

Resumen de la 1º clase:

- **ERP:** Enterprise resource planning. Es un software que sirve para administración de negocios. Las compañías lo utilizan para recolectar, almacenar, administrar e interpretar información de diversas actividades de negocios (utiliza una bases de datos convencionales).
- **BI:** Business Intelligence, inteligencia empresarial, inteligencia de negocios.
La inteligencia de negocios es término popularizado por Howard Dresner, para describir un conjunto de conceptos y métodos para mejorar la toma de decisiones de negocio mediante el uso de sistemas de apoyo computarizados basado en hechos.
Actualmente, en las actividades diarias de cualquier organización, se generan datos como producto secundario, que son el resultado de todas las transacciones que se realizan.
Aquí es donde la BI entra en juego, ya que al obtener conocimiento del negocio una vez capturada la información de todas las áreas en la empresa, es posible establecer estrategias y determinar cuáles son las fortalezas y las debilidades.
Por lo tanto, BI provee de un conjunto de tecnologías capaces de satisfacer a una gran gama de usuarios, junto a sus necesidades específicas en cuanto al análisis de información, para convertir datos en información que enriquezcan sus decisiones.
Para Daniel Power, los vendedores y analistas de sistemas de información tienden a usar el término BI para una categoría de herramientas de software que pueden ser utilizados para extraer y analizar los datos de las bases de datos corporativas.
Debido a que para llevar a cabo BI, es necesario gestionar datos guardados en diversos formatos, fuentes y tipos, para luego depurarlos e integrarlos, además de almacenarlos en un solo destino o base de datos que permita su posterior análisis y exploración, es imperativo y de vital importancia contar con un proceso que satisfaga todas estas necesidades. Este proceso se denomina Data Warehousing.
- **DSS:** Decision Support System, sistema de soporte a las decisiones.
No hay una definición universalmente aceptada de lo que es un DSS. En general, podemos decir que un DSS es un sistema informático utilizado para servir de apoyo, más que automatizar, el proceso de toma de decisiones.
El autor Daniel Power, propone utilizar el término Data-driven DSS, es lugar de Business Intelligence. Pues, para él, un sistema de inteligencia de negocios es un DSS basado datos (Data-driven DSS) que soporta principalmente la consulta de una base de datos histórica y la producción de informes periódicos.
Data-driven DSS es un tipo de DSS, que hace hincapié en el acceso y la manipulación de una serie temporal interna de la empresa, y a veces, datos externos.
Una serie temporal es una secuencia de datos, observaciones o valores, medidos en determinados momentos y ordenados cronológicamente.
Un sistema de archivos simples, que se accede por herramientas de consulta y recuperación es un Data-driven DSS, en su nivel más elemental de funcionalidad.
Entonces, para Power, BI se refiere a un propósito específico de algunos Data-driven DSS.
Es importante aclarar, que un Data Warehouse no es un DSS. Sino que, es una de las componentes principales de un sistema BI, o Data-driven DSS.
- **DW, ¿Qué es?:** Data Warehouse, almacén de datos.
Bill Inmon fue uno de los primeros autores en escribir sobre el tema de los almacenes de datos, define un DW en términos de sus características.
Ralph Kimball es otro conocido autor en el tema de los data warehouse, define un almacén de datos como: "una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis".
Cuando se define DW en términos de sus características, se dice no volátil por el hecho de

que los datos no son transitorios.

Un DW se trata, sobre todo, de un expediente completo de una organización, más allá de la información transaccional y operacional. Almacenado en una base de datos diseñada para favorecer el análisis y la divulgación eficiente de datos.

- **DW, Arquitectura:**

La preparación de los datos, a menudo llamada la gestión de datos, implica la adquisición de datos y su transformación en información, y en última instancia, se entrega esa información al front room. El acceso a los datos, por parte de los usuarios, está restringido a el front room.

El entorno de un almacén de datos incluye varios componentes, cada uno con su propio conjunto de diseños, técnicas, herramientas y productos. Es importante recordar que ninguna de estas cosas por sí solo constituye un almacén de datos.

- **Bases de datos transaccionales:** Una base de datos transaccional suele ser una base de datos relacional, cuyo SGBD (Sistema de Gestión de Bases de Datos) permite transacciones, es decir que las órdenes que se ejecutan forman una unidad de trabajo; son indivisibles o atómicas. Esto permite, mantener la integridad de los datos, haciendo que estas transacciones no puedan finalizar en un estado intermedio.

- **OLTP: Procesamiento de Transacciones En Línea (OnLine Transaction Processing)** es un tipo de procesamiento que facilita y administra aplicaciones transaccionales, usualmente para entrada de datos y recuperación y procesamiento de transacciones.

Las aplicaciones OLTP tienen un alto rendimiento en la inserción o actualización intensiva de una base de datos. Estas aplicaciones se utilizadas simultáneamente por centenares de usuarios. Las aplicaciones OLTP tienen como objetivos clave la disponibilidad, velocidad, concurrencia y recuperación.

Un cajero automático de un banco es un ejemplo de una aplicación de procesamiento de transacciones comerciales.

En definitiva, los sistemas OLTP son bases de datos orientadas al procesamiento de transacciones. Una transacción genera un proceso atómico (que debe ser validado con un *commit*, o invalidado con un *rollback*), y que puede involucrar operaciones de inserción, modificación y borrado de datos.

- **OLTP vs. DW (OLAP):**

Los sistemas OLTP, tienen las siguientes desventajas:

1. Gran rigidez a la hora de extraer datos, de manera que el usuario tiene que ceñirse a los informes predefinidos que se configuraron en el momento de la implantación, y que no siempre responden a sus dudas reales.
2. Necesidad de conocimientos técnicos. Para la generación de nuevos informes o métricas suele resultar ineludible acudir al departamento técnico, solicitando una consulta adecuada para interrogar la base de datos.
3. Largos tiempos de respuesta, ya que las consultas complejas de datos suelen implicar la unión de tablas operacionales de gran tamaño, lo que se traduce en una incómoda espera que dificulta la fluidez del trabajo.
4. Deterioro en el rendimiento del SI. Cuando la base de datos consultada, para generar informes o ratios de negocio, es la misma que la que soporta el operativo de la empresa, el funcionamiento del sistema puede degradarse hasta afectar y paralizar a todos los usuarios conectados.
5. Falta de integración que implica islas de datos. Muchas organizaciones disponen de

múltiples sistemas de información, incorporados en momentos distintos, para resolver problemáticas diferentes. Sus bases de datos no suelen estar integradas, lo que implica la existencia de islas de información.

6. Datos erróneos, obsoletos o incompletos. El tema de la calidad de los datos siempre es considerado como algo importante, pero esta labor nunca se lleva al extremo de garantizar la fiabilidad de la información aportada.
7. Problemas para adecuar la información al cargo del usuario. No se trata de que todo el mundo tenga acceso a toda la información, sino de que tenga acceso a la información que necesita para que su trabajo sea lo más eficiente posible.
8. Ausencia de información histórica. Los datos almacenados en los sistemas operacionales están diseñados para llevar la empresa al día, pero no permiten contrastar la situación actual con una situación retrospectiva de años atrás.

Los sistemas OLAP son bases de datos orientadas al procesamiento analítico. Este análisis suele implicar, generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil: tendencias de ventas, patrones de comportamiento de los consumidores, elaboración de informes complejos... etc. Este sistema es típico de los DataMarts.

- **Data mart:**

Top-Down: Como se puede apreciar, el DW es cargado a través de procesos ETL y luego este alimenta a los diferentes DM, cada uno de los cuales recibirá los datos que correspondan al tema o departamento que traten.

Esta forma de implementación cuenta con la ventaja de no tener que incurrir en complicadas sincronizaciones de hechos, pero requiere una gran inversión y una gran cantidad de tiempo de construcción.

Bottom-Up: Los DM se cargan a través de procesos ETL, los cuales suministrarán la información adecuada a cada uno de ellos. En muchas ocasiones, los DM son implementados sin que exista el DW, ya que tienen sus mismas características pero con la particularidad de que están enfocados en un tema específico. Luego de que hayan sido creados y cargados todos los DM, se procederá a su integración con el depósito.

La ventaja que trae aparejada este modelo es que cada DM se crea y pone en funcionamiento en un corto lapso de tiempo y se puede tener una pequeña solución a un costo no tan elevado. Luego que todos los DM estén puestos en marcha, se puede decidir si construir el DW o no. El mayor inconveniente está dado en tener que sincronizar los hechos al momento de la consolidación en el depósito.

- **Hechos:** Son las medidas de negocios (o indicadores de negocios) a los que el usuario final quiere observar y sobre los cuales quiere tener conocimientos.
- **Dimensiones:** Éstas son las que le dan contexto a lo que se quiere observar (hechos). Por ejemplo, si uno quiere observar el hecho total vendido, las posibles dimensiones que le dan contexto a esa métrica son el tiempo, el negocio y la ciudad en que se ubica. Si bien cada dimensión hace referencia a un concepto específico que aporta al contexto en el que se quiere observar el hecho, éste puede estructurarse con un orden jerárquico. Por ejemplo, si tenemos una dimensión país, dentro de la misma podríamos tener provincia y ciudad como dos niveles de granularidad. Las dimensiones contienen datos cualitativos.
- **Esquemas dimensionales:** Son los esquemas de bases de datos que separan físicamente las medidas que cuantifican el negocio (**hechos**) de los elementos que describen el contexto en que queremos observar esas medidas cuantificadas (**dimensiones**). El esquema dimensional

puede ser físico o lógico.

- Un esquema dimensional **físico** generalmente los objetos que contiene son en realidad tablas de base de datos.
- Un esquema dimensional **lógico** los hechos y las dimensiones se representan como entidades y atributos independientes a una base de datos y por lo tanto, se pueden transformar en un esquema dimensional físico para cualquier base de datos.

- **Esquema de estrella:** En este tipo de esquemas nos encontraremos con una tabla de hechos, rodeada de tablas de dimensiones. En la tabla de hechos tendremos un campo por cada dimensión, el cual hace referencia a la clave primaria de la misma (por medio de éstos luego se podrá recuperar la información necesaria para cada hecho) y las métricas que quieren ser evaluadas. La combinación de estas forman, generalmente, la clave primaria de la tabla de hechos. Por otra parte, cada dimensión tendrá su campo clave y otros campos en los que se almacenan las correspondientes descripciones que son de utilidad en el ambiente de negocios. EJEMPLO
- **Esquema de copo de nieve:** Este otro tipo de esquemas, es similar al esquema de estrella salvo por cierto detalle, las dimensiones pueden, a su vez, estar conectadas con otras tablas de dimensiones. Es decir, tendremos una tabla de hechos rodeada de tablas de dimensiones que nuevamente están conectadas con otras tablas de dimensiones con una relación muchos a uno. Cada una de las dimensiones conectadas con una dimensión representan un nivel jerárquico para la dimensión en si misma (por ejemplo, para la dimensión país, podríamos tener una tabla para país, otra para provincia y otra para ciudad). En el ejemplo anterior se ve claramente que una dimensión se implementa con mas de una tabla. La razón de esto es buscar una normalización de las tablas, con lo cual se logra reducir espacio de almacenamiento al eliminar redundancia en los datos. Al tener una dimensión (la cual sería representada por una tabla en la base de datos) representada por varias tablas, se logra facilitar el mantenimiento de la dimensión.
- **Esquema de constelación:** En este tipo de esquemas se compone de varios esquemas estrella y/o de copo de nieve. Pueden tener más de una tabla de hechos y comparten dimensiones. La posibilidad de tener más de una tabla de hechos permite que una de ellas sea la principal y las demás contener diversas sumalizaciones de información contenida en la principal.
- **Dimensión tiempo:** En un DW, la creación y el mantenimiento de una tabla de dimensión Tiempo es obligatoria, y la definición de granularidad y estructuración de la misma depende de la dinámica del negocio que se este analizando. Toda la información dentro del depósito, como ya se ha explicado, posee su propio sello de tiempo que determina la ocurrencia de un hecho específico, representando de esta manera diferentes versiones de una misma situación. Al momento de implementar la dimensión como una tabla en la base de datos se debe tener especial cuidado con el campo clave que se elige ya que esto puede influir mucho en el espacio que se ocupara en disco y los tiempos de respuesta. Las posibilidades son usar un valor de tipo **Date** o un valor de tipo **Entero**:
 - Si usamos el tipo **Date** para el campo clave, costaría mucho espacio en comparación con un entero y relentizarían los tiempos de respuesta ya que para los motores de consulta resulta más costosa la comparación con este tipo de campos que con uno entero.
 - Si elegimos un valor de tipo **Entero** tenemos dos opciones:
 - Utilizar un campo autoincremental, el cual tiene la siguiente forma: 1, 2, 3, ..., 5140, ... con lo cual el costo de almacenamiento en disco se reduciría considerablemente y los tiempos de respuesta bajan ya que la comparación entre enteros no es algo

costoso. Con esta representación se torna confuso con que fecha se está tratando.

- Utilizar un campo de la forma yyyyMMdd, por ejemplo 20150315, se obtienen los mismos beneficios que los recién planteados y se agrega una mejor comprensión a la hora de visualizar los registros.