

**INTELIGENCIA DE NEGOCIOS, ALMACEN  
DE DATOS Y MODELADO DIMENSIONAL**

# Tabla de contenido

Introducción.....	4
<b>1. Los sistemas de información en las organizaciones</b>	<b>4</b>
Herramientas de negocio.....	6
Sistemas OLTP y sistemas OLAP.....	6
Control de la información.....	8
Retos de la democracia de información.....	8
Inteligencia de negocios.....	9
Historia de la inteligencia de negocios.....	9
Los cinco estilos de la inteligencia de negocios.....	10
Comparación entre tipos de aplicaciones.....	11
Demostración con Pentaho.....	11
Instalación rápida del servidor.....	11
Ingreso al portal de Pentaho.....	12
Instalación de Saiku Analytics.....	13
Crear una vista de análisis con Saiku.....	14
Funciones básicas.....	16
Ordenamiento.....	16
Límites.....	17
Filtros.....	18
Totales y sub-totales.....	19
Gráficos.....	21
Laboratorio.....	22
<b>2. Almacenes de datos</b>	<b>23</b>
Ventajas de un sistema DW.....	24
Arquitectura/componentes de un sistema de BI.....	25
Metáfora del restaurante.....	26
Diseño de un almacén de datos.....	27
Análisis de requisitos y diseño conceptual.....	27
Diseño lógico.....	27
Diseño físico.....	27
Implementación.....	28
<b>3. Modelo de dato dimensional</b>	<b>28</b>
Modelo de dato relacional.....	28
Modelo entidad-relación.....	29
Esquema estrella.....	30
Técnicas de modelado dimensional.....	31
Tablas de hechos.....	31
Aditividad en las medidas.....	32
Características físicas de una tabla de hechos.....	32
Técnicas de las tablas de hechos.....	33
Tablas de hechos transaccionales.....	33
Tablas de hechos periódica.....	33
Tablas de hechos acumulativas.....	33
Tablas de hechos sin hechos.....	36
Tablas agregadas.....	36
Nulos en la tabla de hechos.....	36

Tabla de dimensión.....	37
Calidad de la dimensión.....	37
Jerarquías en las dimensiones.....	38
Técnicas de dimensiones.....	39
Dimensiones Degeneradas.....	39
Dimensiones JUNK.....	39
Dimensiones Snowflake.....	40
Dimensiones Outtrigger.....	41
Dimensiones que cambian lentamente (SCD).....	42
SCD Tipo 0 (Mantener valores originales).....	42
SCD Tipo 1 (Sobrescribir valores).....	42
SCD Tipo 2 (Agregar nueva fila).....	42
SCD Tipo 3 (Agrega un nuevo atributo).....	43
Roles en una dimensión.....	43
Dimensiones multi-valuadas y tablas puentes.....	44
Valores nulos en las dimensiones.....	45
Manejando jerarquías en las dimensiones.....	45
Jerarquías con profundidad fija.....	45
Jerarquías ligeramente des-balanceadas con profundidad variable.....	45
Jerarquías des-balanceadas con tablas puentes.....	46
Extensibilidad del esquema.....	49
Arquitectura en bus del almacén de datos.....	51
Dimensiones conformadas.....	51
Dimensiones encogidas.....	51
Matriz en bus del almacén de datos.....	52
Data marts.....	52

## 4. Fases en el diseño dimensional 53

Los procesos de negocio.....	53
El grano.....	53
Diseño del modelo dimensional de Steel Wheels.....	54
Seleccionar proceso de negocio.....	54
Declarar el grano.....	55
Seleccionar las dimensiones.....	55
Seleccionar los hechos.....	56
Matriz bus comidá.....	57
Base de datos.....	57
Restaurar bases de datos.....	57

## Introducción

Las organizaciones están haciendo uso de los Sistemas de Recurso Empresariales (*Enterprise Resource Planning* o ERP, por sus siglas en inglés) para automatizar las típicas áreas dirigidas por los gerentes operacionales, tales como: producción, logística, distribución, inventario, envíos, facturas, contabilidad, entre otras. Dichas áreas, generan una enorme cantidad de registros de operaciones, que son mostrados por los módulos de reportes, para apoyar la toma de decisiones rutinarias; que no excedan los procedimientos preestablecidos por la organización.

Sin embargo, tal flujo de registros de operaciones proveniente del ERP o de otros sistemas de la organización, resulta difícil de manipular para la gerencia de alto nivel, ya que esta necesita, que dicho flujo este unificado, estandarizado y resumido, para poder profundizar en todos los niveles de detalle de la información, de modo de tener una mejor comprensión del negocio y poder tener soporte para la toma de decisiones estratégicas poco rutinarias y difíciles de analizar.

En la actualidad, muchas empresas no le dan la adecuada importancia a la incorporación de nuevas tecnologías y herramientas para el desarrollo de sistemas que apoyen la toma de decisión estratégica debido al rechazo por parte de la comunidad empresarial de la organización o por la creencia de que los ERP pueden responder las preguntas de la gerencia de alto nivel. Aunque estos sistemas pueden generar consultas e informes, no están diseñados para recolectar, unificar y consultar la información de distintas fuentes y de forma eficiente, generando así las siguientes preguntas:

- ¿Se dedica la mayor parte del tiempo a elaborar informes?
- ¿Depende a menudo del departamento de TI para obtener información de tus aplicaciones?
- ¿No te fías de la información que presentas mediante informes?
- ¿Sientes que realizas tareas mecánicas y que no te queda tiempo para la toma de decisiones?

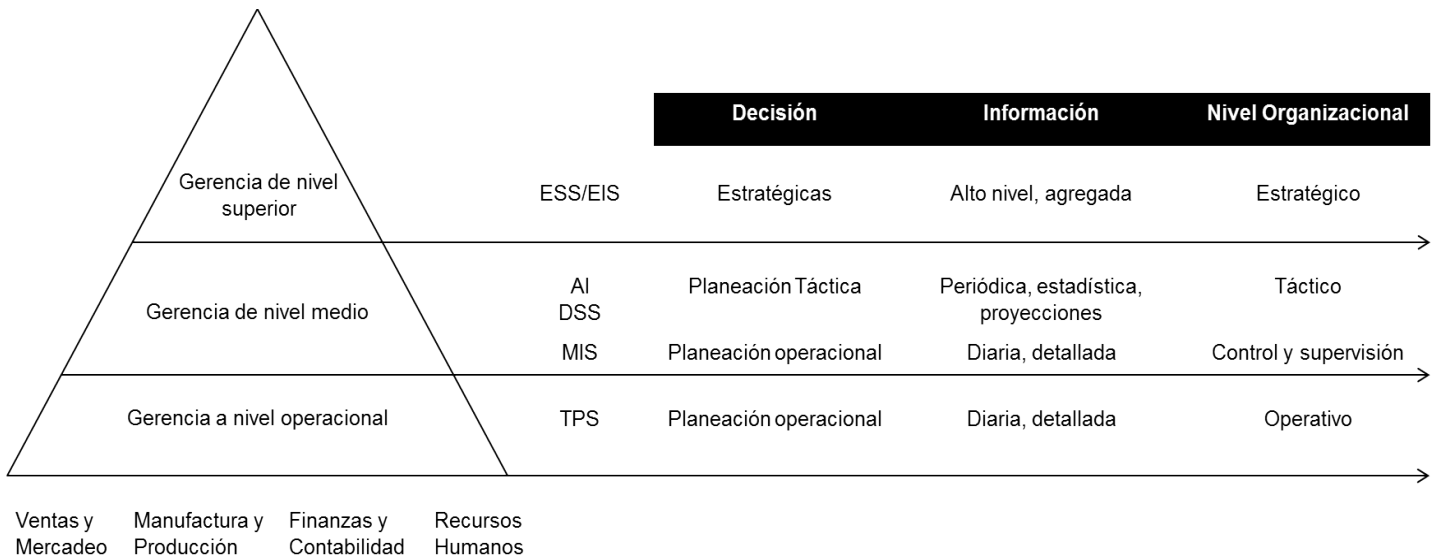
## 1. Los sistemas de información en las organizaciones

Los sistemas de información (SI) han sido utilizados desde 1990 por organizaciones de todo tipo ya que proporcionan un acceso rápido y centralizado a bases de datos de información personal, lectura de referencia, mejores prácticas, y son fácilmente adaptables para cubrir necesidades de la organización. Los SI en una organización poseen las siguientes características:

- **Automatiza los procesos:** Hacer lo mismo, más rápido, con menos gente y a menor costo.
- **Mejora la productividad:** Facilitar a los empleados a hacer más, en menos tiempo.
- **Produce ventajas competitivas:** Genera nuevas capacidades para la empresa.

Según la función a la que vayan destinados o el tipo de usuario final del mismo, los SI pueden clasificarse en:

- **Sistema de procesamiento de transacciones (TPS):** Gestiona la información referente a las transacciones producidas.
- **Sistemas de información gerencial (MIS):** Orientados a solucionar problemas empresariales en general.
- **Sistemas de soporte a decisiones (DSS):** Herramienta para realizar el análisis de las diferentes variables de negocio con la finalidad de apoyar el proceso de toma de decisiones.
- **Sistemas de información ejecutiva (EIS):** Herramienta orientada a usuarios de nivel gerencial, que permite monitorizar el estado de las variables de un área o unidad de la empresa a partir de información interna y externa a la misma. Es en este nivel cuando los sistemas de información manejan información estratégica para las empresas.



**Figura 1.1 - Niveles Organizacionales.**

## Herramientas de negocio

Han aparecido diferentes herramientas de negocio, como por ejemplo: EIS, OLAP, consultas e informes, minería de datos, etc. A continuación se muestran las diferencias entre los más usados:

**Sistemas de información ejecutiva (EIS):** Son un conjunto de herramientas asociadas que:

- Proporciona a los directivos acceso a información y sus actividades de gestión.
- Está especializado en analizar el estado diario de la organización (mediante indicadores clave) para informar rápidamente sobre cambios a los directivos.
- La información solicitada suele ser, en gran medida, numérica (ventas semanales, nivel del inventario, balances parciales, etc.) y representada de forma gráfica al estilo de las hojas de cálculo.

**Consultas e informes:** Los sistemas de informes o consultas avanzadas están basados, generalmente, en sistemas relacionales u objeto-relacionales. Utilizan los operadores clásicos: concatenación, proyección, selección, agrupamiento (en SQL y extensiones). Además, el resultado se presenta de una manera tabular.

**Herramientas OLAP:** Las herramientas OLAP funcionan sobre un sistema de información (transaccional o almacén de datos) y permiten realizar agregaciones y combinaciones de los datos de maneras mucho más complejas y ambiciosas, con objetivos de análisis más estratégicos. Estos sistemas ayudan a analizar los datos debido a que producen diferentes vistas de los mismos.

**Minería de datos:** La minería de datos es un conjunto de técnicas de análisis de datos que permiten extraer patrones, tendencias y regularidades para describir y comprender mejor los datos y predecir comportamientos futuros. Debido al gran volumen de datos, este análisis debe ser semi-automático.

La minería de datos se diferencia claramente del resto de herramientas, ya que **NO** transforma y facilita el acceso a la información para que el usuario la analice más fácilmente, sino que la minería de datos “analiza” los datos.

## Sistemas OLTP y sistemas OLAP

**Sistemas de Procesamiento de transacciones en línea (OLTP):** Son aplicaciones que ejecutan operaciones del día a día, como por ejemplo compras, inventario, nóminas. Estos sistemas definen el comportamiento habitual de un entorno operacional de gestión.

**Sistemas de Procesamiento analítico en línea (OLAP):** Son aplicaciones que se encargan de analizar el negocio, interpretar lo que ha ocurrido y tomar decisiones, como por ejemplo, mejorar los servicios al cliente, incrementar ventas, etc. Estos sistemas definen el comportamiento de un sistema de análisis de datos y elaboración de información.

Se pueden destacar las siguientes diferencias por medio de los siguientes criterios:

Criterio/Sistema	OLTP	OLAP
Orientación de los datos	Los datos son organizados inherentemente por aplicación.	Los datos son organizados por dimensiones definidas por el negocio.
	Focalizado en encontrar requerimientos de aplicaciones específicas para tareas específicas.	Focalizado en encontrar requerimientos de análisis empresarial.
Integración	Típicamente no integradas.	Debe ser integrada.
	Cada tema de negocios puede tener información en diferentes sistemas.	Toda información referente a un tema, es alimentado por varios sistemas reunidos en una sola bases de datos.
	Diferentes sistemas contienen diferentes tipos de datos.	Todos los tipos de datos integrados en un mismo sistema.
	Diferentes convenciones de nomenclatura.	Convenciones de nomenclatura estandarizadas.
	Diferentes formatos de archivos.	Formatos de archivos estándares.
	Diferentes plataformas de hardware.	Posee un solo servidor lógico.
Acceso y manipulación de datos por parte de usuarios finales	Los usuarios son los que operan el sistema: ingresar datos nuevos, eliminar, abrir-cerrar registros, corregir datos antiguos.	Los usuarios no modifican datos, sólo llevan a cabo consultas.
	Se ejecutan muchas veces las mismas acciones.	El usuario continuamente cambia el tipo de preguntas que realiza a la base de datos.
Administradores	Manipulación de datos registro a registro.	Carga y acceso de datos en forma masiva.
	Transacciones y/o rutinas de validación a nivel de registro.	Validación realizada antes o después de cada carga.
Transacción	Se manejan cientos de transacciones por día.	Se maneja sólo una transacción que contiene cientos de registros. Esto se hace a través de una carga de datos desde OLTP al repositorio del almacén de datos.
	Si la transacción fue realizada de manera exitosa, se asegura consistencia de ese pedazo de datos.	Si la carga termina exitosamente se tiene consistencia asegurada de todo el conjunto de datos.
Dimensión tiempo	Hay una falta de soporte explícito para reconstruir la historia previa.	La base de datos dimensional puede verse como una serie de capas de datos, compuestas cada una por una impresión del OLTP tomadas en intervalos regulares.
	Datos operacionales son altamente volátiles, cambian en la medida que opera la empresa y sus sistemas computacionales reflejan la operación.	Los datos del almacén de datos son altamente estables, son insertados en intervalos definidos y no son modificados.

**Tabla 1.1 – Comparación entre sistemas OLTP y OLAP.**

Teniendo en cuenta estos criterios, la siguiente tabla muestra una comparación general:

OLTP	OLAP
Almacena datos actuales.	Almacena datos históricos.
Almacena datos de detalle.	Almacena datos de detalle y datos agregados a distintos niveles.
Los datos son dinámicos (actualizables).	Los datos son estáticos.
Las transacciones son repetitivas.	Los procesos no son previsibles.
El número de transacciones es elevado.	El número de transacciones es bajo o medio.
Dedicado al procesamiento de transacciones.	Dedicado al análisis de datos.
Soporta decisiones diarias.	Soporta decisiones estratégicas.

**Tabla 1.2 – Comparación resumida entre los sistemas OLTP y OLAP.**

### Control de la información

La información es esencial para tomar buenas decisiones y se presenta en todos los niveles de la organización. Existen 3 tendencias en cuanto al manejo de la información por la cual la organización puede inclinarse:

- **Democracia de la información:** Cada persona tiene acceso a la información necesaria para el bien de la organización .
- **Comunismo de la información:** Todos acceden a la misma información.
- **Monarquía de la información:** Solamente la gerencia puede acceder a la información.

El comunismo y la monarquía de la información son tendencias simples, que solo tienen utilidad si la organización maneja pequeñas cantidades de información; sin embargo, ninguna es una alternativa real para la organización moderna. Todas tienen grandes (quizás enormes) cantidades de datos que pueden proporcionar una valiosa perspectiva a quienes los utilicen dentro de la organización, ayudándolos a ellos y a sus empresas a solucionar problemas con mayor rapidez y a explotar mejor las oportunidades.

### Retos de la democracia de información

Si todos los trabajadores pudieran acceder a la información que necesitan para desempeñar sus respectivas funciones, y si la información es uniforme y coherente en toda la organización, se lograría tener la verdadera democracia de la información, pero para lograrla, es necesario superar estos retos:

- **Variedad de conocimiento y destrezas de empleados:** No todos deberían ser expertos en SQL o si quiera conocerlo.
- **Transparencia de cambios en el esquema físico de datos:** Los cambios en el esquema físico deben ser propagados sin afectar la información.
- **Seguridad de datos:** Los usuarios deben ser restringidos a usar los datos apropiados a sus respectivas áreas.



- **Auditoría y cumplimiento:** Se deben cumplir todas las regulaciones en cuanto a auditoría y cumplimiento de reglas de acceso.

## Inteligencia de negocios

La Inteligencia de Negocios (*Business Intelligence* o BI, por sus siglas en inglés) se basa en aplicaciones, tecnología y procesos de recolección, almacenamiento y presentación de los datos para ayudar a los usuarios a tomar mejores decisiones. Esta incluye:

- La estandarización en la captura y consolidación de la información.
- Procesos de transformación de datos e interfaces con otros sistemas.
- Mayor control y confiabilidad en el manejo de Información.
- Repositorio único para la comparación de datos actuales versus datos históricos.
- Facilidad para la realización de Procesos Analíticos en Línea (OLAP).
- Visualización de indicadores que muestren la realidad de la gestión.
- Flexibilidad en la construcción de reportes ejecutivos.
- Presentación de la información en forma gráfica, análisis estadísticos y proyecciones.
- Detección de patrones de comportamiento de los datos operacionales (Minería de datos).

## Historia de la inteligencia de negocios

**Año 1958:** Una investigación realizada por IBM en 1958 define un sistema de BI como una colección de máquinas para el procesamiento de datos usada “para identificar información conocida, encontrar quien necesita saberla y diseminarla eficientemente.” El sistema llena una “necesidad creciente para mejores decisiones a niveles inferiores a los típicos en el pasado” .

**Años 70:** consultas y reportes .

Se vendían herramientas para “no-programadores” para acceder y analizar los datos. Estos sistemas eran cerrados (sin interoperabilidad), operaban contra extractos de datos típicamente no actualizados y eran limitados en cuanto a poder de procesamiento.

**Años 80:** Sistemas de Información.

Nacen los sistemas de soporte de decisiones (DSS) y de información ejecutiva (EIS). Los computadores aumentan su poder de procesamiento, surgen las hojas de cálculo y la computación cliente-servidor.

**Años 90:** Almacenes de Datos (DW) y Procesos analíticos en línea (OLAP).

En 1997, Wal-Mart creó un DW de 24 TB. Aparecieron los sistemas OLAP y modelos multidimensionales que facilitaron el análisis de grandes colecciones de datos. Por otra parte, hubo grandes avances en el poder del computador y redes, lo cual liberó mas datos a las masas .

**Pleno siglo 21:** BI en Masa y Democracia de Información.

Los proveedores de BI tienen productos basados en Web. Se busca agregar más funciones Ad Hoc (sistemas que permiten al usuario personalizar una consulta en tiempo real, en vez de estar atado a las consultas prediseñadas para informes), y de proveer BI a

todas las áreas de una empresa .

## Los cinco estilos de la inteligencia de negocios

**Notificaciones de Alertas:** Para lograr que los procesos de envío de alertas y avisos proactivos sean efectivos, es preciso contar con una aplicación de BI flexible y muy bien diseñada, que sea capaz de distribuir grandes cantidades de reportes y alertas a grandes comunidades de usuarios, tanto internos como externos.

**Análisis OLAP:** Mediante la funcionalidad OLAP es posible llevar adelante la forma más sencilla de análisis, permitiendo que cualquier persona pueda ver de manera minuciosa subconjuntos de datos interrelacionados o "cubos", simplemente con un clic. Para ello, el acceso y la manipulación de los datos se deben llevar a cabo a la "velocidad del pensamiento". Los usuarios pueden analizar los datos empleando características OLAP básicas, tales como:

- **Agregación Dinámica:** Agregación de datos en tiempo real.
- **Navegación:** Habilidad de movimiento entre los diferentes niveles de datos (Drill Down y Drill Up).
- **Segmentación:** Habilidad para combinar y re-combinar varias dimensiones con el fin de obtener distintas vistas de la información.
- **Pivote:** Habilidad de ofrecer comparaciones, revelar patrones y relaciones, y analizar tendencias.
- **Ordenamiento:** Ordenamiento de los datos, ya sea por dimensiones o medidas.

**Reportes:** Las funciones de reportes permiten que el BI llegue al público poniendo a su disposición información con alto nivel de detalle, lo que impacta fuertemente a los encargados de la toma de decisiones de las organizaciones. Independientemente del cargo o del tipo de trabajo que desarrollen, los miembros de la organización, como así también los socios de negocios y clientes, confían en el poder y la flexibilidad de sus sistemas de reportes que presentan los datos seleccionados en formatos cómodos y prácticos para realizar las operaciones diarias.

**Indicadores de gestión:** Los indicadores de control (o Tableros de Control) brindan información instantánea sobre el rendimiento del negocio. Habitualmente se construyen para gerentes y ejecutivos que necesitan tener una visión general del negocio. Para ellos es muy valioso poder ver muestras oportunas y visualmente intuitivas de la información estratégica, tanto financiera como operativa de la organización.

**Análisis predictivo:** La posibilidad de realizar análisis avanzados y predictivos, brinda tanto a los usuarios de negocios como a los analistas de información, capacidades completas y muy poderosas para investigaciones profundas de cualquier sector del datawarehouse para encontrar los detalles que se esconden tras los resultados.

## Comparación entre tipos de aplicaciones

Tipo de Aplicación	Enfoque	Manejo de datos	Interactividad	Ejemplos
Sistemas transaccionales (OLTP)	Ejecutar procesos de negocios	Cambia el estado actual de los datos detallados	Alto grado de interactividad	ERP CRM
Soluciones analíticas (OLAP)	Medir el rendimiento de la organización	Examina los datos de manera agrupada a través del tiempo	Alto grado de interactividad	Análisis de Ventas Análisis Financiero Tableros de control
Reportes Operacionales	Apoyar la toma de decisión y excepciones en procesos de negocio	Muestra datos detallados y actualizados	Estático	Reporte de excepciones Lista de tareas diarias Listado de selección y paquetes
Análisis predictivo (Minería de datos)	Identificar variables para predicciones utilizando algoritmos especializados	Utiliza datos detallados para el reconocimiento de patrones y ejecución de modelos embebidos	Embebido	Segmentación de clientes Siguiendo mejor oferta Detección de fraude

**Tabla 1.3 – Comparación entre tipos de aplicaciones.**

### Demostración con Pentaho

Una vez entendido los conceptos básicos de BI, comencemos con una demostración de un conjunto de programas libres para generar inteligencia de negocios llamado Pentaho BI Suite. **IMPORTANTE:** Pentaho está desarrollado en JAVA, por lo que es necesario tener instalado el JRE (versión 1.7 preferiblemente) y establecer la variable de entorno de JAVA\_HOME, para cualquier sistema operativo.

### Instalación rápida del servidor

Una vez su ambiente de ejecución esté preparado, siga los siguientes pasos:

1. Descargue la última versión estable (*Stable version*) del servidor de Pentaho: <http://sourceforge.net/projects/pentaho/files/Business%20Intelligence%20Server/>
2. Descomprima el archivo en la ruta de su preferencia.
3. Inicie el servidor de Pentaho utilizando el script de arranque *start-pentaho* que se encuentra en la carpeta descomprimida *biserver-ce*. Tome en cuenta que debe utilizar el script correspondiente a su Sistema Operativo: ejecutar el comando *sh start-pentaho.sh* para sistemas UNIX o presionar doble click al script *start-pentaho.bat* para sistemas Windows. Aparecerá en el terminal la siguiente información (presione Enter cuando aparezca el mensaje [OK]):

```

ricardo@ricardo-VIT-P2400:~/Pentaho/servers/biserver-ce$ sh start-pentaho.sh
DEBUG: Using JAVA_HOME
DEBUG: _PENTAHO_JAVA_HOME=/opt/java/jre1.7.0_55
DEBUG: _PENTAHO_JAVA=/opt/java/jre1.7.0_55/bin/java
-----
The Pentaho BI Platform now contains a version checker that will notify you
when newer versions of the software are available. The version checker is enabled by default.
For information on what the version checker does, why it is beneficial, and how it works see:
http://wiki.pentaho.com/display/ServerDoc2x/Version+Checker
Press Enter to continue, or type cancel or Ctrl-C to prevent the server from starting.
You will only be prompted once with this question.
-----
[OK]:

Using CATALINA_BASE:   /home/ricardo/Pentaho/servers/biserver-ce/tomcat
Using CATALINA_HOME:   /home/ricardo/Pentaho/servers/biserver-ce/tomcat
Using CATALINA_TMPDIR: /home/ricardo/Pentaho/servers/biserver-ce/tomcat/temp
Using JRE_HOME:        /opt/java/jre1.7.0_55
Using CLASSPATH:       /home/ricardo/Pentaho/servers/biserver-ce/tomcat/bin/bootstrap.jar
ricardo@ricardo-VIT-P2400:~/Pentaho/servers/biserver-ce$ █

```

4. Abra un navegador y visite la siguiente página: <http://localhost:8080>. De preferencia, utilice Firefox.

El servidor de Pentaho cuenta con una variedad de herramientas para la visualización de la información:

#### **Preinstaladas**

- JPivot, sólido, pero sin soporte desde el 2008.

#### **No preinstaladas**

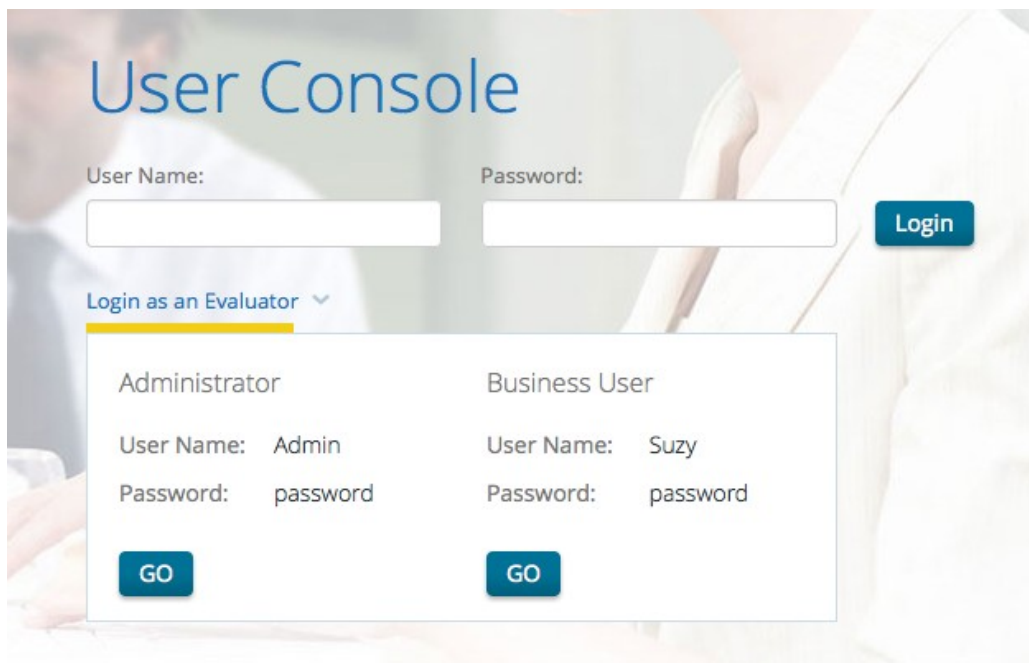
- Saiku, estable e innovador al ser unos de los primeros visores de código abierto que ofrece la navegabilidad de los datos usando la función de arrastre.
- Pivot4J, estable/incompleto, surge de la motivación de actualizar el Jpivot.

### **Ingreso al portal de Pentaho**

Ingrese a la plataforma utilizando un usuario que posea funciones de administrador:

Usuario: Admin

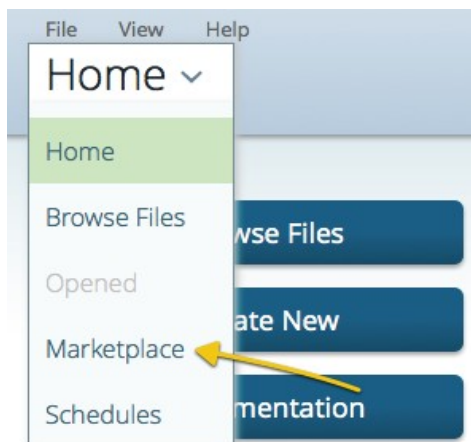
Clave: password



## Instalación de Saiku Analytics

Es posible extender las funcionalidades del servidor de Pentaho a través de complementos (plugin) los cuales pueden ser descargados desde el Marketplace, el gestor de complementos de Pentaho:

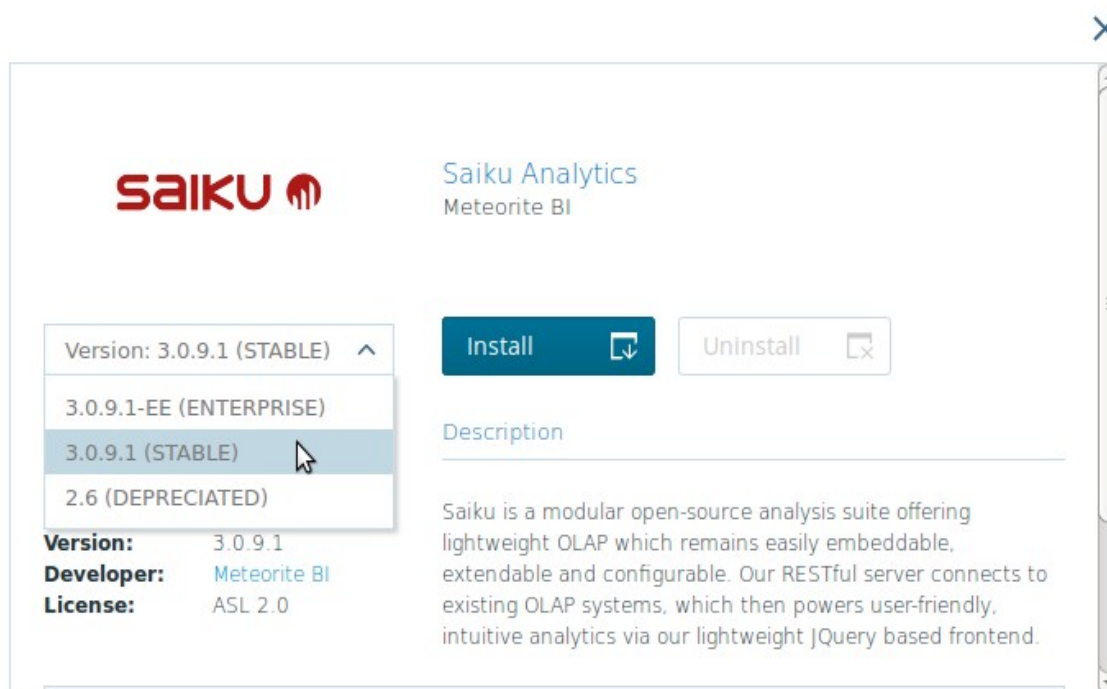
1. Ingrese al portal de Pentaho con el usuario administrador.
2. Cambie la perspectiva del portal presionando sobre el menú desplegable en la parte superior izquierda de la ventana justo debajo de la barra de herramientas y seleccione Marketplace.



### 3. Busque en la lista de complementos Saiku Analytics.

D3 Component Library Webdetails	2	Available 14.06.18 (RELEASE)	Install
Saiku Analytics Meteorite BI	4	Available 3.0.9.1-EE (ENTERPRISE)	Install
Saiku Chart Plus IT4biz	2	Available ChartPlusStable (vSaiku3)	Install

### 4. En la opción *Available*, seleccione la versión que desea instalar. Escoja una diferente a la *Edición Empresarial (EE)*.

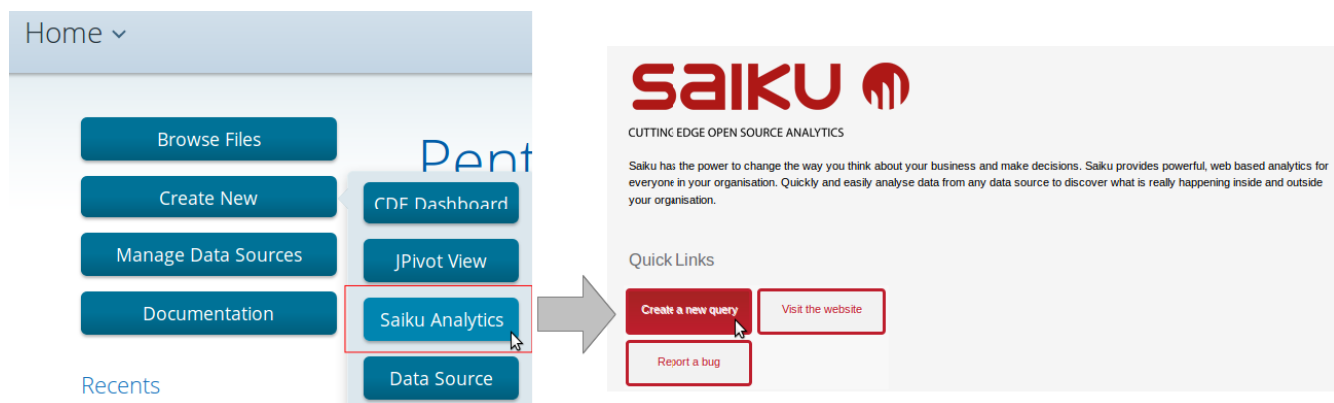


5. Presione el botón de *Install* y confirme la instalación presionando el botón *Ok*.
6. Una vez que termine debe reiniciar el servidor de Pentaho ejecutando el script *stop-pentaho*, localizado en la misma dirección que el script *start-pentaho*. Una vez ejecutado el script para detener el servidor, espere 20 segundos aproximadamente para iniciarlo nuevamente.

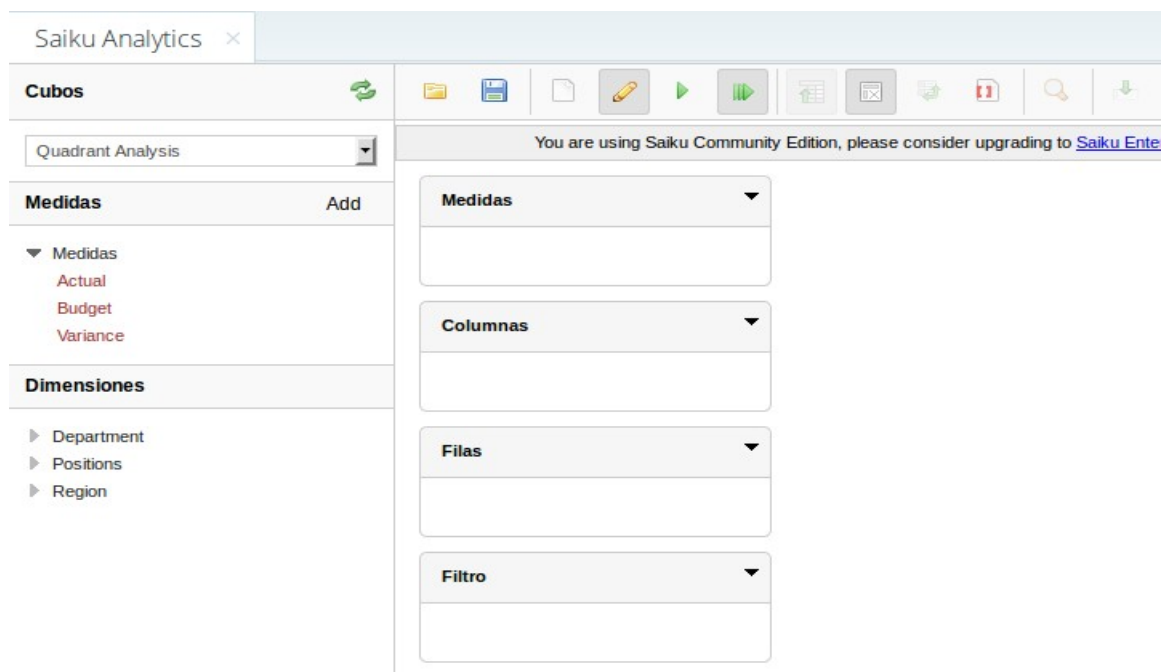
## Crear una vista de análisis con Saiku

Una vez reiniciado el servidor e ingresado al portal de Pentaho, es posible crear una vista de análisis con saiku:

1. Seleccione la opción "Create New" en el panel lateral izquierdo, y luego la opción "Saiku Analytics". Aparecerá una pantalla de presentación y para continuar, seleccione la opción "Create a new query".



2. En el panel lateral izquierdo de Saiku, presione el menú desplegable y escoja el cubo “Quadrant Analysis” perteneciente al esquema “SampleData”, un ejemplo de un modelo de ventas de Pentaho. Aparecerá el siguiente ambiente:



3. En el panel lateral izquierdo, en la sección “Dimensions”, despliegue la dimensión “Region” y arrastre el nivel “Region” a la sección de “Rows”, en el panel central.
4. En el panel lateral izquierdo, en la sección “Measures”, arrastre las tres medidas (“Actual”, “Budget” y “Variance”) a la sección de “Measures”, en el panel central.

Cubes

Quadrant Analysis

Measures Add

- Measures
  - Actual
  - Budget
  - Variance

Dimensions

- Department
- Positions
- Region
  - (All)
  - Region

Measures

- Actual
- Budget
- Variance

Columns

Rows

Region

Region	Actual	Budget	Variance
Central	37,893,162.00	38,397,600.00	504,438.00
Eastern	35,248,940.00	35,487,861.00	238,921.00
Southern	35,248,940.00	34,803,861.00	-445,079.00
Western	35,248,940.00	34,510,067.00	-738,873.00

El resultado de la vista se desplegará en el panel derecho, mostrando los costos actuales ("Actual"), el presupuesto ("Budget") y la varianza ("Variance") por región del negocio modelado.

## Funciones básicas

### Ordenamiento

Por defecto, los datos vienen ordenados alfabéticamente en relación a la dimensión. Sin embargo, es posible ordenarlos por las medidas. Para ello, en el panel central, en el menú desplegable de la sección "Rows", seleccione: **Sort** → **Ascending** → **Variance**.

Measures

- Actual
- Budget
- Variance

Columns

Rows

Region

Filter

Sort

Ascending

Descending

Ascending (Breaking Hierarchy)

Descending (Breaking Hierarchy)

Custom...

Clear Sort

Actual

Budget

Variance

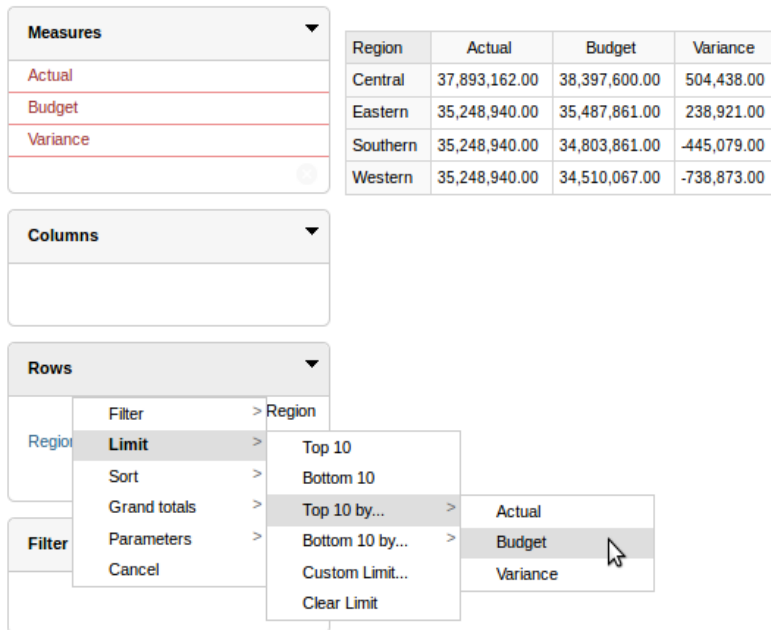
Region	Actual	Budget	Variance
Western	35,248,940.00	34,510,067.00	-738,873.00
Southern	35,248,940.00	34,803,861.00	-445,079.00
Eastern	35,248,940.00	35,487,861.00	238,921.00
Central	37,893,162.00	38,397,600.00	504,438.00



La opción seleccionada anteriormente ordena la medida “Variance” de forma ascendente. En las mismas opciones, se pueden ordenar las otras medidas, y de manera descendente. Para eliminar el ordenamiento, seleccionamos: **Sort** → **Clear Sort**.

## Límites

No siempre son necesario todo el conjunto de datos. Para limitar los registros bajo ciertas condiciones (por ejemplo, los diez primeros o los diez últimos) eliminamos el ordenamiento anterior y seleccionamos: **Rows** → **Limit** → **Top 10 by...** → **Budget**.

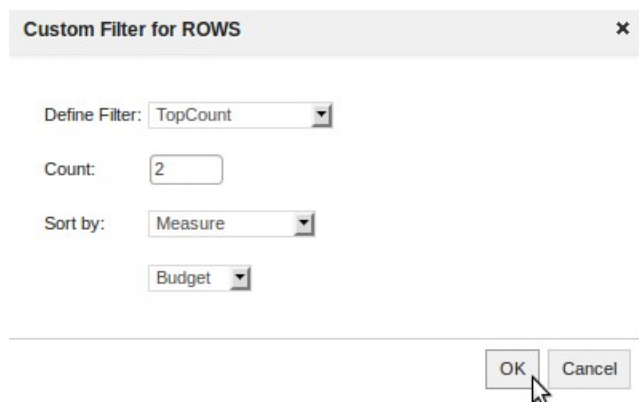


The screenshot shows the Tableau interface. On the left, the 'Measures' shelf contains 'Actual', 'Budget', and 'Variance'. The 'Columns' shelf is empty. The 'Rows' shelf contains 'Region'. A context menu is open over the 'Region' pill, showing options: Filter, Limit, Sort, Grand totals, Parameters, and Cancel. The 'Limit' option is selected, opening a sub-menu with: Top 10, Bottom 10, Top 10 by..., Bottom 10 by..., Custom Limit..., and Clear Limit. The 'Top 10 by...' option is selected, opening another sub-menu with: Actual, Budget, and Variance. The 'Budget' option is being selected by the mouse.

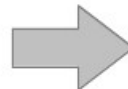
Region	Actual	Budget	Variance
Central	37,893,162.00	38,397,600.00	504,438.00
Eastern	35,248,940.00	35,487,861.00	238,921.00
Southern	35,248,940.00	34,803,861.00	-445,079.00
Western	35,248,940.00	34,510,067.00	-738,873.00

Aunque se haya aplicado el cambio, no hay diferencia debido a que sólo hay 4 registros. Para hacer un límite personalizado, seleccionamos: **Limit** → **Custom Limit...**. Aparecerá una ventana para definir el límite, en la cual agregaremos la siguiente información:

- **Define filter:** TopCount.
- **Count:** 2.
- **Sort by:** Measure → Budget



The screenshot shows the 'Custom Filter for ROWS' dialog box. It has a title bar with a close button (X). Inside, there are three fields: 'Define Filter:' with a dropdown menu showing 'TopCount', 'Count:' with a text input field containing '2', and 'Sort by:' with a dropdown menu showing 'Measure'. Below these fields is a 'Budget' pill. At the bottom of the dialog are 'OK' and 'Cancel' buttons. A mouse cursor is pointing at the 'OK' button.



Region	Actual	Budget	Variance
Central	37,893,162.00	38,397,600.00	504,438.00
Eastern	35,248,940.00	35,487,861.00	238,921.00

En la ventana anterior hemos creado una vista que muestra los dos registros de mayor presupuesto ("Budget") por región.

## Filtros

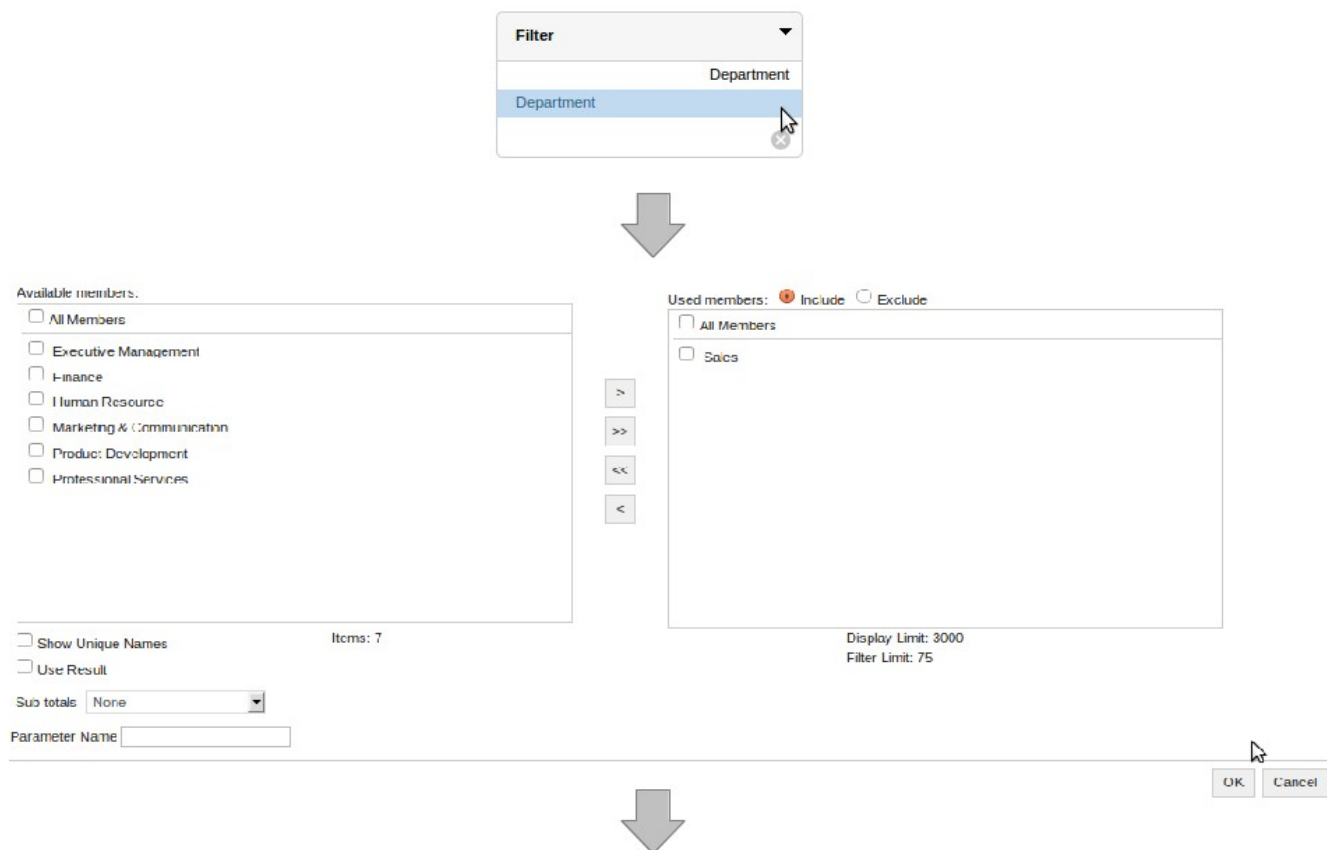
Elimine el límite personalizado para continuar con la siguiente función: **Rows** → **Limit** → **Clear limit**. Los filtros nos permiten seleccionar datos específicos al momento de crear una vista. Pueden ser seleccionados desde las secciones "Columns" y "Rows", o en la sección "Filter". Primero, creemos un filtro desde la sección "Rows":

1. Seleccione el nivel "Region".
2. Aparecerá una ventana que muestra dos listas. Seleccione las regiones "Eastern" y "Central" de la lista izquierda y desplácelo a la lista derecha. Presione el botón OK.



Eliminemos el filtro anterior desplazando las regiones "Eastern" y "Central" a la lista de la derecha. Ahora, agreguemos un filtro en la sección "Filter":

1. En el panel lateral izquierdo, en la sección "Dimensions", despliegue la dimensión "Departement" y arrastre el nivel "Department" a la sección de "Filter", en el panel central.
2. Seleccione el nivel "Department" en la sección de "Filter".
3. Seleccione el departamento "Sales" y desplácelo a la lista derecha. Presione OK.



Los filtros colocados en la sección “Filter” afectarán la vista de análisis a pesar de que no aparezcan explícitamente. En el ejemplo, los datos fueron limitados a los del departamento de ventas (“Sales”) por región.

## Totales y sub-totales

A continuación, elimine el filtro de la sección “Filter” presionando el botón de “x”, situado abajo del nivel “Department”.



Measures	Add	Measures		Region	Department	Actual
▼ Measures		Actual		Central	Executive Management	1,776,282.00
Actual					Finance	3,106,680.00
Budget					Human Resource	3,438,863.00
Variance					Marketing & Communication	3,590,423.00
					Product Development	2,997,702.00
					Professional Services	20,068,039.00
					Sales	2,915,173.00
				Eastern	Executive Management	1,507,580.00
					Finance	3,039,180.00
					Human Resource	3,212,200.00
					Marketing & Communication	3,440,110.00
					Product Development	2,548,800.00
					Professional Services	18,749,870.00
					Sales	2,751,200.00
				Southern	Executive Management	1,507,580.00
					Finance	3,039,180.00
					Human Resource	3,212,200.00
					Marketing & Communication	3,440,110.00
					Product Development	2,548,800.00

Para la siguiente demostración, agreguemos el nivel “Department” debajo del nivel “Region” en la sección de “Rows”, en el panel central, y elimine las medidas “Budget” y “Variance” simplemente presionando sobre ellas en el panel central.

En este tipo de vistas, es necesario conocer los totales y sub-totales de las dimensiones, en este caso, los sub-totales por región y el total de todas las regiones. Primero, creemos los sub-totales:

1. Presione el nivel “Region” en la sección de “Rows”.
2. En la parte inferior de la ventana, en el campo desplegable “Sub-totals”, seleccione “Sum” y presione OK.

Rows	
Region	Region
Region	Department
Department	



☐ Show Unique Names  
☒ Use Result  
 Sub totals Sum  
 Parameter Name



Region	Department	Actual
Central	Executive Management	1,776,282.00
	Finance	3,106,680.00
	Human Resource	3,438,863.00
	Marketing & Communication	3,590,423.00
	Product Development	2,997,702.00
	Professional Services	20,068,039.00
	Sales	2,915,173.00
		<b>37,893,162.00</b>
Eastern	Executive Management	1,507,580.00
	Finance	3,039,180.00
	Human Resource	3,212,200.00
	Marketing & Communication	3,440,110.00
	Product Development	2,548,800.00
	Professional Services	18,749,870.00
	Sales	2,751,200.00
		<b>35,248,940.00</b>
Southern	Executive Management	1,507,580.00
	Finance	3,039,180.00
	Human Resource	3,212,200.00

Ahora, los subtotales por región (o totales por departamento) serán calculados y resaltados al final de cada región.

Para calcular los totales por región, selección: **Rows** → **Grand totals** → **Sum**.

Region	Department	Sales
Central	Sales	2,751,200.00
		<b>35,248,940.00</b>
	Executive Management	1,507,580.00
	Finance	3,039,180.00
	Human Resource	3,212,200.00
	Marketing & Communication	3,440,110.00
	Product Development	2,548,800.00
Eastern	Professional Services	18,749,870.00
	Sales	2,751,200.00
		<b>35,248,940.00</b>
	Executive Management	1,507,580.00
	Finance	3,039,180.00
	Human Resource	3,212,200.00
	Marketing & Communication	3,440,110.00
Southern	Product Development	2,548,800.00
	Professional Services	18,749,870.00
	Sales	2,751,200.00
		<b>35,248,940.00</b>
	Executive Management	1,507,580.00
	Finance	3,039,180.00
	Human Resource	3,212,200.00
Western	Marketing & Communication	3,440,110.00
	Product Development	2,548,800.00
	Professional Services	18,749,870.00
	Sales	2,751,200.00
		<b>35,248,940.00</b>
	Executive Management	1,507,580.00
	Finance	3,039,180.00
Grand Total		<b>143,639,982.00</b>

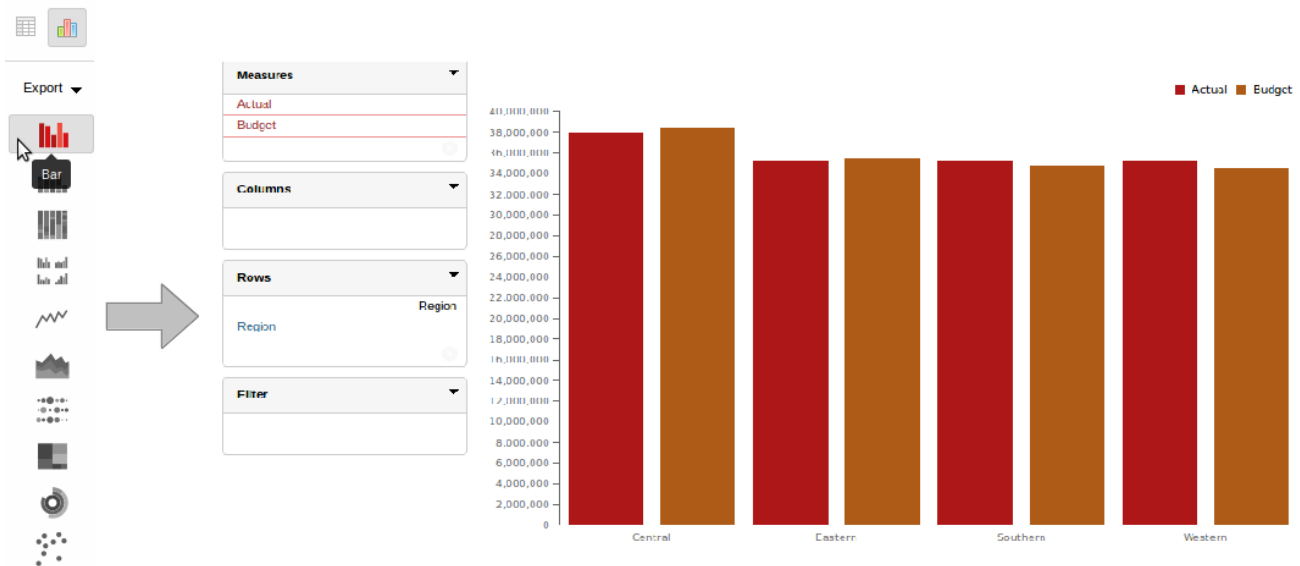
Ahora, el total de todas las regiones se agregará al final de la vista.

## Gráficos

1. Elimine el nivel "Department" de la sección de "Rows" y los totales y subtotales de la vista (Seleccione **None** en el proceso de selección de ambas funciones).
2. Agregue la medida "Budget".

Region	Actual	Budget
Central	37,893,162.00	38,397,600.00
Eastern	35,248,940.00	35,487,861.00
Southern	35,248,940.00	34,803,861.00
Western	35,248,940.00	34,510,067.00

Para crear un gráfico con la información mostrada, seleccione la opción "Chart mode" en el panel lateral derecho y luego el tipo de gráfico.



## Laboratorio

**Objetivo:** Familiarizarse con las opciones básicas de la herramienta Saiku.

Utilizando la herramienta de visualización saiku, cree una nueva vista de análisis utilizando como origen de los datos el modelo dimensional (cubo) de ventas de la empresa Steel Wheels (SteelWheelsSales), que responda las siguientes necesidades:

- Cantidad de artículos vendidos a lo largo de todos los años en operación.
- Cantidad de artículos vendidos por año, trimestre y mes.
- Cantidad e ingresos por venta de artículos vendidos a lo largo de todos los años en operación.
- Cantidad e ingresos por venta de artículos vendidos por el territorio del mercado a lo largo de todos los años en operación ordenados descendientemente por ingreso.
- En un gráfico de torta, muestre los ingresos por ventas correspondientes al territorio del mercado de los últimos 2 trimestres del año 2004 .

## 2. Almacenes de datos

Un almacén de datos (*Data Warehouse* o DW) es un sistema orientado a temas de negocio, diseñado especialmente para el soporte en la toma de decisiones del mismo. El ambiente del DW organiza y provee información de forma tal que el usuario final la entienda con facilidad. Los objetivos de un DW pueden ser deducidos fácilmente caminando por los pasillos de cualquier organización y escuchando a los gerentes del negocio. Estos temas tan recurrentes han existido por más de tres décadas:

- “Hemos reunido toneladas de datos, pero no podemos acceder a ella”
- “Tenemos que ver los datos desde todas las perspectivas posibles”
- “Las personas del negocio necesitan obtener los datos fácilmente”
- “Sólo muéstrame lo que es importante”
- “Pasamos reuniones enteras discutiendo sobre quién tiene los números correctos en lugar de tomar decisiones”
- “Queremos que las personas utilicen la información para apoyar la toma de decisiones basada en los hechos”

Estas citas pueden convertirse en los requerimientos que debe cumplir un sistema de BI/DW:

- **El sistema BI/DW debe hacer la información fácilmente accesible:** El contenido de este sistema debe ser comprensible. Los datos deben ser intuitivos para el usuario, no sólo para el desarrollador. La estructura de los datos y las etiquetas deberían imitar la forma de pensar y el vocabulario del usuario. Los usuarios del negocio quieren separar y combinar los datos analíticos de formas infinitas. Las herramientas del sistema BI deben ser simples y fáciles de usar, además de arrojar resultados de una consulta del usuario con tiempos de espera mínimos.
- **El sistema BI/DW debe presentar la información coherentemente:** Los datos en el sistema BI deben ser creíbles. Deben ser cuidadosamente ensamblados a partir de una variedad de fuentes, limpias, de calidad garantizada y sólo publicados cuando sea apto para el consumo del usuario. La coherencia también implica la definición de etiquetas dentro del sistema BI que se diferencien fácilmente entre sí.
- **El sistema BI/DW se debe adaptar al cambio:** Las necesidades del usuario, las condiciones del negocio, y las tecnologías están sujetas al cambio. El sistema BI debe estar diseñado para manejar inteligentemente estos cambios inevitables de modo que no invalide los datos o aplicaciones existentes. Estos datos y aplicaciones no deben cambiar o ser alterados cuando la comunidad empresarial genere nuevas preguntas o se agreguen nuevos datos al almacén. Por último, si los datos deben ser modificados, se debe manejar de forma apropiada estos cambios y hacerlos transparentes para el usuario.
- **El sistema BI/DW debe presentar la información de una manera oportuna:** A medida que el sistema es usado más y más para las decisiones operacionales, puede ser necesario convertir los datos sin procesar en información útil en cuestión de horas, minutos o incluso segundos. El equipo de BI y los usuarios del negocio necesitan tener una expectativa realista sobre lo que significa entregar los datos cuando hay poco tiempo para limpiarlos y validarlos.

- **El sistema BI/DW debe ser un lugar seguro para proteger la información:** La información más importante de la organización está guardada en el almacén de datos. Por ejemplo, el almacén probablemente contiene información de lo que estás vendiendo, a quién y a qué precio (detalles potencialmente dañinos para la empresa si caen en manos equivocadas). El sistema BI debe controlar efectivamente el acceso a la información confidencial de la organización.
- **El sistema BI/DW debe servir como base de autoridad y de confianza para la mejor toma de decisiones:** El DW debe tener los datos correctos para apoyar la toma de decisiones. Los resultados más importantes de un sistema BI son las decisiones que se toman en base a la evidencia analítica presentada; éstas decisiones proporcionan el impacto en el negocio y el valor atribuible al sistema. La mejor descripción del sistema BI que se va a diseñar es: un sistema para el apoyo de las decisiones.
- **La comunidad empresarial debe aceptar el sistema BI para determinar su éxito:** No importa si se construyó una solución elegante utilizando las mejores plataformas y productos de su clase, sí la comunidad empresarial no abraza el sistema BI ni lo usa activamente, habrá fracasado la prueba de aceptación. A diferencia de un sistema operacional donde el usuario no tiene más remedio que usarlo, el uso de un sistema BI a veces es opcional. Los usuarios del negocio adoptarán el sistema BI si es la fuente “más rápida y simple” para obtener la información procesada.

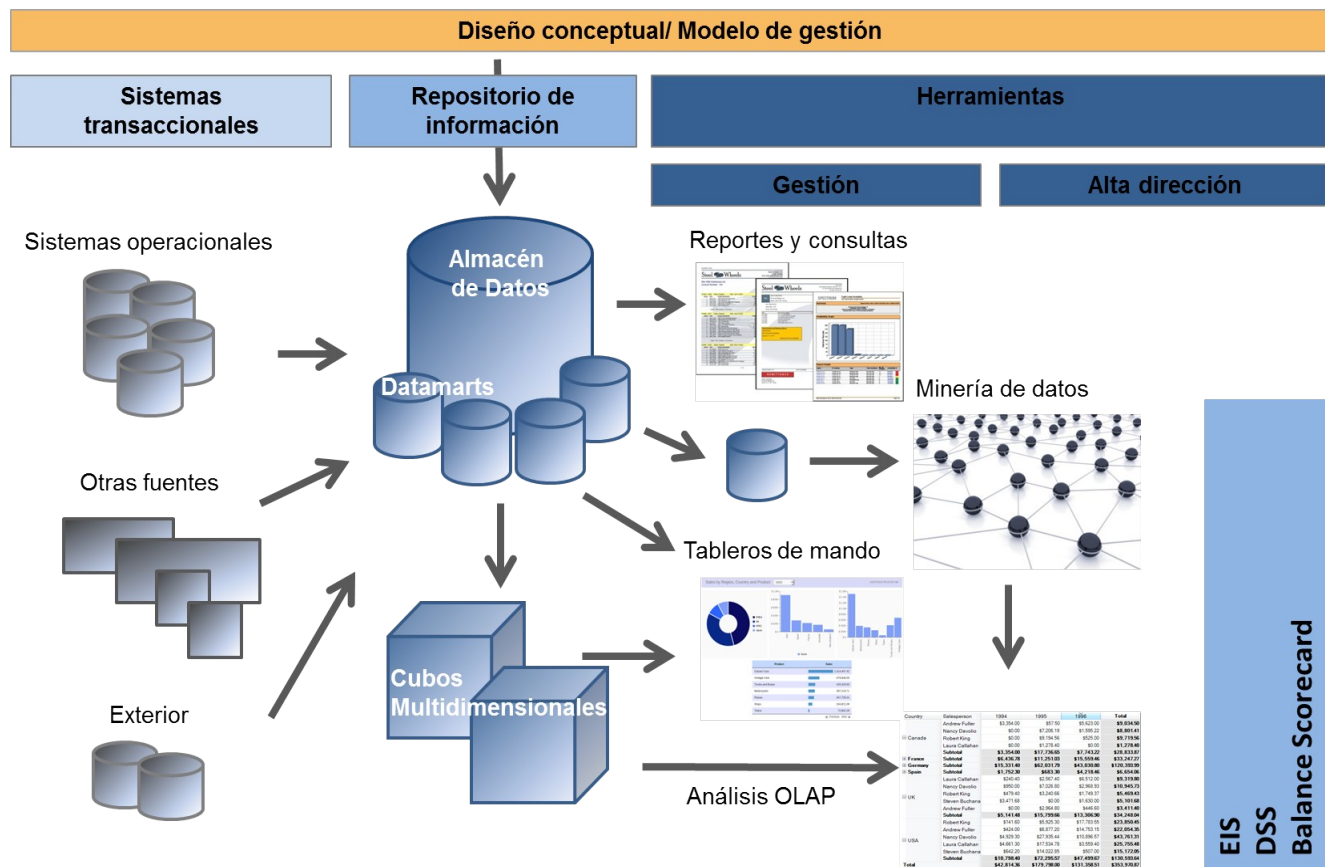
Aunque cada requerimiento en esta lista es importante, los últimos dos son los más críticos, y desafortunadamente, a menudo los más olvidados. El sistema BI exitoso exige más que una arquitectura estelar, técnicos, modeladores o administradores de base de datos.

### **Ventajas de un sistema DW**

- Proporciona una herramienta para la toma de decisiones en cualquier área funcional, basándose en información integrada y global del negocio.
- Facilita la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén; obteniendo un valor agregado para el negocio.
- Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.
- Supone una optimización tecnológica y económica en entornos de Centro de Información, estadística o de generación de informes con retornos de la inversión espectaculares.
- Relación total con el cliente.
- Facilidades en la gestión y análisis de recursos.
- Reaccionar rápidamente a cambios del mercado.



## Arquitectura/componentes de un sistema de BI



**Figura 2.1 – Componentes de un sistema de BI**

Cada componente del DW realiza una función específica y es importante conocer el significado estratégico de cada uno de ellos y cómo manejarlos de manera efectiva. En la **Figura 2.1**, muestra los siguientes componentes:

- **Sistemas transaccionales:** Son los sistemas que capturan los registros de las operaciones. Estos pueden estar en bases de datos centralizadas, hojas de cálculos, archivos de texto, entre otros.
- **Repositorios de información:** El área de repositorio es tanto el almacén de datos y el conjunto de procesos comúnmente llamados Extracción, Transformación y Carga (*Extract, Transformation and Load* o ETL, por sus siglas en inglés).
- **Herramientas:** Es el área donde los datos son organizados y puestos a disposición para la consulta directa por parte de los usuarios. Las herramientas utilizadas pueden ser simples consultas Ad Hoc o complejas aplicaciones de minería de datos.

## **Metáfora del restaurante**

### *ETL EN LA COCINA*

Los sistemas ETL son análogos a la cocina de un restaurante, donde los chefs obtienen los materiales y los transforman en deliciosos y apetitosos platillos para los clientes. Pero mucho antes de que la cocina empiece sus operaciones, se debe dedicar una gran cantidad de tiempo en la planificación del diseño del espacio de trabajo y de los componentes. La cocina tiene los siguientes objetivos:

- El diseño del lugar debe ser altamente eficiente. Cuando el restaurante está lleno de clientes con hambre, no hay tiempo para desperdiciarlo en movimientos.
- Entregar una calidad consistente es el segundo objetivo más importante. El chef debe crear su salsa en la cocina, en vez de simplemente enviar los ingredientes a las mesas.
- Por último, se debe entregar al cliente una comida de alta integridad. Los chefs no querrán que alguien se enferme por la comida del restaurante.

Una vez el diseño de la cocina esté listo, los chefs comienzan las operaciones, es decir, crean los diferentes platillos del menú. Ellos obtienen los materiales, los transforman en las diferentes comidas y se lo envían a los clientes, al igual que los sistemas ETL que extraen, transforman y cargan los datos.

### *PRESENTACIÓN DE LOS DATOS EN EL COMEDOR*

Centremos la atención al comedor. Según las estadísticas, los restaurantes se califican por los siguientes aspectos:

- Comida (Calidad, sabor y presentación).
- Decoración (Agradable, rodeada de patrones).
- Servicio (Entrega rápida, personal competente, comida entregada como se ordeno).
- Precios.

Estos puntos tienen que estar en equilibrio, ya que uno afecta al otro. Por ejemplo, si la comida es excelente, pero el ambiente es pobre, influye en la decisión del cliente sobre como calificar el restaurante. Los sistemas DW entregan los datos por medio de metadata, reportes publicados y operaciones analítica parametrizadas, al igual que los menús, que describen los platillos que están disponibles y la forma como se presentarán. La decoración del comedor debe tener patrones agradables. Debe estar diseñado basados en las preferencias de los clientes BI, no del personal programador. El servicio es un punto crítico en los sistemas DW. Los datos deben ser entregados, como se ordeno, de manera rápida. Por último, el costo es un factor en los sistemas de DW. Los chefs pueden elaborar excelentes y costosas comidas, pero si no hay mercado en ese precio, el restaurante no sobrevivirá.

## Diseño de un almacén de datos

El diseño de un DW consta de 5 etapas:

1. Recolección y análisis de requisitos.
2. Diseño conceptual.
3. Diseño lógico específico.
4. Diseño físico.
5. Implementación.

### Análisis de requisitos y diseño conceptual

Durante la etapa de análisis se deben distinguir las fuentes necesarias del sistema de información de la organización (OLTP) y las fuentes externas. También se deben recaudar los requisitos del usuario (consultas de análisis necesarias, nivel de agregación, indicadores claves de rendimiento y de metas, entre otros). Una vez obtenido los dos puntos anteriores, se elabora el diseño conceptual.

### Diseño lógico

Para elaborar el diseño lógico, se debe crear el modelo de dato dimensional, que es representado por lo general con un esquema estrella, donde se definan las relaciones, normalización e integridad adecuada.

### Diseño físico

El diseño físico surge a partir del diseño lógico. Esta fase se enfoca en la eficiencia del almacenamiento (Particionamiento de las tablas, tipo de indexación, entre otros) de la arquitectura OLAP que se utilizará:

- **OLAP Relacional (ROLAP):** La premisa de los sistemas ROLAP es que las capacidades OLAP se soportan en una base de datos relacional por medio de un modelo dimensional. La arquitectura ROLAP, accede a los datos almacenados en un almacén de datos para proporcionar los análisis OLAP. El sistema ROLAP utiliza una arquitectura de tres niveles:
  - La información es almacenada en una base de datos relacional.
  - El nivel de base de datos se encarga del manejo, acceso y obtención de los datos por medio de una base de datos relacional.
  - El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios.
  - El motor ROLAP se integra con los niveles de presentación, a través de los cuales los usuarios realizan los análisis OLAP.
- **OLAP Multidimensional (MOLAP):** Los sistemas MOLAP usan bases de datos multidimensionales, almacenando y manipulando los datos en estructuras especiales (sistema de matrices donde cada eje es una dimensión) para proporcionar el análisis. El sistema MOLAP utiliza una arquitectura de dos niveles, la base de datos multidimensional y el motor analítico:
  - La base de datos multidimensional es la encargada del manejo, acceso y

- obtención del dato.
- El nivel de aplicación es el responsable de la ejecución de los requerimientos OLAP.
- El nivel de presentación se integra con el de aplicación y proporciona una interfaz a través de la cual los usuarios finales visualizan los análisis OLAP.
- **OLAP Híbrido (HOLAP):** El sistema HOLAP incorpora las ventajas de los sistemas ROLAP y MOLAP: Los datos resumidos se almacenan dentro de una base de datos multidimensional (Sistemas MOLAP) y el detalle de los datos se almacena en una base de datos relacional (Sistemas ROLAP).

## Implementación

Una vez definidos diseño lógico y las herramientas OLAP que se utilizarán, se comienza la implementación del almacén de datos, comenzando por el sistema ETL, que es el encargado del mantenimiento del almacén de datos. La construcción del Sistema ETL es responsabilidad del equipo de desarrollo del almacén de datos y es construido específicamente para cada almacén de datos. Aproximadamente 50% del esfuerzo de desarrollo está centrado en la creación del sistema ETL. Las funciones principales del sistema ETL son:

- La carga inicial de los datos.
- Mantenimiento o refresco periódico: inmediato, diario, semanal, mensual, etc.

## 3. Modelo de dato dimensional

Una vez entendido los objetivos del sistema DW/BI, empezamos a considerar los principios del modelado dimensional. El modelado dimensional es ampliamente aceptado como la técnica preferida para la presentación de los datos analíticos, ya que aborda simultáneamente estos 2 requerimientos:

- Entregar datos que sean entendibles para el usuario.
- Generar consultas rápidas.

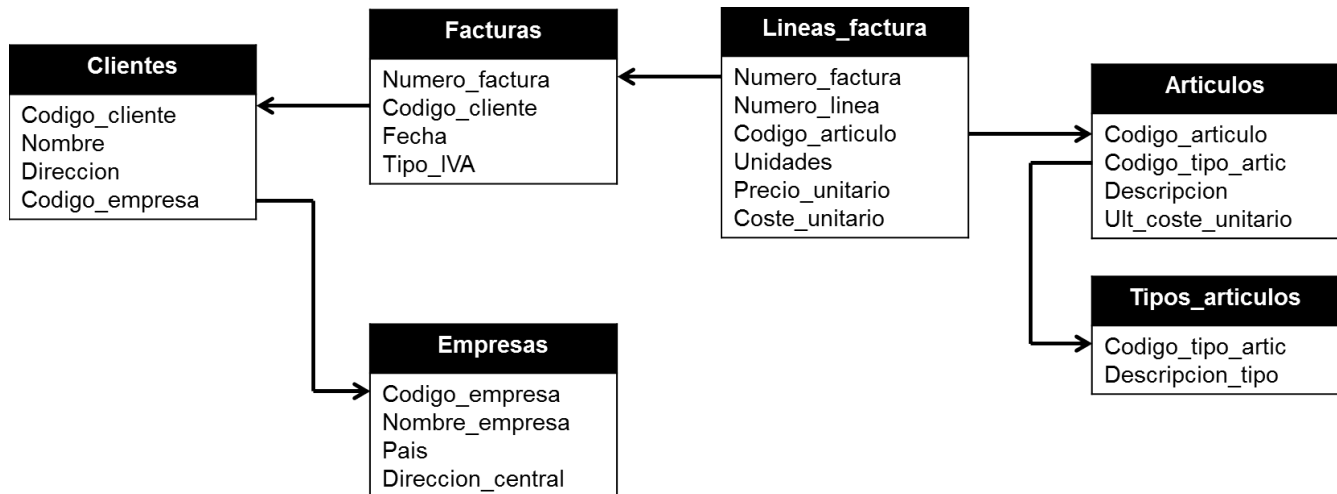
El modelado dimensional es una técnica desarrollada hace muchos años para el diseño de bases de datos simples. La sencillez es importante ya que garantiza que los usuarios puedan fácilmente entender los datos, así como también permite que el software navegue y entregue los resultados eficientemente.

### Modelo de dato relacional

Aunque los modelos dimensionales están desarrollados en Sistemas de Gestión de Base de Datos (SGBD), son muy diferentes a los modelos en la tercera forma normal (3FN) los cuales tienen como objetivo eliminar la redundancia. Las estructuras normalizadas hasta la 3FN, dividen los datos en diferentes entidades, las cuales se convierten en tablas relacionadas. Una base de datos de facturas de ventas, basada en la 3FN, puede ser un modelo complejo, compuesto de cientos de tablas normalizadas.

## Modelo entidad-relación

Los diagramas de entidad-relación (*Entity-Relationship Diagramming* o ERDs, por sus siglas en inglés) generan relaciones entre las tablas. Tanto como la 3FN y los modelos dimensionales pueden ser representados en ERDs debido a que ambos unen tablas relacionadas; la diferencia entre la 3FN y los modelos dimensionales es el grado de normalización. Ya que ambos modelos pueden ser representados como ERDs, nos referiremos a la 3FN como modelos normalizados. Las estructuras normalizadas son de mucha ayuda en el procesamiento operacional debido a que la actualización o inserción de una transacción toca la base de datos en sólo un lugar.



**Figura 3.1 – Modelo relacional básico de ventas.**

Sin embargo, los modelos normalizados son muy complicados para las consultas de BI. Los usuarios no podrían comprender, navegar o recordar los modelos normalizados que representen, por ejemplo, un mapa del sistema de autopistas de Los Ángeles. Del mismo modo, la mayoría de los SGBD no pueden consultar eficientemente un modelo normalizado; la complejidad de las consultas impredecibles de los usuarios sobrepasa los optimizadores de base de datos, lo que resulta en consultas de bajo rendimiento. Por ejemplo, formule la siguiente pregunta con respecto a la **Figura 3.1**: “¿Cuál es la ganancia de las ventas en el año 2014 por parte de las empresas en Panamá según tipo de artículos?”. Para poder responder a esa pregunta desde el modelo relacional en 3FN, es necesario hacer el siguiente SQL:

```

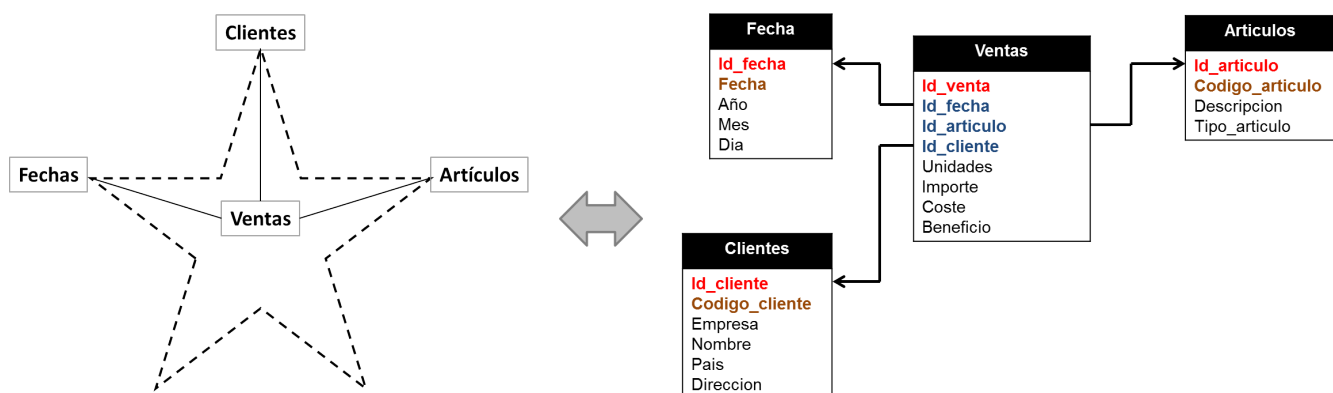
1 SELECT
2     Tipos_articulos.descripcion_tipo ,
3     SUM (unidades * (precio_unitario – coste_unitario))
4 FROM
5     Empresas, Clientes, Facturas, Lineas_factura , Articulos, Tipos_articulos
6 WHERE
7     Empresa.Codigo_empresa = Clientes.Codigo_empresa AND
8     Clientes.Codigo_cliente = Facturas.Codigo_cliente AND
9     Facturas.Numero_factura = Lineas_factura.Numero_factura AND
10    Lineas_factura.Codigo_articulo = Articulos.Codigo_articulo AND
11    Articulos.Cod_tipo_artic = Tipos_articulos.Cod_tipo_artic AND
12    Facturas.Fecha Between '2014/01/01' AND '2014/12/31' AND
13    Empresas.Pais = 'Panamá'
14 GROUP BY
15    Tipos_articulos.descripcion_tipo

```

La complejidad de la sentencia requiere un alto conocimiento en SQL por parte de los usuarios analíticos, además del costo en tiempo y recursos del SGBD en el procesamiento de cálculos y agrupaciones.

## Esquema estrella

Una forma de abordar estos problemas relacionados a esquemas complejos, es el desarrollo de modelos dimensionales implementados en un SGBD, usualmente llamados esquemas estrellas debido a su semejanza con la estructura de una estrella. Estos esquemas están conformados por dos tipos de tablas: las tablas de hechos, que contienen los valores de las medidas de negocio (o indicadores del negocio), y las tablas dimensiones, que contienen los atributos descriptivos utilizados para agrupar los datos de la tabla de hechos.



**Figura 3.2 – Representación física del esquema estrella.**

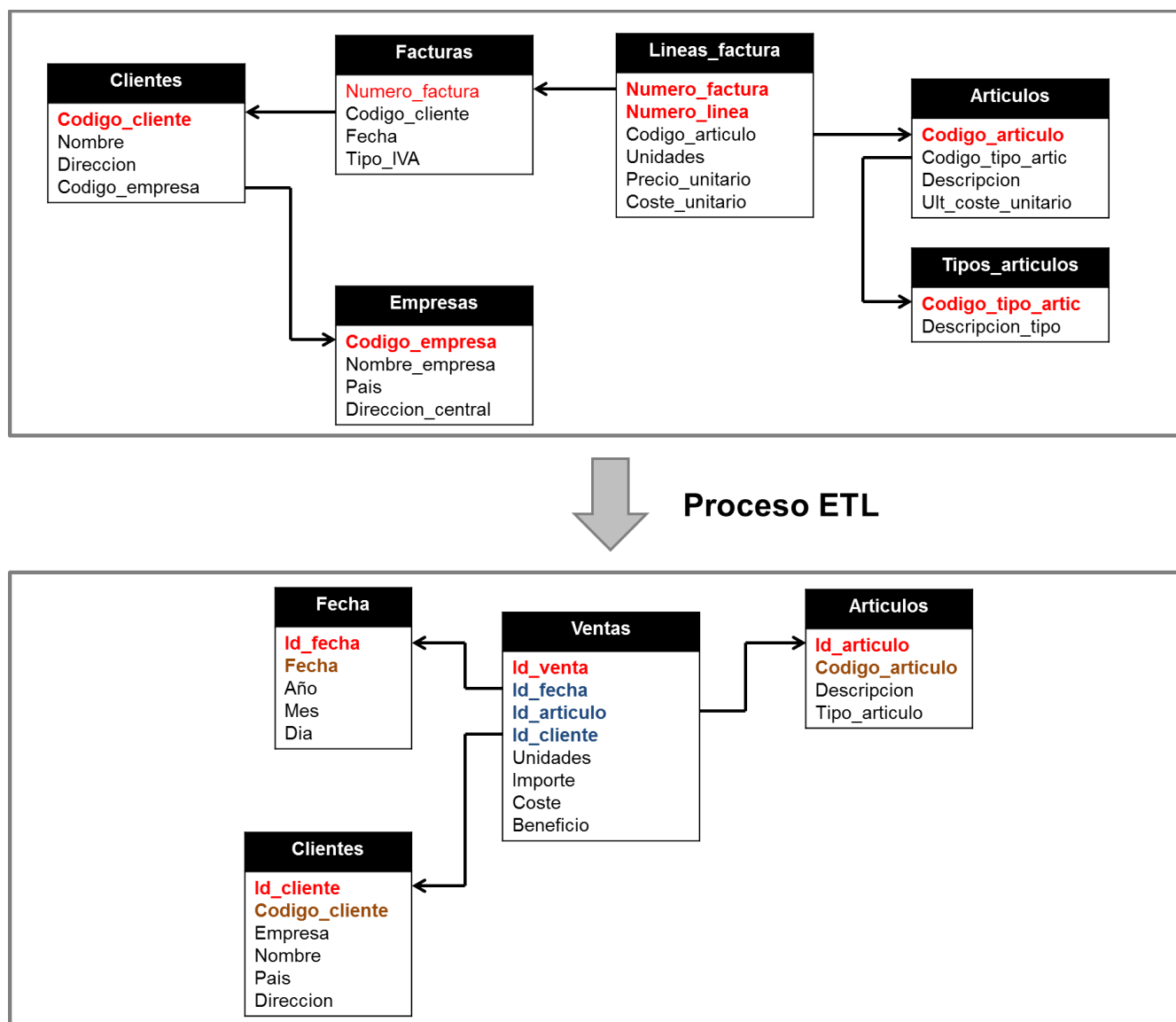


Figura 3.3 – Del esquema relacional al esquema estrella.

## Técnicas de modelado dimensional

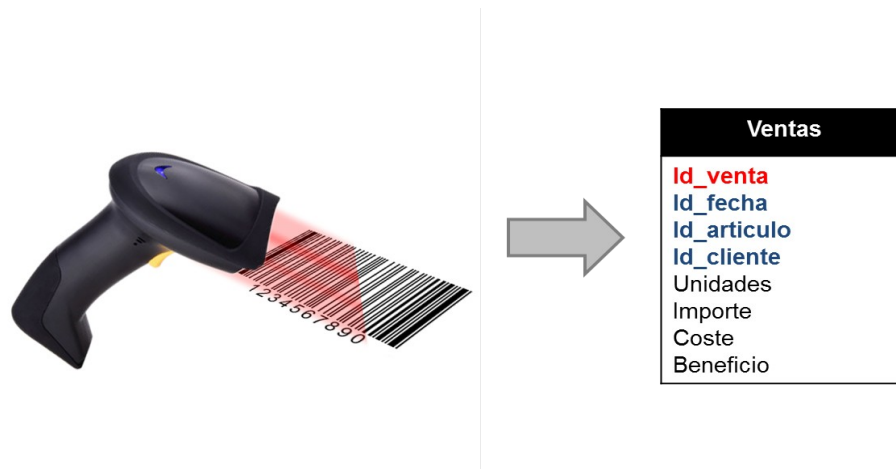
### Tablas de hechos

Las tablas de hechos en un modelo dimensional almacenan las medidas de rendimiento, como resultado de un evento del proceso de negocio de una organización. Estas medidas deben ser de bajo nivel y se deben almacenar en un solo modelo dimensional. Debido a que los datos de medición son el conjunto de datos más grande, no deben ser replicados en varios lugares por las diferentes áreas organizacionales de una empresa. Las tablas de hechos permiten que los usuarios de negocio puedan acceder a un único repositorio centralizado para cada conjunto de datos medibles, lo que asegura el uso de datos coherentes en toda la empresa.

El término *hecho* representa una medida de negocio. Imagine estar en el mercado, observando como los productos son vendidos en la caja registradora. Cuando los productos

son escaneados, se procesa la cantidad y el monto del producto en cada transacción de la venta.

Cada fila de la tabla de hechos corresponde a un evento de medición. Los datos en cada fila están a un nivel específico de detalle (Grano), que en este caso, es el producto vendido en una transacción. Uno de los principios básicos del modelado dimensional es que todas las filas de medición en una tabla de hechos deben estar en el mismo grano.



**Figura 3.4 - Eventos de medición del proceso de negocio.**

### **Aditividad en las medidas**

Los hechos mas útiles son numéricos y aditivos, tales como el monto de la venta. La aditividad es crucial debido a que las aplicaciones BI raramente consultan una sola línea de hechos; por el contrario, estos sistemas traen cientos, miles o incluso millones de hechos, y lo más útil que hacer con tantos datos es sumarlos. No importa como el usuario visualice los datos de la **Figura 3.4**, la cantidad y el monto sumarán un total válido. A veces, los hechos son semi-aditivos, o incluso no-aditivos. Los hechos semi-aditivos, tales como el saldo de cuenta, no pueden ser sumados a través de la dimensión tiempo. Los hechos no-aditivos, tales como el precio unitario, no deben ser sumados. Para este caso, se podrían realizar conteos, promedios, mínimo y máximo, o imprimir cada fila de la tabla de hechos (un ejercicio poco práctico cuando se tiene millones de filas).

### **Características físicas de una tabla de hechos**

Todas las tablas de hechos tienen dos o más claves foráneas (FK) que se conectan con la clave primaria de la tabla de dimensión. Por ejemplo, la clave del producto en la tabla de hechos siempre coincidirá con la clave del producto de la tabla de dimensión. Cuando todas las llaves en la tabla de hechos coinciden correctamente con sus respectivas claves primarias en las correspondientes tablas de dimensiones, las tablas satisfacen la integridad referencial.

La tabla de hechos por lo general tiene una clave primaria compuesta por todas las claves foráneas. Esta clave es a menudo llamada *clave compuesta*. Toda tabla que contenga una clave compuesta es una tabla de hechos. Las tablas de hechos expresan relaciones muchos-a-muchos. Las demás son tablas dimensiones.

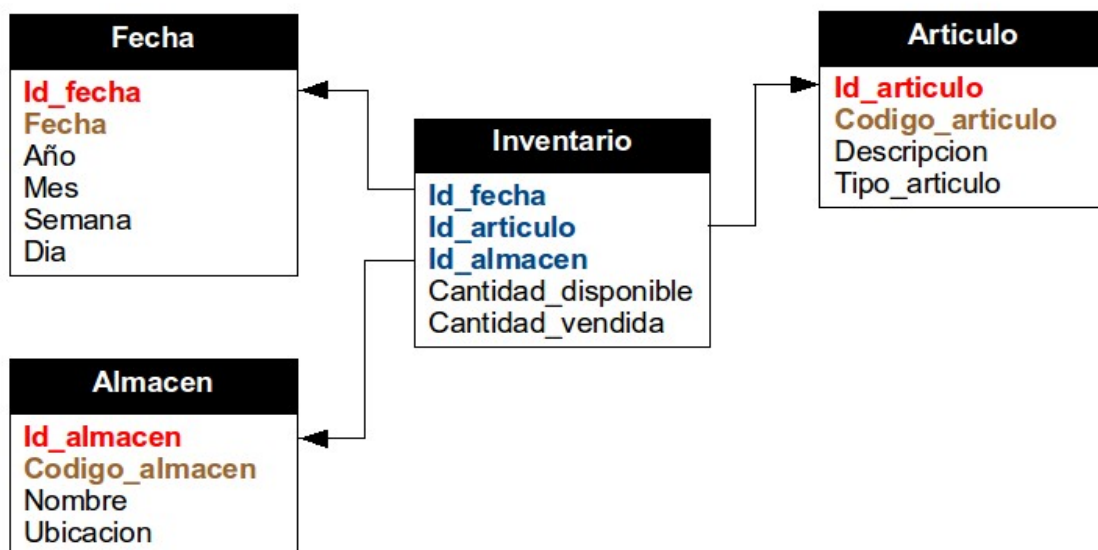


## Técnicas de las tablas de hechos

Existen muchas técnicas para definir las tablas de hechos; sin embargo, en este curso analizaremos las más comunes:

**Tablas de hechos transaccionales:** Representan eventos que suceden en un determinado tiempo y espacio. Se caracterizan por permitir analizar los datos con el máximo detalle. Reflejan las transacciones relacionadas con nuestros procesos de negocio (ventas, compras, contabilidad, entre otros). Es la tabla de hecho más común (**Figura 3.2 y 3.3**).

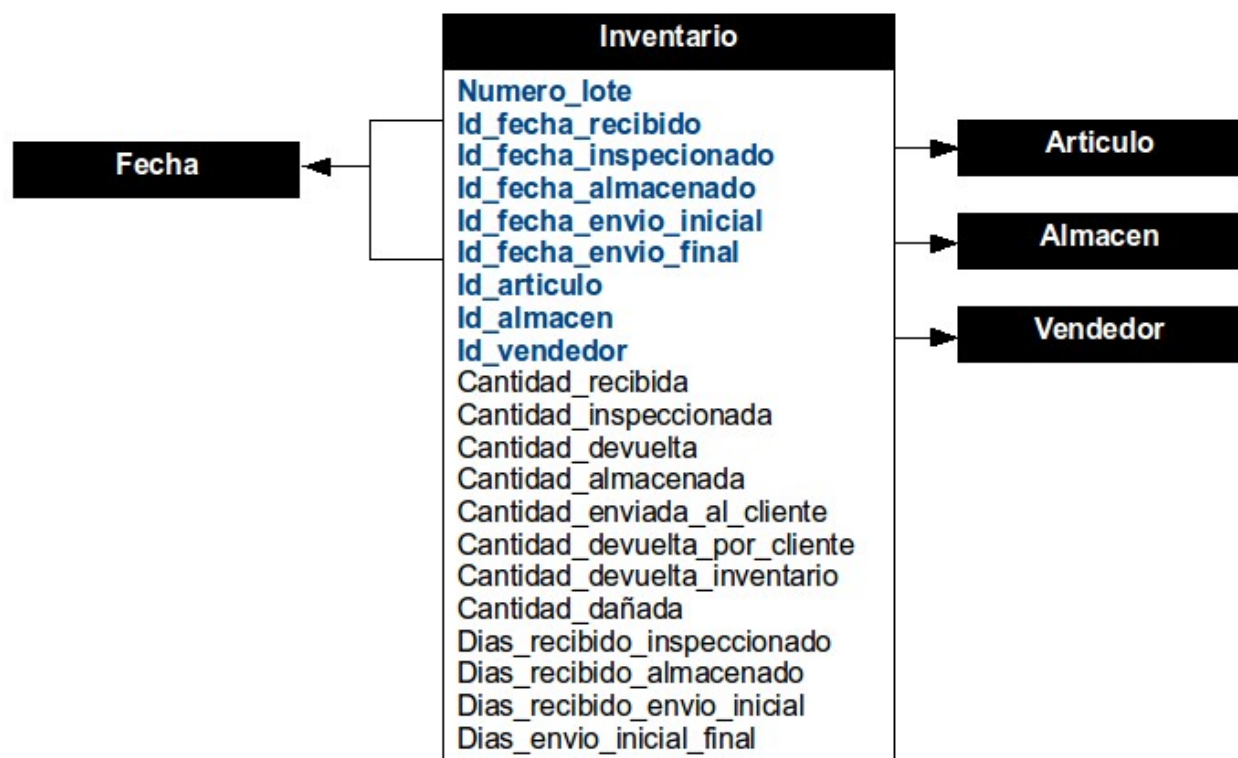
**Tablas de hechos periódica:** Una fila en una tabla de hechos periódica resume muchos eventos medibles ocurridos en un instante de tiempo, ya sea día, semana o mes. El grano es el período, no la transacción individual. Estas tablas son constantes con respecto a las claves foráneas ya que, incluso si no hay actividades durante el período, una fila se inserta normalmente en la tabla de hechos, colocando “ceros” o “nulos” en cada medida.



**Figura 3.5 – Proceso de inventario representado por una tabla de hechos periódica.**

En la **Figura 3.5**, el nivel de granularidad es el inventario diario por almacén por producto. Es posible que el grano sea por semana, por mes o incluso por año. Si el grano fuese por mes, el nivel de detalle de la dimensión fecha cambiaría, y sólo guardaría las fechas hasta el nivel de mes.

**Tablas de hechos acumulativas:** Representan el ciclo de vida completo de una actividad o proceso, que tiene un principio y final. Cada fila representa un proceso y se la actualiza repetidamente manteniendo actualizado su estado a medida que pasa por distintas etapas. La tabla de hecho resultante se caracteriza por tener múltiples fechas y hechos. Esta técnica resulta útil para realizar seguimiento a un nivel muy detallado.



**Figura 3.6 – Proceso de inventario representado por una tabla de hechos acumulativa.**

La **Figura 3.6** muestra el modelo de inventario final por medio de una tabla de hechos acumulativa. Esta tabla muestra el estatus actualizado de un lote en el inventario, representado por múltiples claves de fechas. Cada fila es actualizada repetidamente desde que un lote es recibido hasta que los productos del lote hayan sido completamente despachados.

Fila insertada cuando el lote es recibido

Numero de lote	ID Fecha recibido	ID Fecha inspeccionado	ID Fecha almacenado	ID Producto	Cantidad recibida	Dias transcurridos recibido-inspeccion	Dias transcurridos recibido-almacenado
101	20150101	0	0	1	100		

Fila actualizada cuando el lote fue inspeccionado

Numero de lote	ID Fecha recibido	ID Fecha inspeccionado	ID Fecha almacenado	ID Producto	Cantidad recibida	Dias transcurridos recibido-inspeccion	Dias transcurridos recibido-almacenado
101	20150101	20150103	0	1	100	2	

Fila actualizada cuando el lote fue almacenado

Numero de lote	ID Fecha recibido	ID Fecha inspeccionado	ID Fecha almacenado	ID Producto	Cantidad recibida	Dias transcurridos recibido-inspeccion	Dias transcurridos recibido-almacenado
101	20150101	20150103	20150104	1	100	2	3

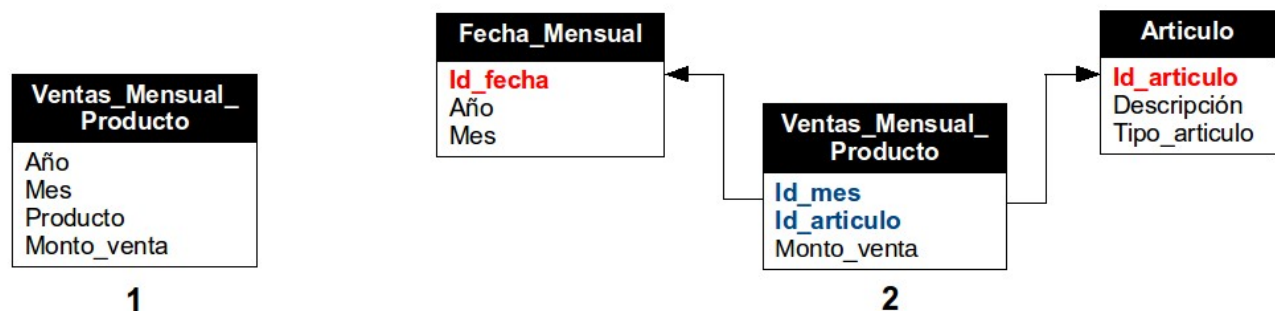
**Figura 3.7 – Proceso de almacenado en la tabla de hechos acumulativa.**

Características	Transacción	Periódico	Acumulativo
Grano	Una fila por transacción	Una fila por período de tiempo	Una fila para el tiempo de vida total de un suceso
Dimensión	Dimensión de fecha en el nivel más bajo de granularidad	Dimensión de fecha en la granularidad de fin de período	Varias dimensiones de fecha
Número de dimensiones	Más que el tipo de hechos periódicos	Menos que el tipo de hechos de transacciones	Número más alto de dimensiones cuando se compara con otros tipos de tablas de hechos
Dimensiones conformadas	Utiliza dimensiones conformadas compartidas	Utiliza dimensiones conformadas compartidas	Utiliza dimensiones conformadas compartidas
Medidas	Se relaciona con actividades de transacciones	Se relaciona con actividades periódicas	Se relaciona con actividades que tienen un tiempo de vida definitivo
Tamaño de base de datos	Es el mayor tamaño. En el nivel de grano más detallado, tiende a crecer muy rápido.	Más pequeña que la tabla de hechos de transacciones, ya que el grano de la dimensión de fecha y hora es significativamente mayor.	La de tamaño más pequeño cuando se compara con las tablas de hechos periódicos y de transacciones.
Rendimiento	Funciona bien y se puede mejorar eligiendo un grano por encima del más detallado	Funciona mejor que otros tipos de tablas de hechos, ya que los datos se almacenan en un grano menos detallado	Funciona bien
Insertar	Sí	Sí	Sí
Actualizar	No	No	Sí, cuando se alcanza un objetivo en una actividad determinada.
Suprimir	No	No	No
Crecimiento de tabla de hechos	Muy rápido	Lento en comparación con una tabla de hechos basada en transacciones	Lento en comparación con la tabla de hechos periódicos y de transacciones
Necesidad de agregación de tablas	Alta, principalmente porque los datos se almacenan a un nivel muy detallado	Baja o muy baja, principalmente porque los datos ya están almacenados en un nivel alto de agregación	Media, porque los datos se almacenan principalmente en el nivel diario. Sin embargo, los datos de las tablas de hechos acumulativos se encuentran en un nivel inferior al nivel de transacción.

**Tabla 3.1 – Comparación entre tablas de hechos.**

**Tablas de hechos sin hechos:** Aunque la mayoría de los eventos capturan resultados numéricos, es posible que el evento solamente colectione un conjunto de entidades en un momento determinado. Por ejemplo, un estudiante asistiendo a clases en un día no tiene ningún hecho numérico; sin embargo, es posible analizar la asistencia por medio de las dimensiones de calendario, estudiante, profesor, lugar y clase.

**Tablas agregadas:** El rendimiento de un almacén de datos depende de tres factores importantes: diseño, hardware y cantidad de datos. Usualmente, las tablas de hechos guardan el nivel de detalle más alto para poder responder a todas las preguntas del negocio; sin embargo, es posible que el volumen de dato almacenado en las tablas de hechos sobrepase el diseño y el hardware, afectando el rendimiento de las consultas. Una técnica que permite abordar este tipo de problema son las tablas agregadas: Tablas pre-calculadas con datos resumidos de las tablas de hechos. La mayoría de las consultas no necesitan acceder al máximo nivel de detalle, por lo que las tablas agregadas podrían guardar datos calculados de, por ejemplo, las ventas mensuales, por cliente, por producto o por combinaciones.



**Figura 3.8 – Tablas agregadas de las ventas mensuales por productos.**

La **Figura 3.8** muestra dos formas diferentes de definir una tabla agregada. La primera muestra los atributos de las dimensiones en la misma tabla, mientras que la segunda muestra la relación con la dimensión producto y la dimensión encogida de fecha mensual (Dimensión explicada más adelante) .

**Nulos en la tabla de hechos:** A pesar que en las medidas, los nulos no afectan las funciones de agrupaciones, se deben evitar en las claves foráneas debido a que esos nulos causarían una violación a la integridad referencial. En lugar de una clave foránea nula, la tabla de dimensión asociada debe tener una fila por defecto (y clave sustituta) que represente la condición “desconocido” o “no aplica”.

## Tabla de dimensión

Las tablas de dimensiones son las que integran el modelo dimensional junto con la tabla de hechos. Estas tablas contienen el contexto asociado a los eventos medibles del proceso de negocio. Describen el “quién, qué, dónde, cuándo, cómo y por qué” asociado con el evento.

Como ilustra la **Figura 3.9**, las tablas de dimensiones a menudo tienen muchas columnas, o atributos. Es común para una tabla de dimensión, tener de 50 a 100 atributos; aunque, algunas tablas de dimensiones pueden tener pocos atributos. Las tablas de dimensiones tienden a tener menos filas que la tabla de hechos, pero pueden ser más amplias, con columnas de texto largo. Cada dimensión es definida por una sola clave primaria (llamada PK en la **Figura 3.9**), la cual sirve como base para la referencia integral con cualquier tabla de hechos a la que este ligada.

Articulos	
<b>Id_articulo (PK)</b>	
<b>Codigo_articulo (Clave Natural)</b>	
Marca	
Categoria	
Departamento	
Tipo_paquete	
Tamaño_paquete	
Peso	
Tipo_almacenado	
Fecha_expiracion	
:	
:	

**Figura 3.9 - Tabla de dimensión que contiene las características descriptivas del proceso de negocio.**

Los atributos de las dimensiones sirven como fuente principal para las restricciones en las consultas, agrupaciones y etiquetas de reportes. En una consulta o solicitud de informe, los atributos son identificados por las palabras. Por ejemplo, cuando un usuario desea ver el monto de las ventas por marca, las marcas deben estar disponibles como un atributo de la dimensión. Los atributos de las tablas de dimensiones juegan un papel vital en el sistema BI. Debido a que son la fuente de prácticamente todas las restricciones y etiquetas de reportes, los atributos de las dimensiones son fundamentales para lograr que el sistema BI sea utilizable y comprensible.

## Calidad de la dimensión

Los atributos de las dimensiones sirven como fuente principal para las restricciones en las consultas, agrupaciones y etiquetas de reportes. En una consulta o solicitud de informe, los atributos son identificados por las palabras. Por ejemplo, cuando un usuario desea ver el monto de las ventas por marca, las marcas deben estar disponibles como un atributo de la dimensión. Los atributos de las tablas de dimensiones juegan un papel vital en el sistema BI,

debido a que son la fuente de prácticamente todas las restricciones y etiquetas de reportes. Los atributos de las dimensiones son fundamentales para lograr que el sistema BI sea utilizable y comprensible.

Los atributos deben estar compuestos de palabras concretas en vez de abreviaturas. Se debe tratar de minimizar el uso de códigos en las tablas de dimensiones y reemplazarlos con atributos de texto más detallado. Para ello, se debe hacer una decodificación estándar de los códigos operacionales y disponerlos como atributos de la dimensión para proporcionar un etiquetado consistente en las consultas, reportes y aplicaciones BI.

A veces, los códigos operacionales o identificadores tienen un significado legítimo en el negocio para el usuario o son necesarios para la comunicación en el mundo operacional. En estos casos, los códigos deben aparecer como atributos explícitos de la dimensión, junto con una descripción textual fácil de entender. Los códigos operacionales a veces tienen “inteligencia” integrada en ellos. Por ejemplo, los dos primeros dígitos pueden significar el tipo de negocio, mientras que los siguientes dos dígitos pueden identificar la región global. En vez de forzar al usuario a preguntar o separar manualmente el código, se debe extraer el significado implícito y presentarlo como atributos separados de la dimensión, para que puedan ser fácilmente filtrados, agrupados o reportados. En muchos sentidos, la eficacia del almacén de datos dependerá de los atributos de las dimensiones; el poder analítico del ambiente BI es directamente proporcional a la calidad y profundidad de los atributos de la dimensión. Cuanto más tiempo se invierta en agregarle a los atributos la terminología detallada del negocio y en asegurar la calidad de los valores, mejor. Atributos robustos ofrecen capacidades de análisis robustas.

## **Jerarquías en las dimensiones**

La **Figura 3.10** muestra que las tablas de dimensiones a menudo representan relaciones jerárquicas. Por ejemplo, los productos pertenecen a una marca y luego a una categoría. Para cada fila en la dimensión de producto, se debe almacenar la descripción de la marca y la categoría. La información jerárquica se almacena de forma redundante para mejorar el rendimiento en las consultas y la facilidad de uso, por lo que se debe evitar la normalización. Debido a que, por lo general, las tablas de dimensiones tienen muchas menos filas que las tablas de hechos, no tendrá impacto alguno mejorar la eficiencia del almacenamiento mediante la normalización.

Clave del Producto	Descripción	Marca	Categoría
1	PowerAll 20 oz	PowerClean	Limpiador multiuso
2	PowerAll 32 oz	PowerClean	Limpiador multiuso
3	PowerAll 48 oz	PowerClean	Limpiador multiuso
4	PowerAll 64 oz	PowerClean	Limpiador multiuso
5	ZipAll 20 oz	Zippy	Limpiador multiuso
6	ZipAll 32 oz	Zippy	Limpiador multiuso
7	ZipAll 48 oz	Zippy	Limpiador multiuso
8	Shiny 20 oz	Clean Fast	Limpiador de vidrios
9	Shiny 32 oz	Clean Fast	Limpiador de vidrios
10	ZipGlass 20 oz	Zippy	Limpiador de vidrios
11	ZipGlass 32 oz	Zippy	Limpiador de vidrios

**Figura 3.10 – Jerarquías en una dimensión.**

## Técnicas de dimensiones

Existen muchas técnicas para definir las dimensiones; sin embargo, en este curso analizaremos las más comunes:

**Dimensiones Degeneradas:** Son campos almacenados en la tabla de hechos. Esto sucede cuando un campo que se utilizará como criterio de análisis posee el mismo nivel de granularidad que los datos de la tabla de hechos, y que por lo tanto no se pueden realizar agrupaciones o cálculos a través de este campo. Los “*números de orden*”, “*números de ticket*”, “*números de transacción*”, etc., son algunos ejemplos de dimensiones degeneradas. La inclusión de estos campos en las tablas de hechos, se lleva a cabo para reducir la duplicación y simplificar las consultas. Se podría plantear la opción de simplemente incluir estos campos en una tabla de dimensión, pero en este caso estaríamos manteniendo tantas filas en la dimensión como en la tabla de hechos, por consiguiente obtendríamos la duplicación de información y complejidad, que precisamente se pretende evitar.

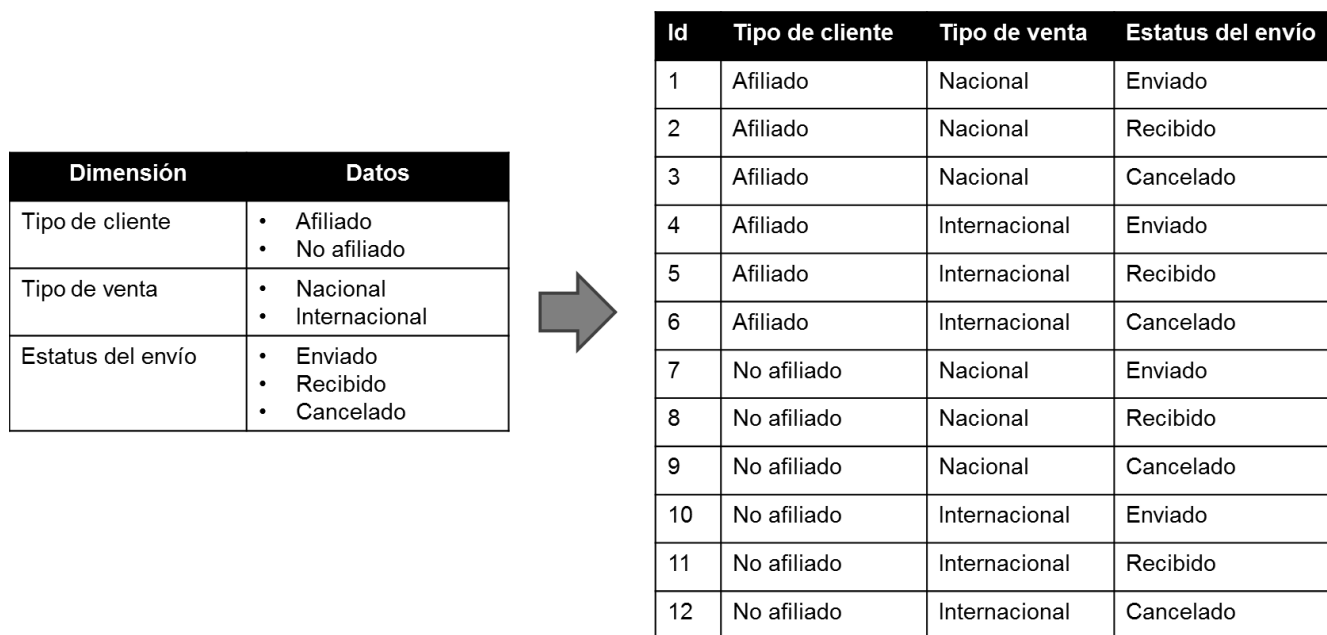
**Dimensiones JUNK:** Los procesos de negocios a menudo producen un número de banderas e indicadores de baja cardinalidad, por ejemplo, un estatus de compra. En vez de hacer una dimensión separada para cada uno de esas banderas y atributos, se puede crear una dimensión que combine todos los valores. Esta dimensión no necesariamente debe tener un producto cartesiano de todos los posibles valores de los atributos, sino que debería tener todas las combinaciones posibles que sucedan en el sistema fuente.

Imaginemos una empresa que guarde la siguiente información de sus ventas: cliente

afiliado o no afiliado, venta nacional o internacional, estatus del envío (enviado, recibido o cancelado). Para este caso, es posible:

- Dejar los atributos en la tabla de hechos (Problema: mucho espacio consumido).
- Hacer dimensiones separadas por cada uno de ellos (Problema: Mientras más dimensiones, más enlaces tendrá la tabla de hechos).

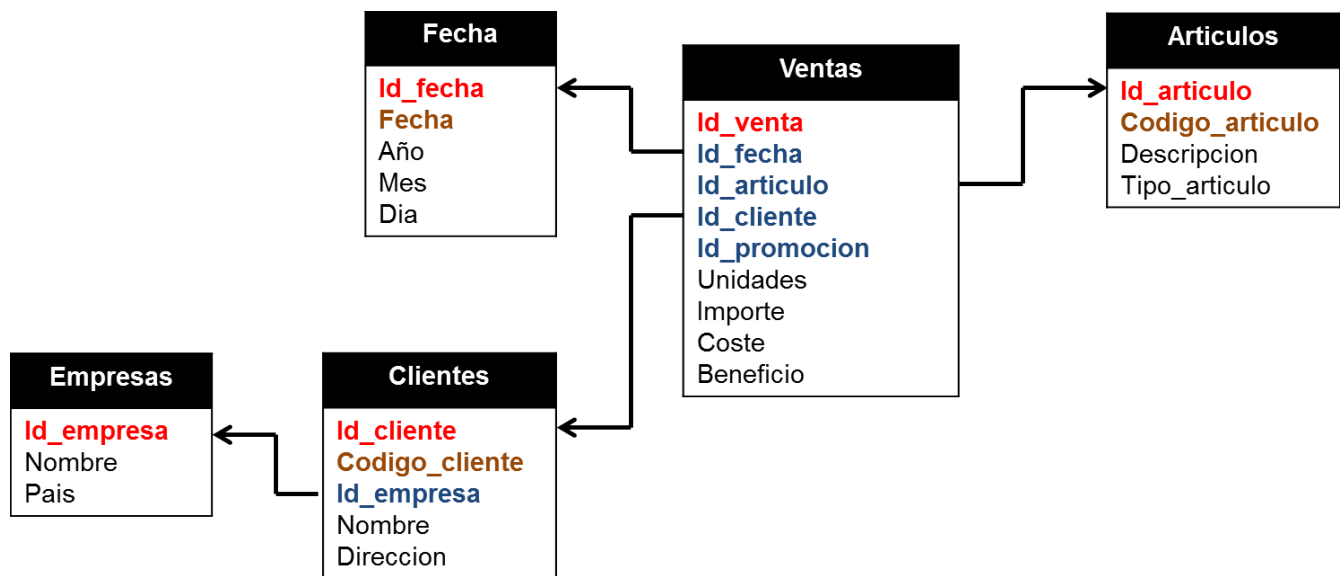
Sin embargo, las soluciones anteriores no son recomendadas, sobretodo si la tabla de hechos contiene cientos de millones de registros. Para abordar este problema, se crearía un producto cartesiano de todos los atributos (o todas las combinaciones posibles dentro del sistema), tal y como se muestra en la **Figura 3.11**.



**Figura 3.11 – Dimensión JUNK.**

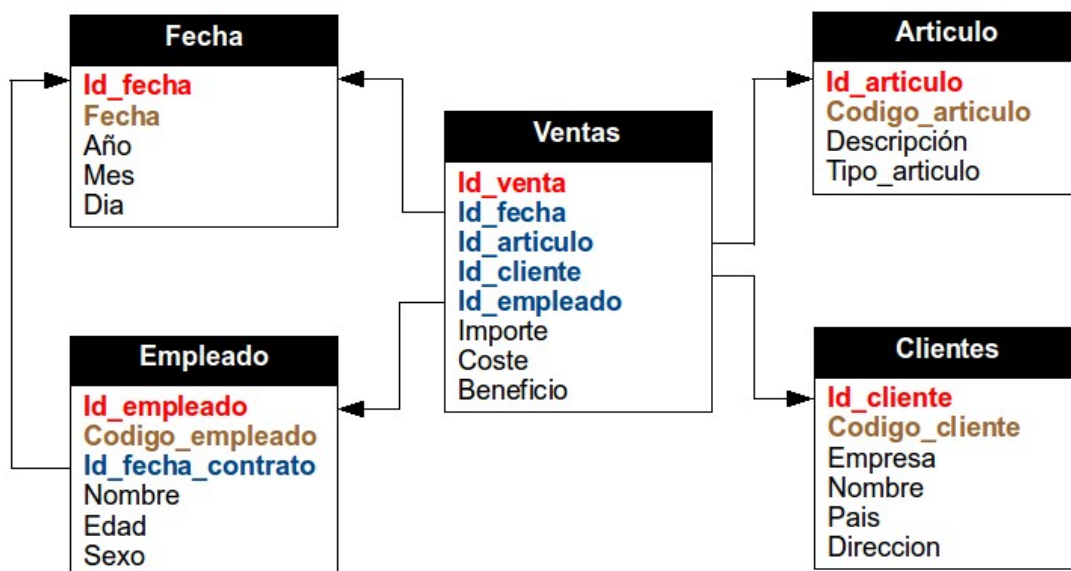
**Dimensiones Snowflake:** Cuando una relación jerárquica de una dimensión es normalizada, atributos de baja cardinalidad aparecen como una tabla secundaria, conectada a la dimensión base por una llave. Cuando este proceso es repetido con todas las jerarquías de las dimensiones, una estructura de multi-nivel es creada y es llamada “Snowflake” (Copo de nieve). Aunque los Snowflake representan los datos jerárquicos con más exactitud y ahorran espacio de almacenamiento, deberían evitarse ya que es difícil para el usuario del negocio entender y navegar este tipo de estructura; por otra parte, agregan mayor complejidad a los ETL encargados de procesar aquellas jerarquías sujetas al cambio y pueden afectar negativamente el rendimiento de las consultas.





**Figura 3.12 – Dimensión “Clientes” separada en una dimensión Snowflake.**

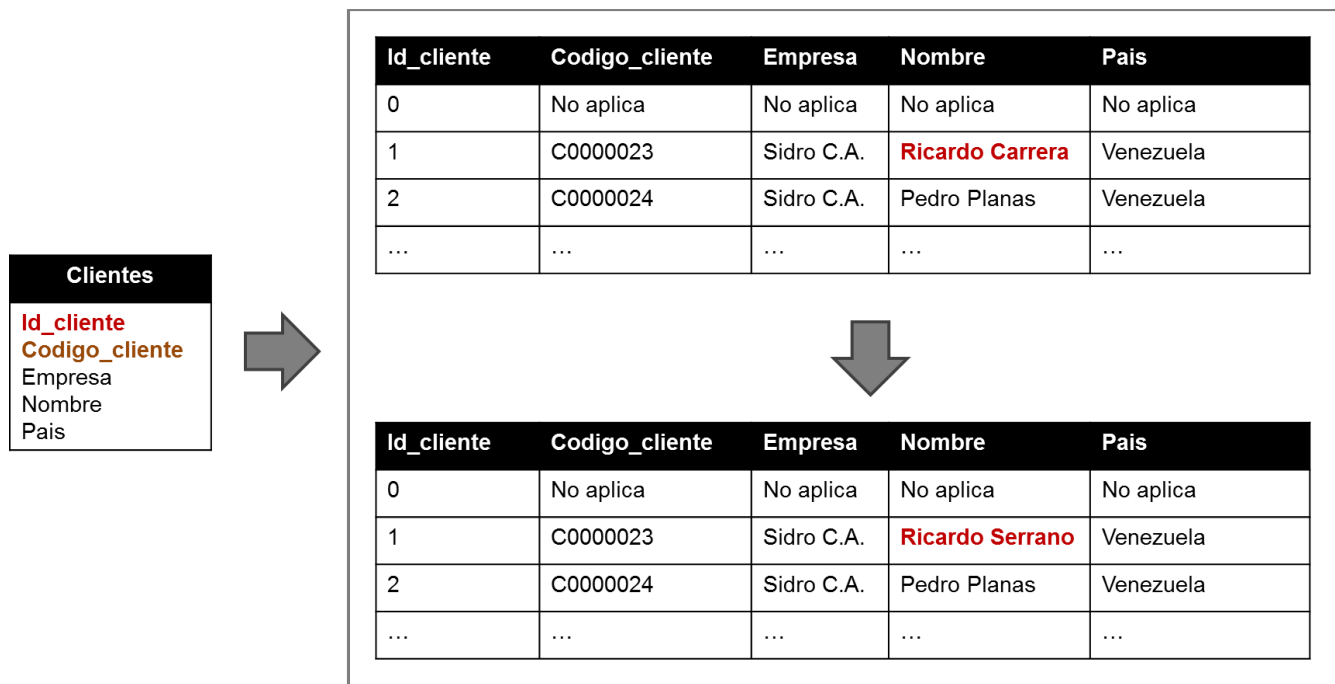
**Dimensiones Outrigger:** Las dimensiones Outrigger son tablas de dimensiones que tienen referencias a otras tablas de dimensiones. Son frecuentemente usadas cuando una dimensión estándar es referenciada en una dimensión, como por ejemplo, la fecha de contratación de un empleado. A diferencia de las dimensiones Snowflake, las dimensiones Outrigger no normalizan el modelo, sino que “retiran” una relación de la tabla de hechos para agregarla a la dimensión. En la mayoría de los casos, las relaciones entre dimensiones deberían ser relegadas a la tabla de hechos, donde ambas dimensiones son representadas como claves foráneas separadas.



**Figura 3.13 – Dimensión “Empleado” referenciando a la dimensión “Fecha”.**

**Dimensiones que cambian lentamente (SCD):** Estas dimensiones manejan técnicas de seguimiento para los diferentes cambios que pueden tener uno o más atributos de una dimensión:

- **SCD Tipo 0 (Mantener valores originales):** Con el tipo 0, los valores de los atributos nunca cambian, por lo que los hechos son agrupados por estos valores originales. Las tipo 0 son apropiadas para cualquier atributo “original”, tales como la cuenta de crédito de un cliente o un identificador duradero. También aplica a la mayoría de los atributos de una dimensión tiempo.
- **SCD Tipo 1 (Sobrescribir valores):** Con el tipo 1, los valores de los atributos se sobrescriben por los nuevos en la misma fila de la dimensión. Los atributos de las tipo 1 reflejan los valores asignados más recientes, por lo que esta técnica destruye la historia. Aunque este enfoque es fácil de implementar y no crea filas adicionales en la dimensión, es necesario ser cuidadoso ya que las agregaciones en la tabla de hechos y los cubos OLAP serán recalculados, por lo que afectaran los resultados.



**Figura 3.14 – Dimensión SCD de tipo 1.**

- **SCD Tipo 2 (Agregar nueva fila):** Las tipo 2 agregan nuevas filas en la dimensión con los valores de los atributos actualizados. Esto requiere generalizar la clave primaria de la dimensión más allá de la clave natural debido a que habrá varias filas describiendo a un mismo miembro. Cuando una fila nueva es creada para un miembro de la dimensión, una nueva clave primaria sustituta es asignada y usada como clave primaria en todas las tablas de hechos desde el momento de la actualización hasta que un cambio posterior cree una nueva clave de dimensión. Al menos se deben agregar tres columnas adicionales a la dimensión para manejar las tipo 2:
  - Fecha efectiva de la fila.
  - Fecha de expiración de la fila.

- Indicador actual de la fila (Versión).

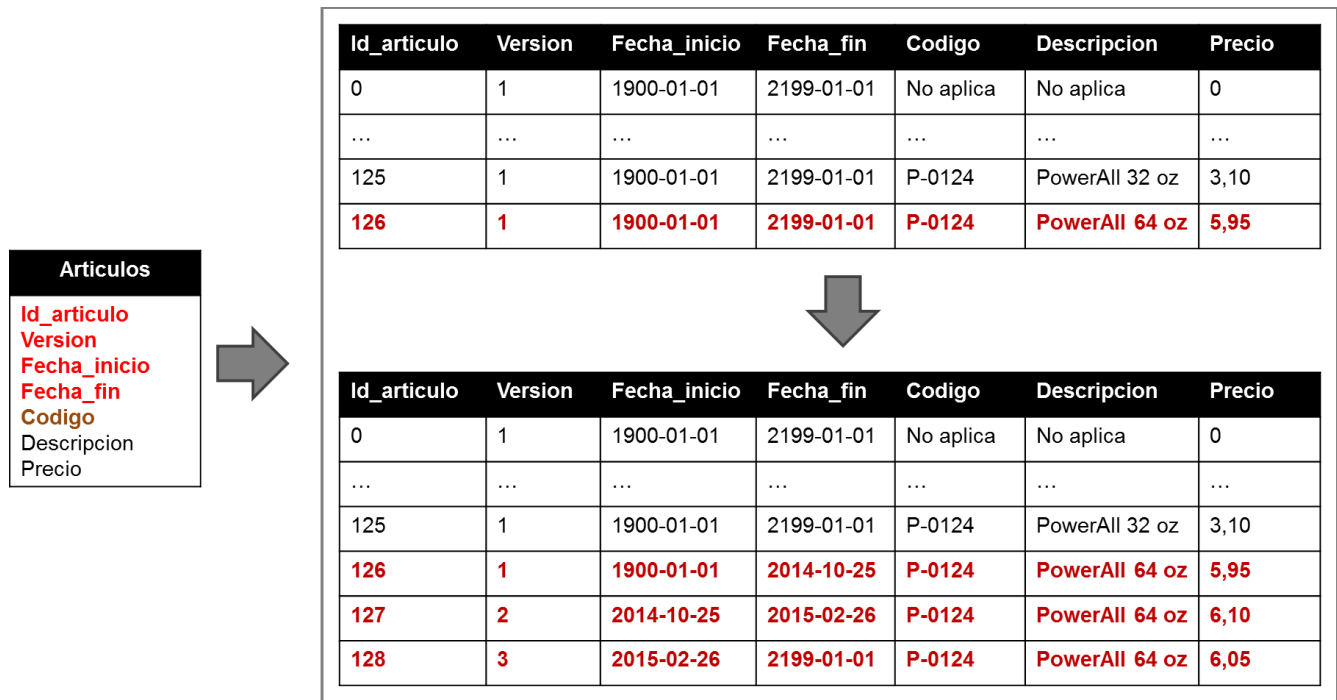


Figura 3.15 – Dimensión SCD de tipo 2.

- **SCD Tipo 3 (Agrega un nuevo atributo):** Las tipo 3 agregan un nuevo atributo (columna) en la dimensión para preservar el viejo atributo. Este tipo de dimensión a veces son llamadas “Realidad alternativa”. Un usuario de negocio puede agrupar y filtrar los datos ya sea por el valor actual o por el valor alternativo. Este tipo de dimensión son relativamente poco frecuente.

**Roles en una dimensión:** Una tabla de hechos puede tener varias referencias a una tabla de dimensión, y cada referencia cumple distintos roles para la dimensión. Por ejemplo, una tabla de hechos puede tener varias fechas, cada una de los cuales está representada por una clave foránea a la dimensión de fecha (ejemplo, fecha de pedido y fecha de entrega). Es esencial que cada clave foránea se refiera a una vista separada de la dimensión fecha, de manera que las referencias sean independientes. Estos puntos de vista diferentes que apuntan a las dimensiones (con nombres de columna de atributos únicos) se denominan roles.

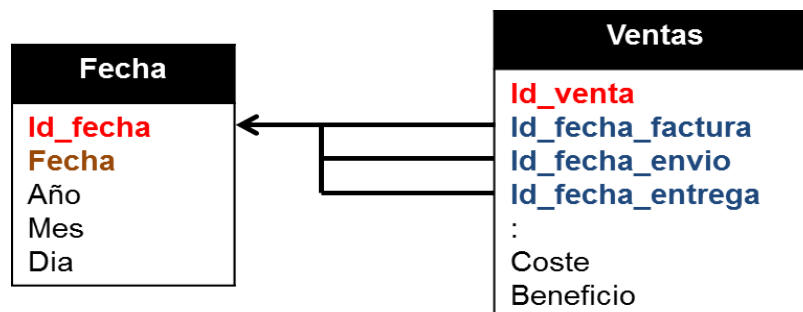
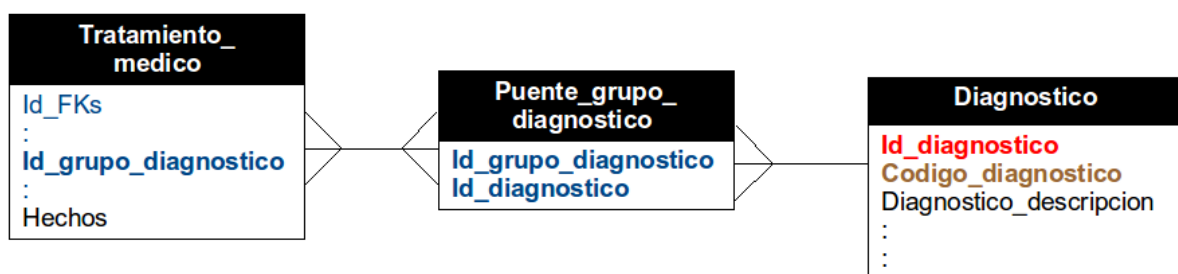


Figura 3.16 – Dimensión “Fecha” con diferentes roles.

**Dimensiones multi-valuadas y tablas puentes:** En un clásico modelo estrella, la relación entre los hechos y las dimensiones son de muchos-a-uno. Sin embargo, existen situaciones en las que la relación puede ser de muchos-a-muchos (dimensiones multi-valuadas). Por ejemplo, un paciente que recibe tratamiento médico puede tener múltiples diagnósticos simultáneos. En estos casos, las dimensiones multi-valuadas deben ser relacionadas con la tabla de hechos a través de una tabla puente que enlace los registros de la tabla de hechos con uno o varios registros de la dimensión, por medio de una clave que agrupe los diferentes registros.



**Figura 3.17 – Diseño de tablas puente para dimensiones multi-valuadas.**

Si un paciente tiene tres diagnósticos, tendrá asignado un grupo en la tabla puente con tres registros correspondientes.

Puente_grupo_diagnostico		Diagnostico		
Id_grupo_diagnostico	Id_diagnostico	Id_diagnostico	Codigo_diagnostico	Diagnostico_descripcion
1	1	1	CA03	Cáncer de pulmón
1	3	2	CZ01	Fibrilación auricular
1	5	3	CA02	Cáncer de hígado
2	4	4	AP07	Apendicitis aguda
2	5	5	TB05	Bronquitis y enfisema por exposición al tabaco

**Figura 3.18 – Relación de los registros entre las tablas puentes y las dimensiones multi-valuadas.**

Las tablas puentes son escalables y flexibles: pueden manejar las dimensiones con múltiples valores asociados con el grano de un evento de la tabla de hechos, y pueden manejar un amplio número de valores sin alterar el diseño de la base de datos. Sin embargo, las tablas puentes tienen sus desventajas. Algunas herramientas de BI luchan por generar el SQL que cruce satisfactoriamente la tabla puente, por lo que la facilidad de uso se ve comprometida. Por otro lado, si la combinación de grupos crece desmesuradamente, el rendimiento se verá afectado. Existen varias técnicas para evitar las tablas puentes; sin embargo, hay que tener en cuenta que también tienen sus desventajas:

- 1. Alterar el grano de la tabla de hechos:** Las relaciones muchos-a-muchos se resuelven mejor en la tabla de hechos. En el caso anterior, es posible que el

diagnóstico sea el grano de la tabla de hechos, por lo que cada registro dentro de la tabla corresponderá a un diagnóstico específico de un paciente. Este grano sólo será lógico para los usuarios del negocio que analicen este escenario.

2. **Designar un valor primario:** Declarar un valor primario, ya sea con una única clave foránea en la tabla de hechos o con un único atributo en la dimensión, elimina la relación muchos-a-muchos. En este escenario todos los nombres de las columnas serían precedidas por la palabra “primario”. Por supuesto, si las reglas del negocio no especifican la relación primaria, sería imposible determinar la misma. Por otro lado, el análisis basado únicamente en la relación primaria sería incompleto y/o engañoso debido que fueron ignorados el resto de los atributos.
3. **Agregar varios atributos a la tabla dimensión:** Por ejemplo, colocar varias columnas de diagnóstico dentro de la dimensión paciente (diagnóstico primario, diagnóstico secundario, diagnóstico terciario, etc.). Esta técnica sólo es apropiada para un número de opciones fijas y limitadas, de lo contrario se podrían obtener N cantidad de columnas por cada atributo diferente. Este enfoque no es escalable ya que agregar nuevos atributos requiere alterar la tabla.
4. **Agregar todos los valores concatenados en un sólo texto con un delimitador:** En el caso de los pacientes, se puede agregar un atributo de tipo texto a la dimensión paciente con los diagnósticos concatenados por medio de un delimitador. Por ejemplo: “Bronquitis|Hipertensión|Apendicitis”. Esta técnica tiene grandes desventajas: las consultas deberán hacer búsquedas con **contains/like**, las cuales son notoriamente lentas. Podría haber ambigüedad con respecto a las mayúsculas y minúsculas de los textos concatenados. No sería apropiado para una larga lista de valores. Por último, no será posible realizar operaciones de agregación por alguno de los valores concatenados.

**Valores nulos en las dimensiones:** Los atributos nulos en una dimensión se producen cuando las filas de una dimensión no han sido totalmente llenadas, o cuando hay atributos que no aplican a todas las filas de la dimensión. En ambos casos se recomienda la sustitución con una cadena descriptiva, ya sea “Desconocido” o “No aplica”, en lugar del valor nulo. Los valores nulos en los atributos de las dimensiones deben evitarse debido a que las diferentes bases de datos manejan la agrupación y las restricciones de los nulos de manera inconsistente.

## Manejando jerarquías en las dimensiones

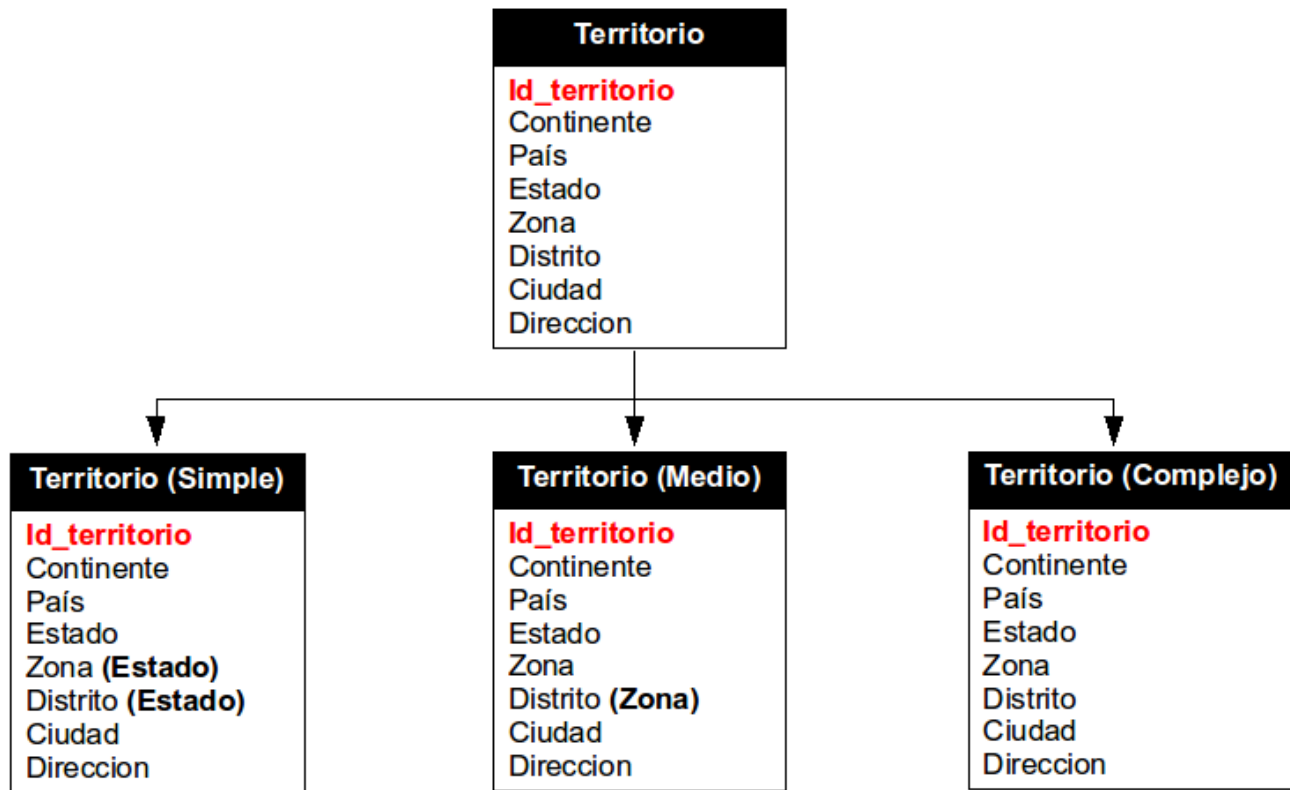
Las jerarquías en las dimensiones son muy comunes, por lo que a continuación se explicarán los diferentes enfoques para tratarlas.

**Jerarquías con profundidad fija:** Las jerarquías con profundidad fija son una serie de relaciones uno-a-muchos, tales como producto, marca, categoría y departamento. Una jerarquía de profundidad fija es definida cuando los niveles de la jerarquía tienen nombres establecidos y cumplen con la misma estructura. Estos niveles deben aparecer como atributos en la tabla de dimensión como se mostró en la **Figura 3.10**. Esta jerarquía es la más simple de entender y navegar.

**Jerarquías ligeramente des-balanceadas con profundidad variable:** Estas jerarquías no tienen un número fijo de niveles, pero el rango de profundidad es pequeño. Por ejemplo, la jerarquía de territorio puede mostrar tres posibilidades:

- Simple: Continente → País → Estado → Ciudad → Dirección.
- Media: Continente → País → Estado → Zona → Ciudad → Dirección
- Compleja: Continente → País → Estado → Zona → Distrito → Ciudad → Dirección.

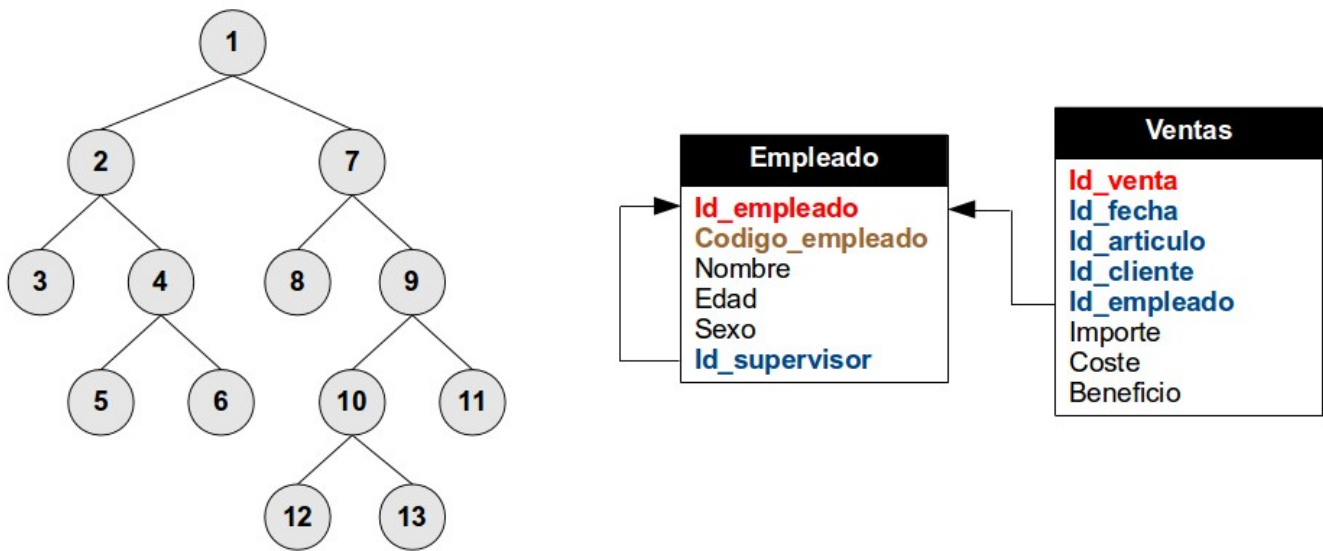
Esta jerarquía se puede diseñar como una jerarquía de profundidad fija, representando los tres tipos de territorios en una sola jerarquía que soporte la posibilidad más compleja. En este caso, los niveles que no existen en las otras jerarquías, se completarían con los valores del nivel superior inmediato. El territorio simple no guarda la zona ni el distrito, por lo que copiaríamos el Estado en esos dos atributos. El territorio medio no guarda el distrito, por lo tanto copiaríamos la Zona en el Distrito.



**Figura 3.19 – Tabla de territorio, representando la jerarquía para los tres tipos.**

Ejemplo de jerarquía simple dentro de la dimensión territorio: Continente: **América** → País: **Venezuela** → Estado: **Distrito Federal** → Zona: **Caracas** → Distrito: **Caracas** → Ciudad: **Caracas**. Esta técnica funciona bien con la jerarquía territorio, ya que tiene entre cuatro y seis niveles. Sí la jerarquía presenta de cuatro a ocho, o diez o más niveles, este enfoque no funcionaría. Es importante que los nombres de los atributos tengan sentido.

**Jerarquías des-balanceadas con tablas puentes:** Las jerarquías des-balanceadas con profundidad variable son difíciles de modelar y consultar en una base de datos relacional. Esta jerarquía es usada para manejar la relación recursiva padre/hijo.



**Figura 3.20 – Representación de una estructura árbol de padre/hijo (Izquierda). Relación de una tabla de hechos con una dimensión padre/hijo con puntero recursivo (Derecha).**

La forma clásica de representar una estructura de árbol padre/hijo es colocando una clave foránea en cada registro que apunte a su padre dentro de la misma tabla. Aunque algunas herramientas OLAP soporten las tablas recursivas, son ineficientes ya que necesitan recorrer toda la tabla para obtener cualquier resultado. Por otra parte, sería difícil mantener atributos de una dimensión que cambia lentamente de tipo 2, ya que un cambio de la clave en un nodo de alto nivel requiere un cambio de clave para cada miembro que esté por debajo de él, hasta el fondo del árbol. La solución a estos problemas es la construcción de una tabla puente que contenga un registro por cada camino posible en la jerarquía, permitiendo recorrer la jerarquía completa desde cualquier punto con el uso de simples consultas SQL. La **Figura 3.21** muestra la estructura de la tabla puente, tomando como ejemplo la **Figura 3.20**. La primera columna es la clave primaria del padre, la segunda columna es la clave primaria del hijo. Cada fila debe contener cada padre e hijo posible, incluyendo la conexión del padre consigo mismo. La tercera columna muestra la distancia de niveles que hay entre el padre y el hijo. Por último, dos columnas booleanas que especifican si el padre es raíz y si el hijo es una hoja.

Mapa_tabla_puente
Id_padre
Id_hijo
Distancia
Raiz
Hoja

Mapa_tabla_puente				
Id_padre	Id_hijo	Distancia	Raiz	Hoja
1	1	0	TRUE	FALSE
1	2	1	TRUE	FALSE
1	3	2	TRUE	TRUE
1	4	2	TRUE	FALSE
1	5	3	TRUE	TRUE
1	6	3	TRUE	TRUE
1	7	1	TRUE	FALSE
1	8	2	TRUE	TRUE
1	9	2	TRUE	FALSE
1	10	3	TRUE	FALSE
1	11	3	TRUE	TRUE
1	12	4	TRUE	TRUE
1	13	4	TRUE	TRUE
2	2	0	FALSE	FALSE
2	3	1	FALSE	TRUE
2	4	1	FALSE	FALSE
2	5	2	FALSE	TRUE
2	6	2	FALSE	TRUE
3	3	0	FALSE	TRUE
4	4	0	FALSE	FALSE
4	5	1	FALSE	TRUE
4	6	1	FALSE	TRUE
5	5	0	FALSE	TRUE
6	6	0	FALSE	TRUE
7	7	0	FALSE	FALSE
7	8	1	FALSE	TRUE
7	9	1	FALSE	FALSE
7	10	2	FALSE	FALSE
7	11	2	FALSE	TRUE
7	12	3	FALSE	TRUE
7	13	3	FALSE	TRUE
8	8	0	FALSE	TRUE
9	9	0	FALSE	FALSE
9	10	1	FALSE	FALSE
9	11	1	FALSE	TRUE
9	12	2	FALSE	TRUE
9	13	2	FALSE	TRUE
10	10	0	FALSE	FALSE
10	12	1	FALSE	TRUE
10	13	1	FALSE	TRUE
11	11	0	FALSE	TRUE
12	12	0	FALSE	TRUE
13	13	0	FALSE	TRUE

**Figura 3.21 – Ejemplo de los registros de la tabla puente.**

El ejemplo muestra que hay 13 caminos desde el nodo número 1; 5 caminos desde el nodo número 2; 1 camino desde el nodo número 3, y así sucesivamente.



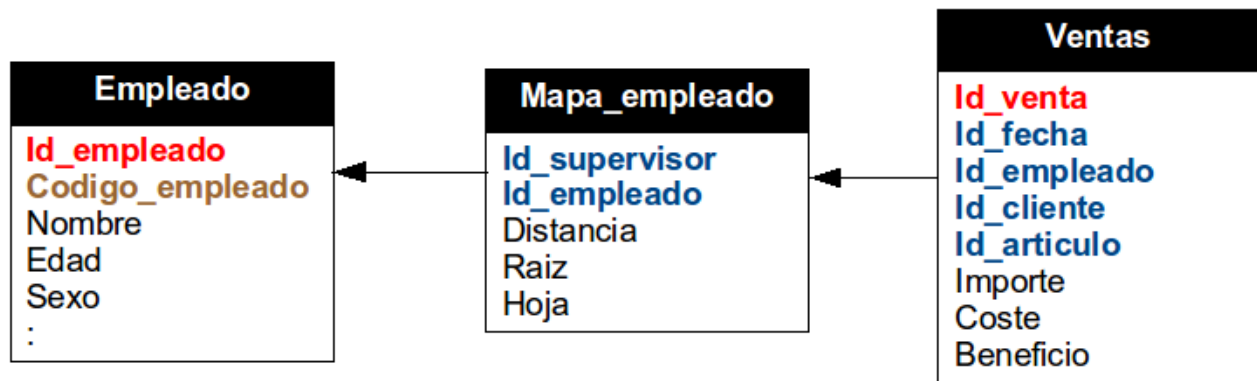


Figura 3.22 – Unión de la tabla puente con la dimensión y la tabla de hechos.

### Extensibilidad del esquema

Los modelos dimensionales son resistentes cuando las relaciones de los datos cambian. Los siguientes cambios pueden ser implementados sin alterar cualquier consulta o aplicación BI existente, y sin alterar los resultados:

- **Nuevas medidas:** Para agregar nuevas medidas, hay que crear nuevas columnas en la tabla de hechos. Es necesario rellenar las filas anteriores al cambio y evitar los valores nulos.
- **Nuevas dimensiones:** Para agregar una nueva dimensión, hay que añadir una clave de referencia en la tabla de hechos y cargar los nuevos valores en la tabla de hechos.
- **Nuevos atributos en la dimensión:** Para agregar nuevos atributos en la dimensión, hay que agregar nuevas columnas en la tabla de dimensión. Si los nuevos atributos sólo están disponibles a partir de una fecha, en las anteriores deben figurar como no disponibles.
- **Redefinir la granularidad:** El grano de una tabla de hechos se puede hacer más atómico al añadir atributos a una tabla de dimensión existente, y luego reiterando la tabla de hechos al grano más bajo, siendo cuidadoso de preservar los nombres de las columnas existentes en las tablas de hechos y dimensiones.

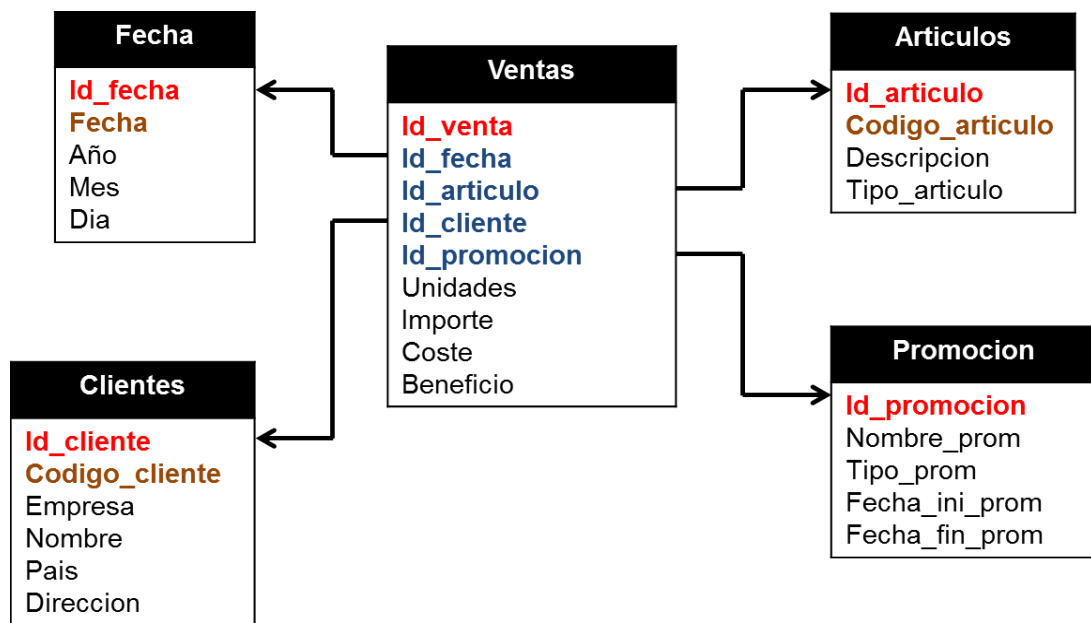
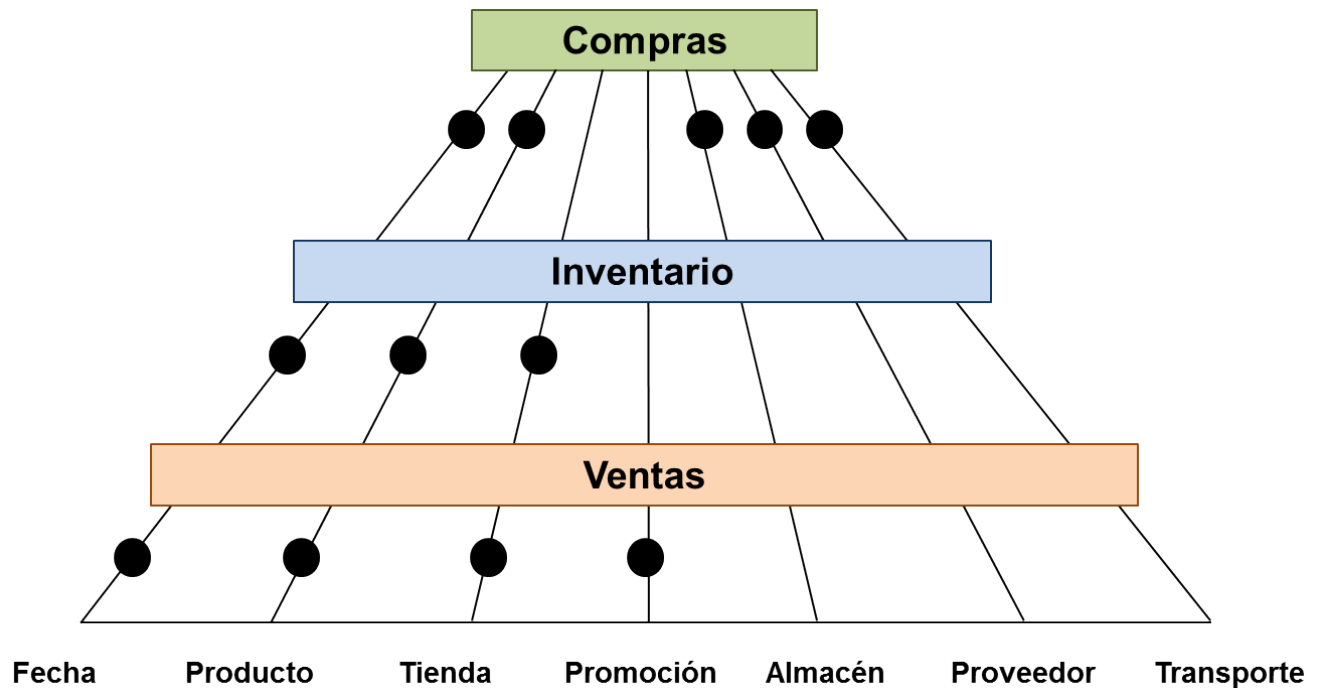


Figura 3.23 – Inserción de la nueva dimensión “Promoción” al esquema estrella.

## Arquitectura en bus del almacén de datos

La arquitectura en bus, creado por el grupo Kimball en los años 90, es un enfoque que permite un desarrollo incremental de los sistemas de Inteligencia de negocio y almacenes de datos, ya que descompone el proceso de planificación en segmentos manejables, centrándose en los procesos de negocios principales de la organización, junto con las dimensiones conformadas asociadas a cada proceso.



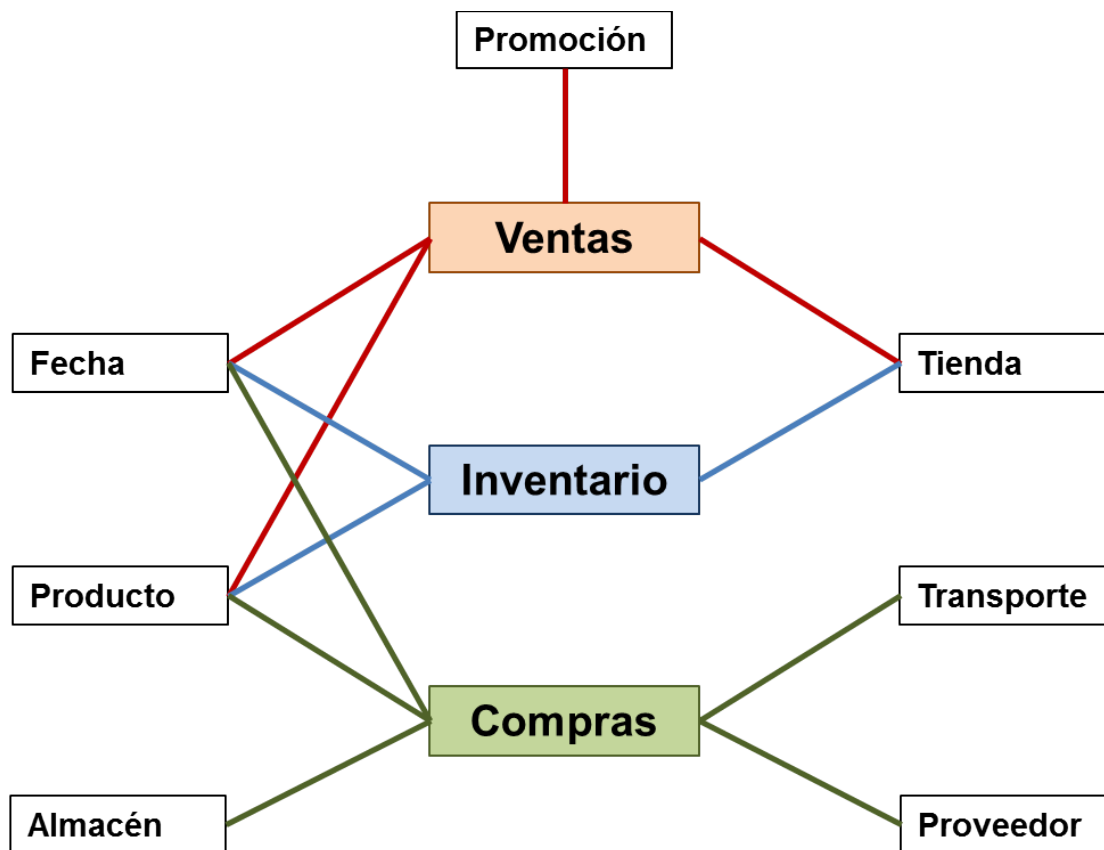
**Figura 3.24– Representación de la arquitectura en bus.**

### Dimensiones conformadas

Las dimensiones conformadas son dimensiones comunes y estandarizadas, administradas en el proceso ETL, que pueden ser utilizadas por varias tablas de hechos. Las dimensiones conformadas entregan atributos descriptivos consistentes entre los diferentes modelos dimensionales, y además, tienen la capacidad de integrar los datos de los múltiples procesos de negocio. Al reutilizar las dimensiones conformadas, se reduce el tiempo de lanzamiento del sistema, eliminando esfuerzos redundantes de diseño y desarrollo.

### Dimensiones encogidas

Son dimensiones conformadas que contienen un subconjunto de filas y/o columnas de una dimensión base. Estas dimensiones son necesarias para las tablas agregadas y los procesos de negocios que capturan los datos en un alto nivel de granularidad, como por ejemplo, una predicción por mes y marca, en vez de la fecha completa y el producto asociado a los datos de la venta.



**Figura 3.25 – Dimensiones conformadas para tres tablas de hechos.**

### Matriz en bus del almacén de datos

La matriz en bus es una herramienta clave de diseño que representa los principales procesos de negocio y las dimensiones asociadas a ellos. Esta arquitectura asegura que los datos en el sistema DW/BI se puedan integrar en toda la organización.

	Fecha	Producto	Tienda	Promoción	Almacén	Proveedor	Transporte
Compras	X	X			X	X	X
Inventario	X	X	X				
Ventas	X	X	X	X			

**Figura 3.26 – Matriz en bus.**

### Data marts

El data mart es un subconjunto del almacén de datos, generalmente de un solo proceso de negocio, que está orientado a un departamento o grupo de usuarios. Normalmente contiene información de un esquema estrella, por lo que se suelen utilizar como sinónimos, aunque conceptualmente son diferentes. Es normal que los diferentes usuarios accedan a un subconjunto específico de los datos, por lo que descomponer el

almacén de datos en diferentes data marts suele mejorar el rendimiento de las consultas ya que reduce el volumen de datos que se requieren para contestar las preguntas de los usuarios. Los data marts se utilizan para:

- Segmentar la información en diferentes plataformas de hardware.
- Facilitar el acceso de las herramientas de consultas.
- Dividir los datos para controlar mejor el acceso.
- Mejorar los tiempos de respuestas.

Para asegurar la consistencia, los data marts deben ser cargados a partir del almacén de datos, y no desde las fuentes de datos.

Aunque los data marts son una solución ante problemas de rendimiento y control de acceso, estos suponen costos adicionales de hardware, software y accesos a la red.

## 4. Fases en el diseño dimensional

Para diseñar el modelo dimensional se deben seguir los siguientes pasos:

1. Seleccionar el proceso de negocio.
2. Declarar el grano.
3. Identificar las dimensiones.
4. Identificar los hechos (Medidas).

Para ello es necesario considerar las necesidades del negocio junto con la realidad de las fuentes de datos durante las mesas de trabajo del modelado. Una vez definido el proceso de negocio, el grano, la dimensión y los hechos, el equipo de diseño se encargará de determinar los nombres de las tablas y columnas, y las reglas del negocio. Los representantes del negocio deben participar en esta actividad de diseño detallado para asegurar el éxito del proyecto.

### Los procesos de negocio

Los procesos de negocio son las actividades operativas realizadas por la organización, por ejemplo, tomar un pedido (facturación) o el registro de los estudiantes para una clase. Los eventos de los procesos de negocios generan o capturan medidas de rendimiento, que se traducen en hechos en la tabla de hechos. La mayoría de las tablas de hechos se centran en los resultados de un único proceso de negocio. La selección del proceso de negocio es importante porque define un objetivo de diseño específico y permite la declaración del grano, las dimensiones y los hechos.

### El grano

El grano (o granularidad) representa el nivel más atómico por el cual se definen los datos. Por ejemplo, no es lo mismo contar el tiempo por horas (grano fino) que por semanas (grano grueso); o en el caso de los productos, se puede considerar cada variante de un mismo artículo como un producto (por ejemplo, en una empresa textil, cada talla y color de pantalón podría ser un producto diferente, en otras palabras, un grano fino) o agrupar todos los artículos de una misma familia considerándolos como un único producto (por ejemplo, el producto pantalón genérico, lo que sería un grano grueso). Como se puede observar, la granularidad afecta a la cardinalidad, tanto de las dimensiones como de la tabla de hechos, a mayor granularidad (grano más fino) mayor será el nivel de detalle y el número de registros final de la tabla de hechos y dimensiones. Los diferentes granos no deben ser mezclados en

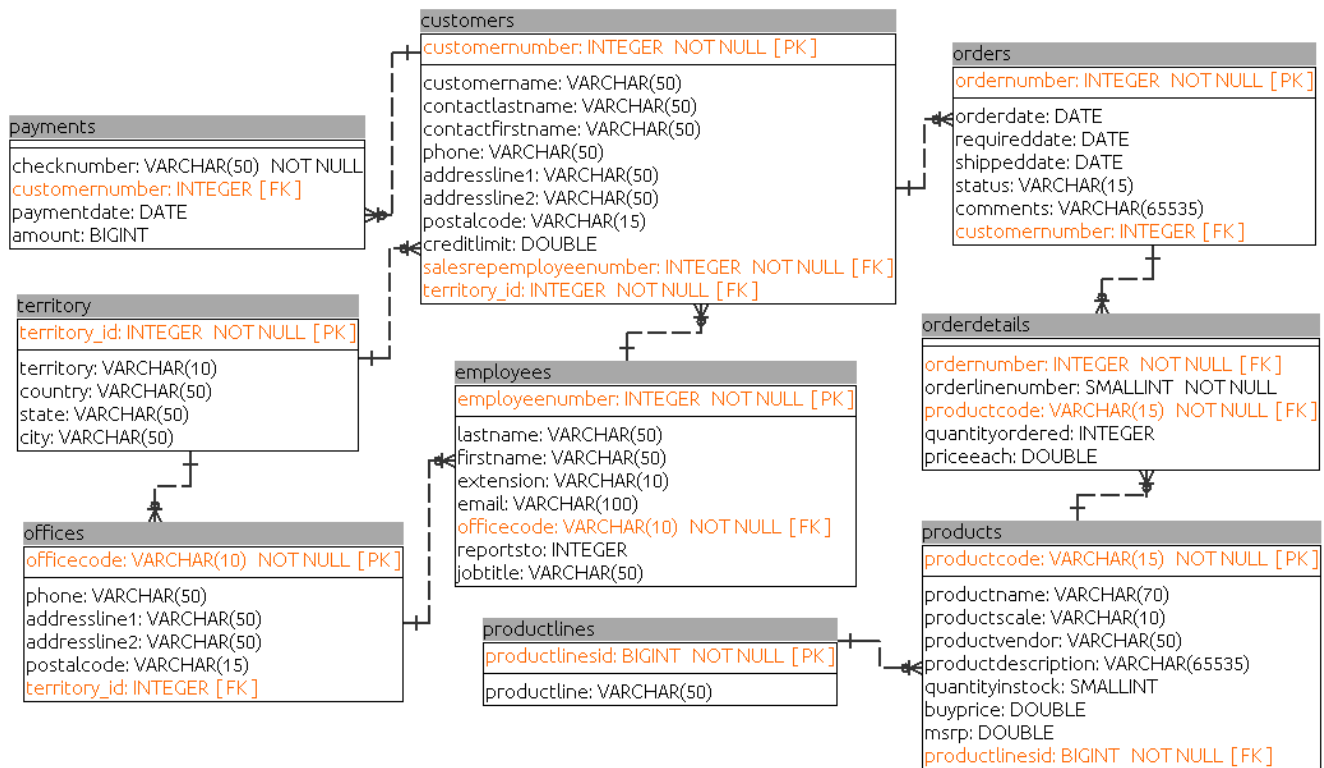
una sola tabla de hechos ya que las agrupaciones y cálculos arrojarán resultados erróneos.

## **Diseño del modelo dimensional de Steel Wheels**

### **Seleccionar proceso de negocio**

Steel Wheels es una empresa ficticia, creada por la comunidad de Pentaho, de ventas al por menor de vehículos a escalas a nivel mundial. La empresa Steel Wheels tiene un sistema de gestión de pedidos realizados por sus clientes desde el año 2003 hasta el año 2005. Entre las necesidades de análisis, se encuentran:

- Producto más vendido, más rentable, número de unidades vendidas por país, por empleado, entre otros.
- Evolución de las ventas realizadas en cada país en los últimos años. ¿Cuál es la tendencia?
- ¿En que países nos hemos introducido y en cuáles hemos perdido cuota en el mercado?
- Ingresos obtenido por cada empleado y por jefe.
- Comparativa de ventas del mismo producto en diferentes meses y países.
- ¿Cuáles productos ofrecen mayor rentabilidad?
- ¿Cuántos clientes nuevos hemos conseguido este año?
- Gasto promedio por cliente.
- ¿Quiénes son nuestros mejores clientes?



**Figura 4.1 – Modelo de Entidad-Relación de la empresa Steel Wheels.**

## Declarar el grano

El proceso de negocio de ventas de Steel Wheels se basa en las facturas generada por cada compra de sus clientes. Las facturas tienen una o más líneas que detallan la cantidad y el costo de cada producto. Con cada línea de detalle, es posible obtener la información necesaria para cumplir las necesidades de análisis de la empresa, por lo que será el grano del modelo dimensional. Es posible colocar la factura como grano, ya que esta guarda el total de la transacción, el cliente y la fecha; sin embargo, el nivel de detalle de la factura no alcanza a obtener la información de las cantidades y montos de los productos vendidos.

## Seleccionar las dimensiones

Las dimensiones son los datos descriptivos del esquema estrella. Podemos identificar las siguientes dimensiones a partir de la **Figura 4.1**:

- Empleado (Employees)
- Oficina (Offices)
- Localidad (Territory)
- Clientes (Customers)
- Fecha de factura, de envío y de entrega (Order date, shipped date, required date)
- Estatus (Orders)
- Productos (Products)

- Línea de productos (Productlines)

Las dimensiones “Productos” y “Línea de productos” se pueden unir en una sola dimensión de producto, siendo la línea de producto una jerarquía interna en la dimensión. Por otra parte, la dimensión “Empleado” y “Oficina” se pueden unir ya que el empleado pertenece a una oficina, y esta información puede ser compactada en una sola dimensión empleado:

- Empleado (Employees, Offices)
- Localidad (Territory)
- Clientes (Customers)
- Fecha: esta dimensión cumple con diferentes roles: la fecha de la factura, la fecha de envío, y la fecha de entrega (Order date, shipped date, required date)
- Estatus (Orders)
- Productos (Products, Productlines)

La dimensión “Localidad” puede ser agregada dentro de las dimensiones “Empleado” y “Cliente”. Las tablas de dimensiones contienen muchos atributos y pueden redundar la información debido a su baja cardinalidad. Por último, empleado tiene una relación recursiva padre/hijo. Esta tabla requiere una tabla puente para representar la jerarquía.

- Empleado (Employees, Offices, Territory) con tabla puente
- Clientes (Customers, , Territory)
- Fecha (Order date, shipped date, required date)
- Estatus (Orders)
- Productos (Products, Productlines)

**NOTA:** Aunque “Estatus” es sólo un atributo, estará en una tabla de dimensión aparte. En la mayoría de los casos, si un atributo es usado para filtrar o agrupar, este pertenece a una tabla de dimensión.

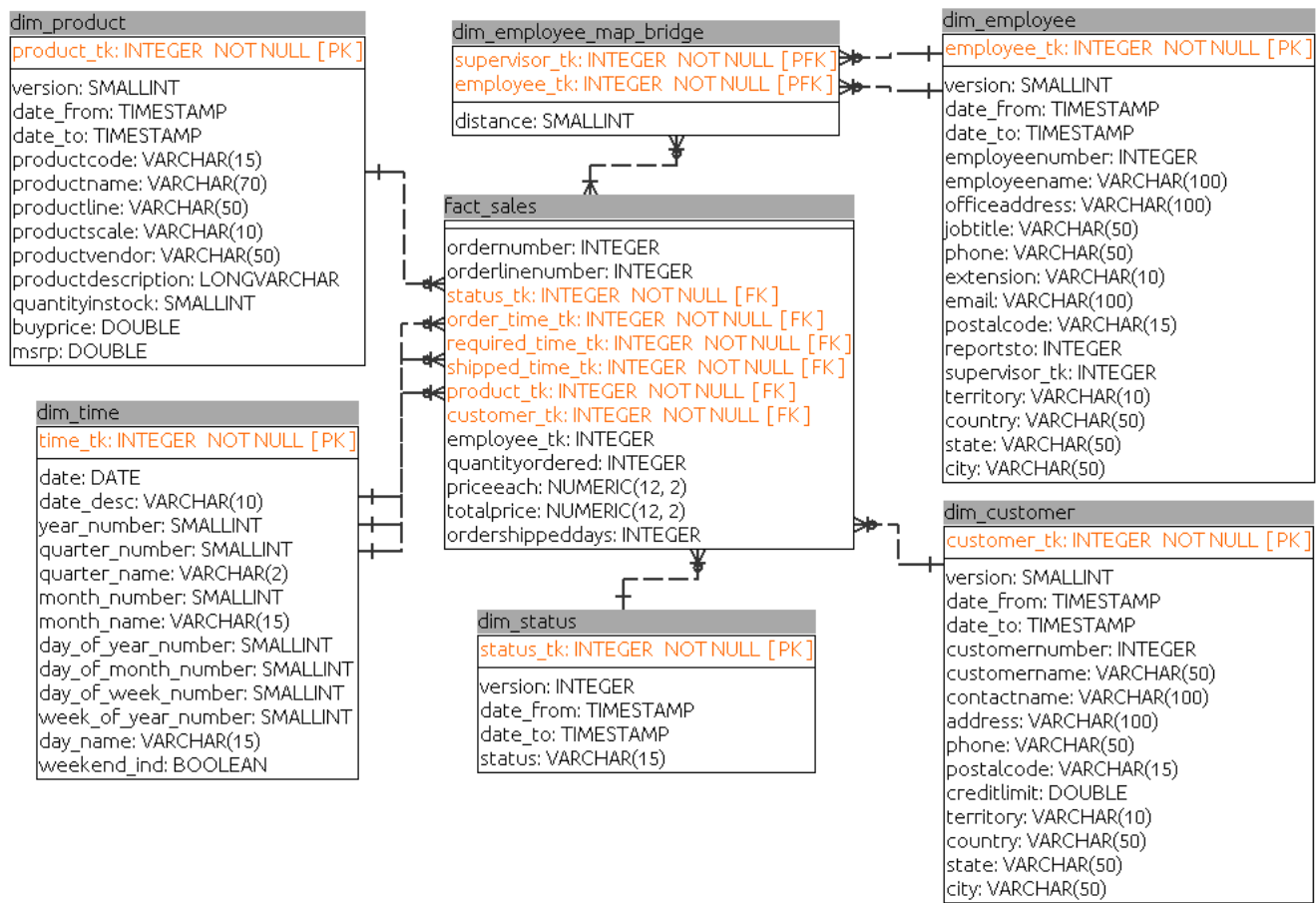
## Seleccionar los hechos

Los hechos son los valores de las medidas de negocio. En el modelo, son los campos por los cuales vamos a medir el rendimiento y obtener los datos numéricos para los análisis:

- Cantidad del producto (Quantityordered)
- Precio de cada producto (Priceeach)
- Cantidad de días entre la fecha de factura y la fecha de envío (shipped date - order date).

Ya definida las dimensiones y los hechos, podemos crear el esquema estrella de la **Figura 4.2:**





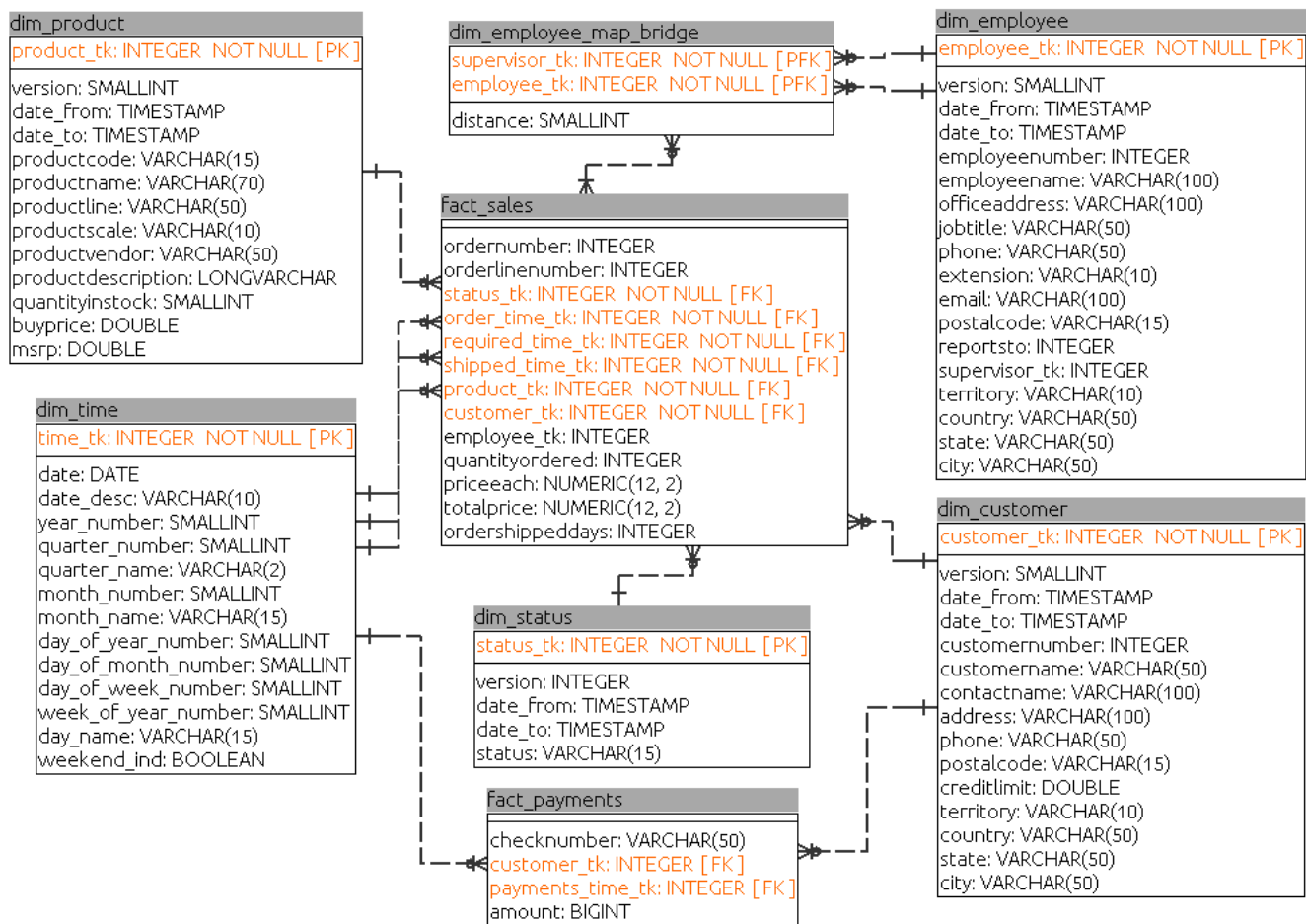
**Figura 4.2 – Esquema estrella del proceso de negocio de ventas de la empresa Steel Wheels.**

## Extender el esquema

Además del proceso de negocio de ventas, el sistema de Steel Wheels también almacena información sobre los pagos realizados por los clientes. Debido a que es un proceso diferente, este requiere el mismo análisis de diseño dimensional:

- **Declarar grano:** Como podemos ver en la **Figura 4.1**, los pagos son representados por una tabla única que contiene el monto pagado por cliente en una fecha determinada, por lo que cada registro de pago será el grano de este proceso.
- **Seleccionar dimensiones:** La tabla de pagos muestra únicamente dos dimensiones: el cliente que realizó un pago y la fecha de ejecución.
- **Seleccionar hechos:** El único hecho medible que se puede obtener de este proceso es el monto del pago.

Ya identificada las dimensiones y los hechos del proceso de negocio de pagos, podemos contruir el esquema estrella. Sin embargo, no será un modelo aparte. Este proceso tiene dimensiones compartidas con el proceso de ventas, por lo que siguiendo la arquitectura en bus descrita en el **Tema 3**, podemos integrar ambos modelos en un almacén de datos:



**Figura 4.3 – Almacén de datos con los procesos de negocios de ventas y pagos**

**Nota:** Debido a la simplicidad de este modelo, es posible notar a simple vista la relación de dimensiones entre modelos. En caso de que el modelo sea mucho mayor, es necesario realizar una matriz en bus para identificar todos los procesos involucrados y las dimensiones que comparten.

## Base de datos

Antes de comenzar el desarrollo del proceso ETL, es necesario tener instalado un sistema de gestión de base de datos para poder acceder al modelo OLTP y el modelo OLAP de Steel Wheels. En el siguiente capítulo, se utilizará como referencia la base de datos **PostgreSQL**; sin embargo, puede utilizarse **MySQL** como una base de datos alterna.

## Restaurar bases de datos

Una vez instalado el sistema gestor de base de datos, procedemos con la restauración de los modelos. Para ello, cree dos bases de datos: **steel\_wheels\_oltp** y **steel\_wheels\_olap** e importe la información de los archivos que se encuentran en el directorio **db**. En este directorio se encuentran respaldos para restaurar la información en **PostgreSQL** y **MySQL**.