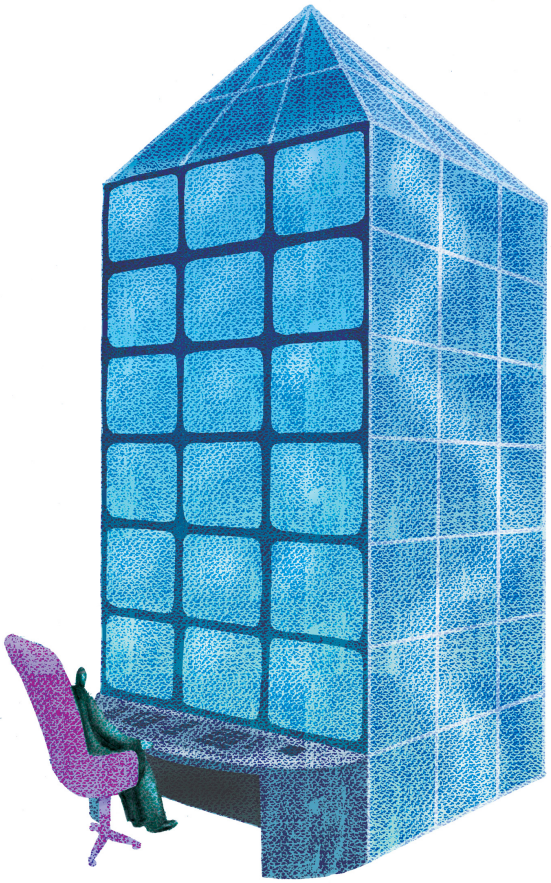
# Generando conocimiento a partir de los datos

Los sistemas de gestión existentes en la empresa (ERP, CRM u otro) captan diariamente grandes cantidades de información de los procesos de negocios. Estos datos y otros, provenientes de distintas fuentes de información, representan la base con la cual será posible generar el conocimiento al interior de la empresa.



El proceso de estructurar la información de manera robusta, consistente y automatizada es fundamental, ya que representa los cimientos sobre los cuales se sustentarán los procesos de BI. La concepción de un buen modelo de información representa uno de los esfuerzos que generan más “rentabilidad” al momento de realizar la explotación de ella.

Podríamos realizar un curso completo sobre **modelamiento de datos**, ya que no es un proceso trivial y requiere de conocimiento técnico. Aunque el especialista en BI no necesariamente debe conocer todos los detalles sobre esta disciplina, es importante que domine los conceptos básicos más relevantes.

No es la pretensión de esta clase abordar complejos detalles técnicos o profundizar ampliamente en su teoría. Para ello existe numerosa documentación para aquellos que requieren ampliar sus conocimientos. Aquí abordaremos los conceptos más importantes que debe conocer el especialista en BI para poder diseñar y explotar la información proveniente de los sistemas de información u otras fuentes.

## Relacionando bases de datos

Los sistemas de gestión están construidos sobre un modelo de datos operacionales. Este diseño del modelo se realiza en forma previa a la implantación de la herramienta y es la base de funcionamiento del sistema.
Para comprender este concepto supondremos un sistema de información muy simple: el sistema **Little ERP**, el cual ha sido diseñado para ingresar información de clientes y registrar sus ventas. Para ello posee dos ventanas donde los usuarios de información interactúan con el sistema: el módulo de clientes y el de ventas.

Es necesario ingresar la siguiente información en cada uno de sus módulos:

| Clientes |
|---|
| Nombre |
| Rut |
| Fecha Nacimiento |
| Dirección |
| Comuna |
| Ciudad |
| Segmento |
| Id Cliente |

| Ventas |
|---|
| Fecha |
| Id Cliente |
| Rut |
| Codigo Producto |
| Descripción Producto |
| Unidades |
| Precio Unitario |
| Total |

Pero, ¿qué sucede al ingresar la información en cada una de las ventanas del sistema?

Normalmente, cada una de las transacciones realizadas sobre el sistema queda guardada en bases de datos que han sido diseñadas para ello. Luego, un nuevo cliente que realiza la compra de un producto quedaría registrado de la siguiente manera en dos tablas distintas:

| Nombre | Rut | Fecha Nacimiento | Dirección | Comuna | Ciudad | Segmento | Id Cliente |
|---|---|---|---|---|---|---|---|
| Pepe Pérez | 123456789 | 23-10-1980 | El Callejón Azul, 1234 | la Florida | Santiago | Alto Valor | 998877 |

| Fecha | Id Cliente | Rut | Codigo Producto | Descripción Producto | Unidades | Precio Unitario | Total |
|---|---|---|---|---|---|---|---|
| 17-10-2011 | 998877 | 123456789 | ME5577 | Zapatillas Mike Running | 1 | 39990 | 39990 |

En los sistemas actuales de información existen cientos de tablas en las que se vuelca la información registrada de cada una de las transacciones realizadas. Estas tablas poseen un diseño que permite guardar la información de manera estructurada y establecer su relación con otras tablas.

El conjunto de bases de datos relacionadas entre sí en una empresa se le denomina **repositorio de datos**. Son los centros de información desde los cuales se alimentará a otras fuentes de información. Son el verdadero corazón del BI.

Siguiendo nuestro ejemplo, las dos tablas de **Clientes** y **Ventas** serían las fuentes de información para el repositorio de la empresa. Las tablas independientes entre sí no representan un repositorio con el cual se podrá trabajar de manera consistente, ágil y robusta. Para lograr aquello es necesario tomar una serie de definiciones sobre ellas que generarán el modelo de datos necesario para poder extraer información fiable del repositorio.

## Buen diseño, buen resultado

La etapa de diseño de su modelo es extremadamente importante. No se debe escatimar tiempo en este proceso. En el ejemplo anterior, hay que resolver algunos temas de importancia:

1) **Formato de los campos:** ¿cuál será el largo del campo nombre? ¿el RUT será numérico o de caracteres? ¿la fecha de nacimiento será formato dd-mm-aaaa? ¿los valores de las ventas podrán contener decimales?
2) **Unicidad de los datos:** ¿podrá haber más de un registro con el mismo ID de cliente en la tabla de clientes? ¿podrá haber más de un registro con el mismo rut? ¿podrá haber dos compras exactamente iguales para un mismo cliente un mismo día y hora?
3) **Relación de Tablas:** ¿cómo se relacionará la tabla clientes con la de ventas? ¿qué campo determina el cruce entre ellas?

Si bien, tal como se señalaba anteriormente, existen muchos conceptos sobre el modelamiento de datos, estos tres puntos son los más relevantes para el especialista de BI y deben ser asimilados como si fuera algo natural y evidente.

Para el especialista BI será de gran utilidad considerar lo siguiente:

**Formato de los campos:** Existen una gran cantidad de formatos distintos para los campos de tablas. Los más utilizados son int (entero), float (punto flotante o decimales), char (caracter), varchar (caracter variable) y fecha. La decisión de cuál elegir dependerá del uso que se le dará y de la regla de minimización de uso de memoria. Por ejemplo, podríamos considerar las ventas como float (que permite decimales) en vez de formato entero, pero si las compras nunca tendrán decimales, no tendría sentido considerar este formato dado que la utilización de memoria es mayor que para el formato "entero".

Por otro lado, es necesario considerar los posibles formatos que tendrá el campo a futuro. En nuestro ejemplo el ID de cliente pareciera tener formato entero, pero ¿habrá ID de cliente que empiece con un cero? ¿será el ID de cliente siempre del mismo largo? Este tipo de consideraciones deben ser tomadas en cuenta al momento de diseñar el formato de cada uno de los campos de la tabla.

**Unicidad de los datos:** Este es un tema que da muchos dolores de cabeza a los usuarios de la información. No es extraño encontrarse con bases que poseen registros totalmente duplicados. Esto es porque el encargado de diseñar no consideró una clave primaria o primary key (PK) sobre la tabla. La clave primaria de una tabla no puede repetirse y la misma base no permitirá el ingreso de registros con duplicidad.

El no considerar una clave primaria es un error bastante común y puede echar por tierra su modelo de datos. Siempre, ¡siempre!, considere que las tablas deben tener una clave primaria. En el ejemplo, en la tabla de clientes la clave primaria debiera ser el ID de cliente y en el caso de la tabla de ventas sería una combinación de Fecha (con hora), ID de Cliente y Código de Producto. Aunque sería mucho más abordable haber creado un campo ID de transacción que fuera único para esta llave o clave primaria.

**Relación de Tablas:** Es evidente que el ejemplo analizado es una gran simplificación de la realidad de una empresa. Normalmente los sistemas de información poseen cientos de tablas que recogen la información de las transacciones realizadas. Luego, es muy importante considerar la relación entre tablas y que ésta sea correcta. En nuestro ejemplo, definir una relación entre la tabla de clientes y la de ventas es relevante a la hora de buscar cruces de información. La relación entre ellas será por el ID de cliente, dado que el éste es la clave primaria de la tabla de clientes y ese campo existe en la tabla de ventas. El resultado de una mala relación entre tablas provocará inconsistencias, duplicidad o imposibilidad de cruces de información.

## Chequeando la integridad de la información

Es muy probable que usted sea usuario de una base de datos en cuyo diseño no ha tenido injerencia. Antes de comenzar con la explotación de esta base, es muy recomendable realizar lo siguiente para conocer la integridad de la información:

1) **Contar con un diccionario de datos:** esto puede acelerar enormemente el conocimiento sobre la base. En proyectos bien ejecutados es práctica obligada la generación de este tipo de documentos.
2) **Conteos y sumas:** al realizar conteos y sumas sobre las tablas, se puede detectar si los resultados tienen concordancia con los números del negocio. Por ejemplo, si el resultado de las ventas de un mes son muy distintos a lo que usted ha visualizado en otros informes, es posible que no esté realizando la consulta correcta o haya problemas con la información contenida en la base.
3) **Revisar resultados históricos:** si bien los resultados sobre un mes en particular pueden ser los correctos, es necesario hacer conteos y sumas sobre las bases históricas para conocer desde cuándo se puede contar con esta exactitud.
4) **Conocer la clave primaria y detección de duplicados:** para conocer la clave primaria si ésta no se encuentra especificada en el diccionario de datos, es necesario visualizar la base y elegir algunos "candidatos". Sobre éstos se deben realizar conteos de manera que el total de registros sobre la base debe ser igual al conteo de todos los distintos valores de la clave primaria. Recuerde que puede haber más de un campo que sea clave primaria. Si una vez detectadas las posibles claves primarias no logra conseguir unicidad de la información, puede concluir que hay problemas en el origen del diseño de la información que es necesario corregir.
5) **Detección de datos faltantes:** es muy posible que detecte datos faltantes en algunos de los campos de la base. Si

fuera así, es necesario conocer la razón de ello. Debe hacerse las siguientes preguntas: ¿significa algo que el campo esté vacío? ¿el vacío en este campo es un valor en sí mismo? ¿se puede interpretar de alguna manera? ¿el dato se puede reconstruir?

Hay que tomar en cuenta que, además de los campos provenientes de los sistemas de información, es posible generar nuevas columnas que no existen en el repositorio fuente. Estos nuevos campos enriquecerán las bases del usuario y harán más ágil su respuesta a determinadas consultas. Por ejemplo, podríamos considerar la creación de un indicador de vigencia del cliente, compras promedio en los últimos tres meses, edad, y muchos otros.

## Cimientos sólidos

El conocimiento correcto de la estructura y contenido de una base es el punto de partida de un proyecto de BI exitoso. El grado de conocimiento de cada una de las tablas que conforman su modelo de información le puede llevar al éxito o al fracaso. Tal como un buen orfebre debe conocer bien las propiedades de los materiales con los que creará su obra, un especialista de BI debe dominar el contenido de la materia prima con la cual realizará la explotación de la información. Cuanto mejor conozca la materia prima, mejor provecho podrá sacar de ella.

Esta etapa del proyecto consume gran parte del tiempo y generalmente existe la tentación de acelerar su término. No apuraríamos irresponsablemente la construcción de los cimientos de nuestra casa sino que tomaríamos todos los resguardos necesarios para tener la seguridad de que estamos construyendo sobre bases sólidas. En el BI no es distinto: construya cimientos sólidos y podrá levantar importantes obras sobre ellos.