ejemplo, si la empresa fue o no rentable el año pasado no significa que sí lo será el próximo año).
- Tendencia a reforzar los silos funcionales. Los estados financieros generalmente están discriminados por área o departamento funcional de la empresa, pero no tienen en cuenta el valor y el costo asociado a las interrelaciones entre los departamentos.
- Sacrificio del pensamiento a largo plazo. Muchas veces, los esfuerzos por la reducción de costos pueden ser muy útiles para mejorar determinados indicadores financieros en el corto plazo, pero pierden de vista el largo plazo y la capacidad para crear valor

agregado en la empresa (ejemplo: reduciendo el presupuesto de investigación y desarrollo).

- Los indicadores financieros no son los adecuados para muchos niveles de la empresa. Los empleados de todos los niveles de la empresa necesitan datos sobre resultados con los que puedan trabajar.

## Concepto de Cuadro de Mando Integral

Como sostiene Paul Niven (2002, p. 33), “se necesita un sistema que equilibre la exactitud histórica de las cifras financieras con los impulsores de los resultados futuros, al mismo tiempo que ayude a las empresas a poner en marcha sus estrategias diferenciadoras”.

En el año 1992, Robert Kaplan y David Norton plantearon el concepto de **Cuadro de Mando Integral** (BSC, *Balanced Scorecard* en Inglés).

El Balanced Scorecard no es más que un conjunto cuidadosamente seleccionado de indicadores derivados de la estrategia de una empresa, es decir, es una herramienta para medir el desempeño de la empresa.

## Perspectivas

El Balanced Scorecard está compuesto de cuatro grandes perspectivas:

- Clientes
- Procesos Internos
- Aprendizaje y Crecimiento
- Financiera

## Perspectiva Clientes

Según Paul Niven (2002, p. 38), “al elegir las medidas (indicadores) que formarán parte de la perspectiva del cliente dentro del cuadro de mando, las empresas deben responder a dos preguntas fundamentales: ¿Quiénes

son nuestros clientes? y ¿cuál es nuestra proposición de valor al servirlos?".

Entre los indicadores de clientes generalmente usados en las organizaciones, podemos encontrar los siguientes:

- Satisfacción de los clientes
- Fidelidad de los clientes
- Cuota de mercado
- Quejas de los clientes
- Quejas resueltas al primer contacto
- Tasa de rentabilidad
- Tiempo de respuesta por solicitud de cliente
- Precio directo
- Precio en relación con la competencia
- Costo total para el cliente
- Duración media de la relación
- Clientes perdidos
- Retención de clientes
- Tasa de adquisición de clientes
- Clientes por empleado
- Porcentaje de ingresos por nuevos clientes
- Número de clientes
- Ventas anuales por cliente
- Tasa de ganancia (ventas cerradas / contactos de ventas)
- Visitas de clientes a la empresa
- Horas pasadas con los clientes
- Costo comercial como porcentaje de las ventas
- Número de anuncios publicados

- Número de propuestas hechas
- Reconocimiento de marca
- Tasa de respuesta
- Número de ferias con participación
- Volumen de ventas
- Gastos compartidos por cliente objetivo con clientes
- Ventas por cada canal
- Tamaño medio económico de los clientes
- Gastos por servicios a los clientes por cliente
- Rentabilidad de los clientes
- Frecuencia (numero de transacciones de venta)

(Niven, 2002, p. 174)

## Perspectiva Procesos Internos

En esta perspectiva, de acuerdo con Paul Niven (2002, p. 39), “se identifican los procesos claves en los que la empresa debe destacar para continuar añadiendo valor para los clientes y finalmente para los accionistas”.

Para poder satisfacer a los clientes, es necesario que los procesos de negocio internos de la organización funcionen eficaz y eficientemente y de esa manera cumplir con los objetivos de la empresa.

Según Paul Niven (2002, p. 39), “nuestra tarea en esta perspectiva es identificar esos procesos y desarrollar las mejoras medidas (indicadores) posibles con las que hacer el seguimiento de nuestros avances”.

Entre los indicadores de procesos internos generalmente usados en las organizaciones podemos encontrar los siguientes:

- Costo medio por transacción
- Entrega a tiempo

- Tiempo de espera medio
- Rotación de inventario
- Emisiones medioambientales
- Gasto de investigación y desarrollo
- Participación de la comunidad
- Patentes pendientes
- Edad media de las patentes
- Relación productos nuevos / oferta total
- Falta de existencias
- Tasas utilización mano de obra
- Tiempo de respuesta a solicitudes de clientes
- Porcentaje de defectos
- Repetición del trabajo
- Disponibilidad base de datos clientes
- Momento de equilibrio
- Mejora de los tiempos cíclicos
- Mejoras continuas
- Reclamaciones de garantías
- Identificación usuario destacado
- Productos y servicios en la red
- Tasa de rentabilidad interna de proyectos nuevos
- Reducción desperdicios
- Utilización del espacio
- Frecuencia de compras devueltas
- Tiempo muerto
- Exactitud de la planificación

- Tiempo necesario para salir al mercado de nuevos productos / servicios
- Introducción de nuevos productos
- Número de historias positivas en los medios

(Niven, 2002, p. 183)

## Perspectiva Aprendizaje y Crecimiento

En esta perspectiva, de acuerdo con Paul Niven (2002, pp. 39-40), “si se quieren alcanzar resultados ambiciosos con respecto a los procesos internos, los clientes y también los accionistas, ¿qué se puede hacer? Las medidas concernientes a la perspectiva de aprendizaje y crecimiento son verdaderos facilitantes de las otras tres perspectivas”.

Debido a lo anterior, también suele decirse que esta perspectiva representa de algún modo los “cimientos” sobre los cuales se estructura el edificio organizacional.

Entre los indicadores de aprendizaje y crecimiento generalmente usados en las organizaciones, podemos encontrar los siguientes:

- Participación de los empleados en asociaciones profesionales o comerciales
- Inversión en formación por cliente
- Promedio años de servicio
- Porcentaje de empleados con estudios avanzados
- Número de empleados con formación cruzada
- Ausentismo
- Tasa de rotación
- Sugerencias de los empleados
- Satisfacción de los empleados
- Participación en planes de propiedad de acciones
- Accidentes y tiempo perdido

- Valor añadido por empleado
- Índice de motivación
- Número de solicitudes de empleo pendientes
- Tasas de diversidad
- Índice de empowerment (número de directivos)
- Calidad del entorno laboral
- Calificación de las comunicaciones internas
- Productividad de los empleados
- Números de cuadros de mando producidos
- Promoción de la salud
- Horas de formación
- Tasa de cobertura de competencias
- Realización de metas personales
- Oportuna conclusión de valoración de actividades
- Desarrollo de liderazgo
- Planificación de la comunicación
- Accidentes dignos de mención
- Porcentaje de empleados con ordenadores
- Ratio de información estratégica
- Tareas interfuncionales
- Gestión del conocimiento
- Violaciones de la ética

(Niven, 2002, p. 191)

## Perspectiva Financiera

Los indicadores financieros siguen siendo, desde luego, un componente central del Balanced Scorecard, pero ya no son las únicas métricas que usaremos.

De acuerdo con Paul Niven (2002, p. 40), “las medidas de esta perspectiva nos dicen si la ejecución de nuestra estrategia, detallada a través de medidas elegidas en las otras perspectivas, nos está llevando a resultados finales mejores”.

Así, los indicadores financieros tradicionales se encuentran en esta perspectiva.

Entre los indicadores financieros generalmente usados en las organizaciones, podemos encontrar los siguientes:

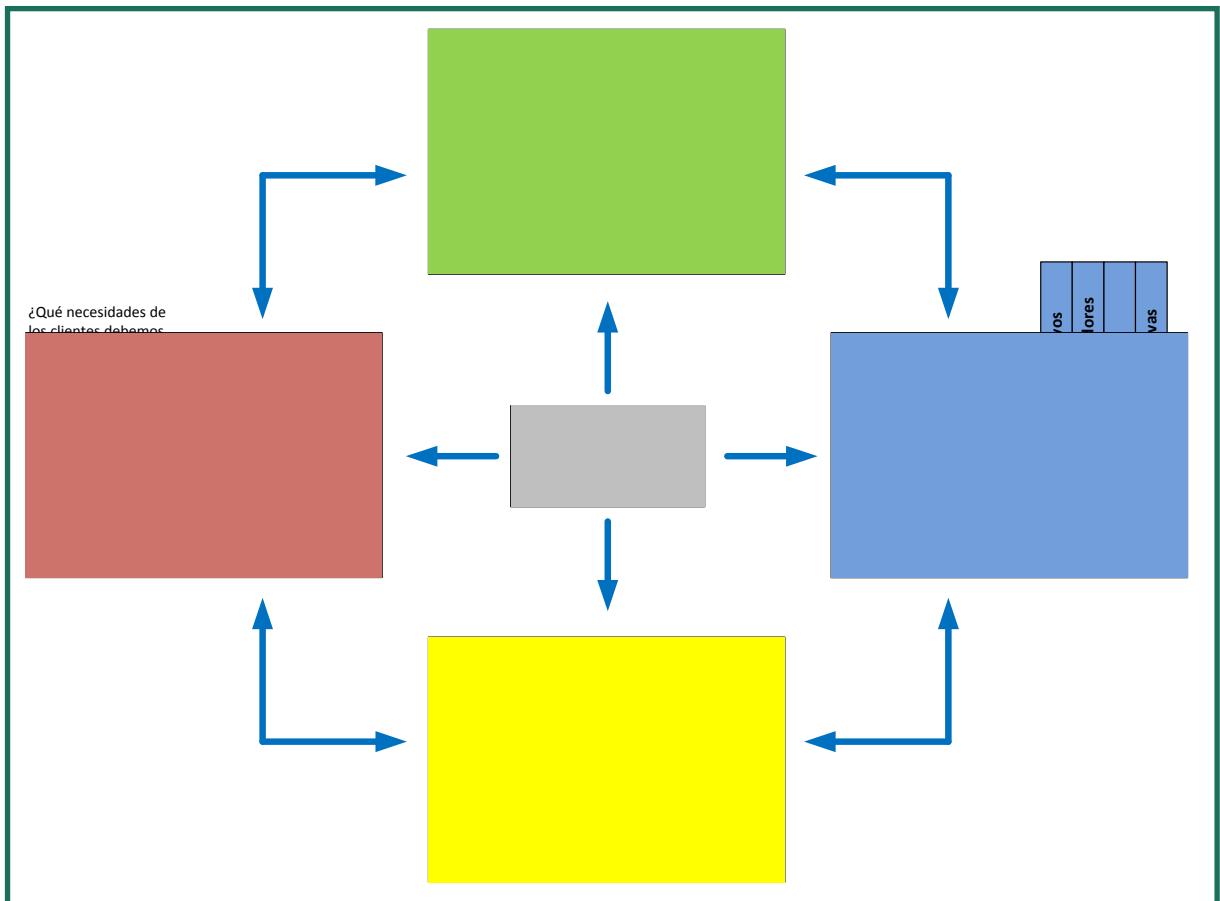
- Activo total
- Activo total por empleado
- Beneficio como % del activo total
- Rentabilidad del activo neto
- Rentabilidad del activo total
- Ingresos / activo total
- Margen bruto
- Beneficio neto
- Beneficio como % de las ventas
- Beneficio por empleado
- Ingresos
- Ingresos por productos nuevos
- Ingresos por empleado
- Rentabilidad de los recursos propios (ROE)
- Rentabilidad del capital empleado (ROCE)
- Rentabilidad de la inversión (ROI)

- Valor económico añadido (EVA)
- Valor añadido de mercado (MVA)
- Valor añadido por empleado
- Tasa de crecimiento compuesta
- Dividendos
- Valor de mercado
- Precio de las acciones
- Mix de accionistas
- Fidelidad de los accionistas
- Flujo de caja
- Costos totales
- Calificación crediticia
- Deuda
- Relación capital ajeno / capital propio
- Intereses ganados
- Días ventas en cuentas a cobrar
- Facturación cuentas a cobrar
- Días en cuentas a pagar
- Días en inventario
- Ratio rotación de existencias

(Niven, 2002, p. 164)



**Imagen 6:** Perspectivas del Balanced Scorecard



Fuente: Elaboración propia en base a Paul Niven (2002, p. 37)

## Equilibrio en el Cuadro de Mando Integral

Como destaca Paul Niven (2002, pp. 47-48), el equilibrio en el sistema del Balanced Scorecard se refiere a tres aspectos:

- “Equilibrio entre indicadores financieros y no financieros.”
- Equilibrio entre constituyentes internos y externos de la empresa.
- Equilibrio entre indicadores posteriores (pasados) y futuros de los resultados”.

## Desarrollo del Cuadro de Mando Integral

Los pasos principales para el desarrollo de un Balanced Scorecard son los siguientes:

- **Paso 1:** Reunir y distribuir material informativo de fondo.
- **Paso 2:** Desarrollar o confirmar misión, valores, visión y estrategia.
- **Paso 3:** Entrevistarse con la dirección.
- **Paso 4:** Desarrollar objetivos y medidas (indicadores) en cada una de las perspectivas del Cuadro de Mando Integral.
- **Paso 5:** Desarrollar relaciones de causa – efecto.
- **Paso 6:** Establecer metas para las medidas (indicadores).
- **Paso 7:** Desarrollar el plan en marcha para implementar el Cuadro de Mando Integral.

(Niven, 2002, pp. 94-96)

Desde luego, de los pasos anteriores debe quedar claro que la Misión, Visión, Valores y la Estrategia de la empresa seguramente ya existen.

El trabajo se concentrará, por lo tanto, en definir los **objetivos de resultados** para cada una de las cuatro perspectivas. Como sostiene Paul Niven (2002, p. 149): “las declaraciones de objetivos son declaraciones concisas que describen las cosas concretas que hay que hacer bien para poder implementar la estrategia con éxito”. Dichos objetivos deberán estar alineados con la Visión y Misión empresarial.

Y por cada objetivo, posteriormente, debemos definir los indicadores. Según Niven (2002, pp. 157-158), “las medidas (o indicadores) de los resultados son las herramientas que usamos para determinar si estamos cumpliendo con nuestros objetivos”.

Es decir, los indicadores deben ser cuantificables, con el objeto de poder medir su grado de avance hacia el cumplimiento de los objetivos pautados.

## Modelización del Negocio

Respecto a esta temática, Josep Lluís Cano expresa lo siguiente:

Cuando definíamos Business Intelligence [...] decíamos que se refiere a un área concreta del negocio: clientes, productos, costes, etc. Cuando nos referimos a un área de negocio siempre pretendemos “medir” un resultado, bien sean ingresos, costos, ventas, etc. La razón que nos lleva a ello está ampliamente desarrollada en la literatura empresarial y se podría resumir como “aquellos que no se mide, no se puede gestionar”. El primer paso, por tanto, es definir qué queremos medir, cómo lo vamos a hacer. Para ello debemos definir cuál es nuestro modelo de negocio y cuáles son las métricas que vamos a utilizar.

Construir un modelo nos permite analizar qué está sucediendo y para poder construirlo debemos documentar, probar, y desarrollar nuestras teorías acerca de cómo funciona el negocio. Los modelos nos ayudan a experimentar de qué manera afectarán los cambios que introduzcamos al resultado.

Es importante construir modelos de negocio, ya que en algunos casos no siempre la mejor solución para un departamento es la mejor para toda la organización.

Hablamos de “silos de información” cuando entre distintos departamentos no fluye la información necesaria, bien para la gestión o bien para el análisis, dando lugar tanto a problemas de operaciones, como de optimización del negocio.

Los “silos de información” pueden ser originados por la no integración de los sistemas de información de los distintos departamentos o por razones políticas dentro de nuestra organización.

Los modelos de negocio son simplificaciones de la realidad que nos sirven para comprender qué está sucediendo.

Para definirlos podemos acudir a distintas metodologías: Contabilidad Analítica o de Costes, EFQM (European Foundation for Quality Management), SixSigma, Análisis de procesos, Modelos Financieros, Análisis de ratios, etc.

Si el modelo de negocio está bien definido nos permitirá responder preguntas clave de la gestión de nuestra organización.

(Cano, 2007, pp. 59-61 y 65)

## Indicadores Claves del Negocio (KPI)

Los Indicadores Claves del Negocio (KPI, Key Performance Indicator en Inglés) representan una herramienta esencial para saber si la organización está alcanzando o no sus objetivos.

Una vez que han analizado su misión, han identificado los grupos de poder y han definido sus objetivos, las organizaciones necesitan un sistema para medir su progreso hacia la consecución de los objetivos. Los KPI son los instrumentos adecuados para llevarlo a cabo.

Los KPI deben ser cuantificables y deben medir las mejoras en aquellas actividades que son críticas para conseguir el éxito de la organización. Los KPI deben estar relacionados con los objetivos y con las actividades fundamentales de nuestra organización (aquellas que nos permiten obtener los resultados).

Distintas empresas de un mismo sector pueden tener distintos KPI's en función de sus modelos de negocio, objetivos o su propia idiosincrasia.

Los KPI's que escojamos deben:

- Reflejar los objetivos del negocio.
- Ser críticos para conseguir el éxito.
- Ser cuantificables y comparables.
- Permitir las acciones correctivas.

Si establecemos KPI's por departamentos deberán estar alineados entre ellos y con los objetivos de la organización.

Los KPI's deben ser establecidos involucrando a los responsables de cada una de las áreas de la organización. Debemos seleccionar aquellos KPI's que estén relacionados con la consecución de los

resultados en la organización, es decir, aquellos que son esenciales para conseguir los objetivos.

Debemos escoger un número reducido de KPI's para facilitar que los distintos miembros de nuestra organización nos centremos en conseguirlos. A tal fin, debemos darles un nombre, una definición, establecer como calcularlos y los valores a conseguir.

(Cano, 2007, pp. 65-67)

## Scorecards

Como vimos en la unidad 1 (Módulo 1), una de las herramientas de visualización de Business Intelligence consiste en los denominados 'Scorecards'. Los mismos están basados en el concepto de Balanced Scorecard.

Consisten, básicamente, en herramientas o pantallas de Tableros de Comando en donde se refleja la estructura del Cuadro de Mando Integral: la visión y misión de la empresa, objetivos, indicadores, etc. En un Scorecard puede analizarse la evolución de cada uno de los KPI definidos dentro de un Mapa Estratégico de Balanced Scorecard.

A diferencia de un *Dashboard* (en donde puede verse la evolución de distintos indicadores), en el Scorecard los KPI están clasificados por objetivos a los que pertenecen y éstos enmarcados dentro de cada una de las cuatro perspectivas.

Sin lugar a dudas, el desarrollo de un Scorecard es mucho más complejo que un simple Dashboard puesto que implica haber llevado adelante previamente una estrategia de definición del Cuadro de Mando Integral de la organización.



## Bibliografía de Referencias

- **Cano, J. L.** (2007). *Business Intelligence: Competir con Información*. España: Banesto Fundación Cultural y ESADE.
- **Inmon, W.** (2005). *Building the Data Warehouse* (4º Edición). Estados Unidos: Wiley Publishing.
- **Kimball Ralph, R. M.** (2013). *The Data Warehouse Toolkit* (3º Edición). Estados Unidos: Wiley Publishing.
- **Niven, P.** (2002). *El Cuadro de Mando Integral Paso a Paso*. España: Ediciones Gestión 2000.
- **Two Crows Corporation** (1999). *Introduction to Data Mining and Knowledge Discovery* (3º Edición). Recuperado el 5 de Mayo de 2014:  
<http://www.stanford.edu/class/stats315b/Readings/DataMining.pdf>