VNIVERSIDAD NACIONAL
AVTNOMA DE
MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

ESTUDIO COMPARATIVO DE TÉCNICAS PARA EL DISEÑO Y CONSTRUCCIÓN DE ALMACENES DE DATOS

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN INGENIERÍA (COMPUTACIÓN)

PRESENTA:
OCTAVIO VÁZQUEZ TOLEDO

TUTOR:
DRA. AMPARO LÓPEZ GAONA
FACULTAD DE CIENCIAS-UNAM

MÉXICO, D. F. ABRIL DE 2014.

Agradecimientos

A mi familia quienes han sido la razón de mí existir y para cumplir un objetivo más en mi vida, la maestría, y que sin su apoyo me hubiese sido muy difícil cumplirlo.

A mis padres

Con mucho amor y cariño a mis padres Georgina y Octavio que han compartido y asumido compromisos y obligaciones más que siempre. Por inculcarme el valor de la educación, y darme más de lo necesario en esta etapa profesional y en mi vida.

A mis hermanitas

A quienes adoro Eunice, Ilce y Narely, por darme su amor, cariño, compresión, confianza, hermandad y su gran apoyo.

A personitas especiales

A mis abuelitos, tíos y primos que me han enseñado a valorar la importancia de la gran familia que tengo. En especial a mi cuñado y a mis sobrinos por recordarme cuantas cosas he pasado para llegar hasta aquí.

A la Dra. Amparo López Gaona

Gracias a la Dra. Amparo por brindarme la oportunidad de realizar éste trabajo bajo su tutela, por confiar en mí, por sus consejos, ánimos y por compartir sus experiencias profesionales, lo valoro mucho y le estaré siempre agradecido. Espero nunca se olvide de mí ya que siempre fui inoportuno en todo momento.

A mis sinodales

A mis sinodales por sus valiosos comentarios y puntos de vista sobre mi trabajo.

A mis amigos

En agradecimientos por su confianza, consejo y por ser una gran compañía, que han estado, y estarán en los mejores y peores momentos junto a mí.

"Por mi raza hablará el espíritu".

Índice

Introducción

Los almacenes de datos hoy en día son el centro de atención de grandes empresas y organizaciones, ya que constituyen uno de los soportes fundamentales para el proceso de toma de decisiones. De ahí la importancia de que la información guardada en ellos sea confiable y de calidad.

Uno de los principales problemas que enfrentan los creadores de almacenes de datos, es el diseño de su almacén. Existen modelos de datos básicos que son: el modelo relacional considerado por William H. Inmon, el modelo multidimensional considerado por Ralph Kimball y el modelo de Bernabeu Ricardo Dario, que consiste en que de acuerdo a las operaciones que se deseen o requieran desarrollar, el almacén de datos puede adoptar el modelo relacional o multidimensional pero con ciertos cambios en los pasos de su metodología.

Las organizaciones disponen de una variedad de opciones en términos de software, herramientas y enfoques de desarrollos. Por lo tanto elegir la metodología adecuada para el desarrollo de un almacén de datos, requiere de conocer y comprender las principales: La de Inmon, la de Kimball y la de Bernabeu.

Objetivo general

El objetivo de este trabajo es hacer un estudio comparativo de las principales técnicas para el diseño y construcción de almacenes de datos. Con base a esta comparación el usuario pueda seleccionar la metodología que más se ajuste a sus necesidades.

Objetivos específicos

- Investigar diferentes metodologías, técnicas usadas para el diseño y construcción de almacenes de datos.
- Investigar las etapas del ciclo de vida de un almacén de datos.
- Comparar las metodologías, técnicas de diseño y construcción de almacenes de datos.
- Analizar un mismo ejemplo para cada metodología.
- Este documento sirva como guía de referencia de diseño y construcción de almacenes de datos.
- Precisar la viabilidad de cada metodología de un almacén de datos.

Relevancia y contribución

- Los resultados de esta investigación proporcionarán un marco de comparación de distintas metodologías y técnicas para el diseño y construcción de almacenes de datos.
- Integrar en un solo documento las diferentes técnicas de diseño.
- Es un tema novedoso lo que permitirá introducirse en aspectos básicos en el diseño de almacenes de datos.
- Proporcionará un mecanismo para adoptar y manejar los conceptos básicos bajo el enfoque de almacenes de datos.

- Precisar la viabilidad de cada metodología de un almacén de datos de acuerdo a las necesidades del usuario.

Organización de la tesis

La tesis está conformada por 5 capítulos.

Capítulo 1. Conceptos generales de almacenes de datos.

En este capítulo se describen los conceptos básicos de almacenes de datos, que se utilizarán en capítulos posteriores.

Capítulo 2. Metodología de William H. Inmon.

En este capítulo se describe la metodología de William H. Inmon y el diseño del almacén de datos.

Capítulo 3. Metodología de Ralph Kimball.

En este capítulo se describe la metodología de Ralph Kimball o también conocida como ciclo de vida de Kimball y el diseño del almacén de datos.

Capítulo 4. HEFESTO: Metodología para la construcción de un Almacén de Datos.

En este capítulo se describe la metodología de Bernabeu Ricardo Dario (HEFESTO) y el diseño del almacén de datos.

Capítulo 5. Estudio comparativo de técnicas para el diseño y construcción de almacenes de datos.

En este capítulo se presenta el resultado del estudio comparativo de las diferentes técnicas para el diseño y construcción de almacenes de datos, a partir de las metodologías antes analizadas.

“Los científicos se esfuerzan
por hacer posible lo imposible.
Los políticos por hacer lo posible imposible.”

- Bertrand Russell (1872-1970); Filósofo, matemático y escritor inglés.

Capítulo 1

Conceptos generales de almacenes de datos

En este capítulo se presentan conceptos y definiciones generales sobre almacenes de datos, que son utilizados en capítulos posteriores.

Los temas desarrollados en este capítulo permiten establecer un marco teórico acerca de cómo diseñar y construir un almacén de datos.

1.1. Necesidad de construir un almacén datos

Una base de datos operacional es manejada por sistemas operacionales o transaccionales, que es muy diferente a la base de datos que soporta un sistema de toma de decisiones. Los sistemas operacionales son sistemas diseñados para controlar el flujo de información que se genera diariamente en una organización.

La toma de decisiones no es únicamente para personas con funciones directivas, ejecutivas, sino también para cualquier persona que se encargue de tomar decisiones en una organización. La mayor parte de la información útil para la toma de decisiones en una organización, se concentran en los sistemas transaccionales, pero cuyo modelo no es el adecuado para proporcionar una visión dimensional requerida; es necesario complementar esa información con otras fuentes de datos, es por ello que es necesario el desarrollo de una tecnología de almacén de datos, esto con el propósito de estructurar los sistemas para la toma de decisiones.

Los usuarios de un almacén de datos, utilizan la información contenida en él análisis de manera oportuna en el comportamiento de una organización. Ya que sin un almacén de datos, la posibilidad de analizar de manera oportuna la información se disminuye por el tiempo que se invierte en extraer, cargar y transformar los datos para después analizarla. El diseño y construcción de un almacén de datos se justifica en poner los datos a disposición de los analistas y listo para su explotación y análisis de la información, para su posterior toma de decisiones.

El objetivo principal de un almacén de datos es concentrar la información, para descubrir tendencias, llevar a tomar mejores decisiones a la organización, etc. El almacén de datos intenta responder a la compleja necesidad de obtener información útil sin sacrificar el rendimiento de las aplicaciones operacionales, debido a lo cual ha tenido un impacto significativo en la administración de la información.

Un almacén de datos se crea cuando se tiene la necesidad de concentrar e integrar la información proveniente de diferentes fuentes de datos y conservar información histórica. Esto con el objetivo de que el acceso a la información sea eficiente y útil en la toma de decisiones. Ya que los problemas con la entrega de la información actual son muchos, incluyendo inconsistencia, inflexibilidad y carencia de integración a través de la organización.

En los últimos años los almacenes de datos en conjunto con la tecnología OLAP (OnLine Analytical Processing), paquetes estadísticos profesionales, son elementos principales en las organizaciones para la toma de decisiones.

La tecnología OLAP son bases de datos orientadas al procesamiento analítico, que facilitan el análisis de datos en línea en un almacén de datos, proporcionando respuestas a consultas para su análisis. Por lo tanto la tecnología OLAP es utilizado para la toma de decisiones presentando los datos a través de modelos de datos que sean fáciles de entender para los usuarios finales. Por otro lado con las fuentes de datos que se utilizan diariamente, muchas empresas hacen el análisis para la toma de decisiones sobre bases de datos transaccionales, conocidas como OLTP (OnLine Transaction Processing); son bases de datos orientadas al procesamiento de transacciones, que pueden involucrar operaciones de inserción, modificación y borrado de datos, que no son fáciles de entender para cualquier tipo de usuario. Es común encontrar que los sistemas transaccionales son accesados por cientos de usuarios simultáneamente, mientras que en un almacén de datos sólo por decenas. Otro factor es que frecuentemente los sistemas transaccionales son de menor tamaño a los de un almacén de datos, esto debido a que los almacenes de datos pueden estar conformados de varios OLTP.

Qué es un almacén de datos o Data Warehouse (DW)

Un almacén de datos es una colección de datos provenientes de diferentes fuentes de datos, que sirven para el análisis de la información y posteriormente la toma de decisiones.

La definición clásica de almacén de datos de W. H. Inmon, es la siguiente:

"Un almacén de datos es una colección de datos orientados a temas, integrada, variante en el tiempo y no volátil, diseñado para dar apoyo al proceso de toma de decisiones en una organización" [1].

Otras definiciones importantes de almacenes de datos son:

Ralph Kimball: *"Un almacén de datos es un sistema que extrae, se limpia, se ajusta y entrega datos de origen en un almacén de datos dimensional y luego apoya, aplica consultas y análisis para el propósito de toma de decisiones"* [2].

Barry Devlin: *"Es un simple, completo y consistente almacenamiento de datos obtenidos de una variedad de fuentes y hecho accesible para usuarios finales en una manera que puedan entender y usar en el contexto de sus negocios"* [3].

Ventajas y desventajas de un almacén de datos

En la tabla 1.1 se muestran las ventajas y desventajas de un almacén de datos.

Almacén de datos	
Ventajas	**Desventajas**
• Proporciona información clave para la toma de decisiones empresariales. • Mejora la calidad de las decisiones tomadas. • Proporciona una comunicación fiable entre todos los departamentos de la empresa. • Permite conocer qué está pasando en el negocio, es decir, estar siempre enterado de los buenos y malos resultados. • Mejora la entrega de información, es decir, información completa, correcta, consistente, oportuna y accesible. • Útil para el almacenamiento de análisis y consultas de históricos. • Elimina la producción y el procesamiento de datos que no son utilizados ni necesarios, producto de aplicaciones mal diseñadas o ya no utilizadas.	• Subestimación de los recursos necesarios para la carga de datos. • No es útil para la toma de decisiones en tiempo real, debido al largo tiempo de procesamiento que puede requerir. • Requiere de continua limpieza, transformación e integración de los datos. • Una vez implementado puede ser complicado añadir nuevas fuentes de datos. • Tiene un diseño complejo. • Incremento de la demanda por parte de los usuarios finales. • Alta demanda de recursos. • Altos costos de implementación y mantenimiento. • Proyecto de larga duración. • Puede quedar obsoleto en cualquier momento.

Tabla 1.1. Ventajas y desventajas de un almacén de datos.

1.2. Características de un almacén de datos

Partiendo de la definición de W. H. Inmon que se presentó anteriormente, contempla 4 características importantes:

- Orientado a temas.
- Integrada.
- Variante en el tiempo.
- No volátil.

Orientado a temas

Orientado a temas significa que un almacén de datos está enfocado a los datos relacionados con un área de actividad del negocio y no por aplicación.

En el ambiente de almacén de datos se organiza alrededor de temas tales como cliente, vendedor, producto y actividad. Por ejemplo, para un fabricante, pueden ser clientes, productos, clasificación, ubicación geografía, etc. Mientras que en el proceso orientado a aplicaciones, se incluyen los datos que son necesarios para satisfacer de manera inmediata los requerimientos funcionales de la actividad. Por ejemplo, los datos comunes de los clientes, como su RFC, dirección, correo electrónico, fax, teléfono, etc. Que son importantes de almacenar en un sistema operacional, pero que no son tomados en cuenta en un almacén de datos por falta de valor para la toma de decisiones.

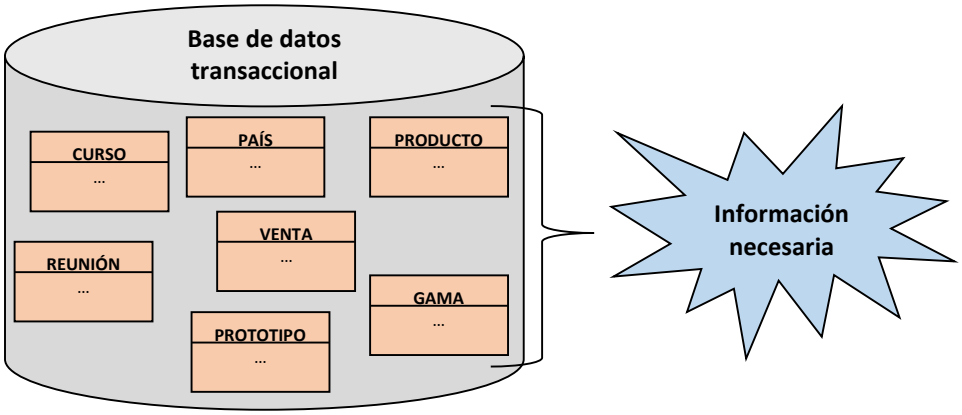

Figura 1.1. Orientado a temas.

Integrada

Integrada significa que los datos independientemente de las bases de datos que provengan son almacenados en un único repositorio de datos, unificando su formato (integración de formato) y estandarizando su significado (integración semántica). La integración de los datos en muchas empresas es un problema en particular cuándo existen muchas tecnologías para su uso. Para ello existen procesos de integración para el proceso de limpieza y estandarización de los datos. A continuación algunos ejemplos:

- **Codificación:** los diseñadores de las aplicaciones representan de varias formas el término género, algunos los ponen como “M” y “F”, otros con “1” y “0”, o como “X” y “Y”.

  No importa como quede representado el género en el almacén de datos, lo importante es que esté en un estado íntegro y uniforme.

- **Medida de atributos:** los diseñadores de aplicaciones utilizan diferentes sistemas de medidas. Por ejemplo: centímetros, pulgadas, metros, pies, etc. Cualquiera que sea la fuente de donde provenga la información debe llegar al almacén de datos de la misma manera para todos.

- **Convenciones de nombramiento:** cuando algún elemento es llamado por diferentes nombres, la transformación asegura que se estandarice.

- **Fuentes múltiples:** un mismo elemento puede provenir de muchos sistemas fuentes, lo que se desea es tomar la información de la fuente más apropiada.

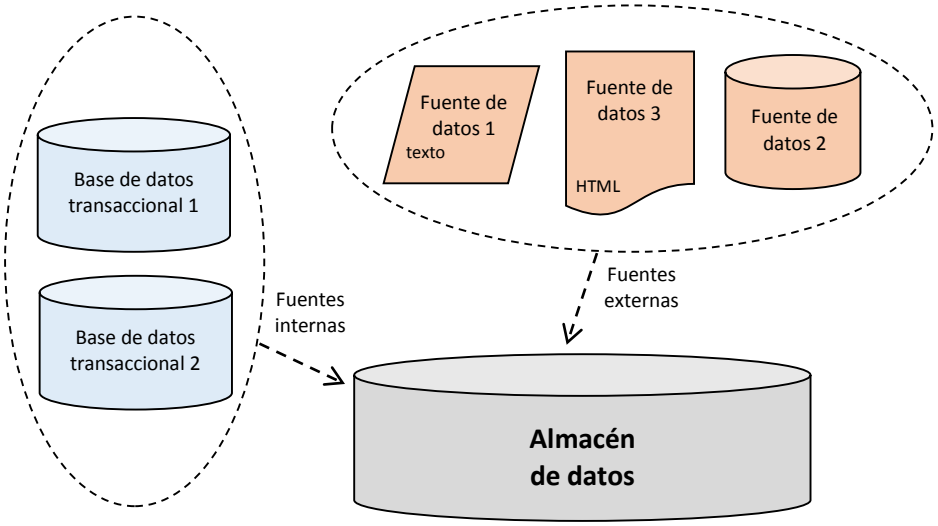

Figura 1.2. Integrada.

Variante en el tiempo

No hay que confundir cuando dice variante en el tiempo, la información no se actualiza. En almacén de datos significa que los datos están asociados a un punto del tiempo. Por ejemplo: diario, mensual, trimestral, semestral o por año. El almacén contiene gran cantidad de información histórica que va recibiendo periódicamente nuevos datos.

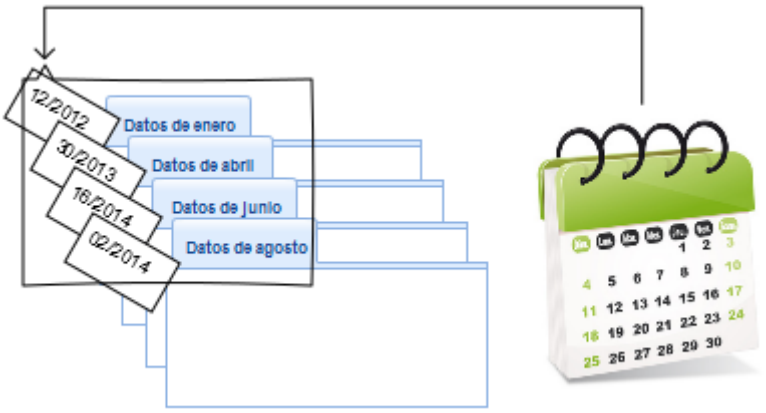

Figura 1.3. Variante en el tiempo.

No volátil

No volátil significa que los datos no cambian (no se actualizan) cuando son depositados al almacén de datos. Cualquiera que necesite extraer información del almacén tiene la seguridad que siempre arrojará el mismo resultado.

La información es útil en un almacén de datos cuando es estable, ya que en una base de datos operacional los datos cambian constantemente.

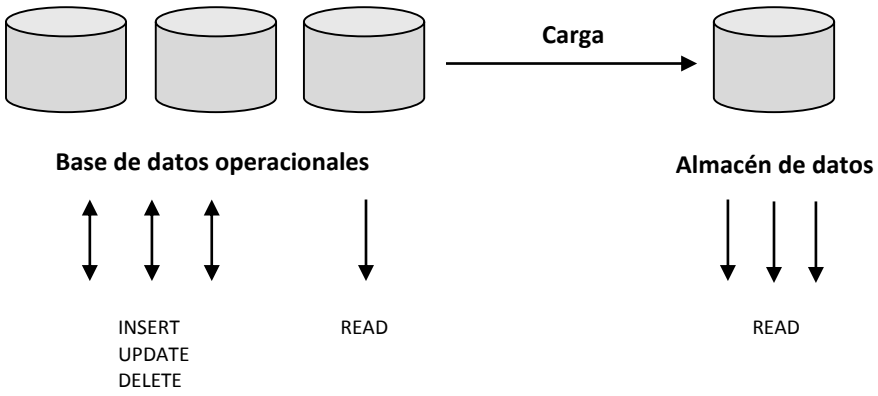

Figura 1.4. No volátil.

1.3. Base de datos operacional

Los sistemas operacionales son importantes ya que ofrecen diferentes tipos de procesos de datos como: simplicidad y generalidad, facilidad de uso para el usuario final, consulta de información de manera sencilla.

Por lo tanto una base de datos operacional es una base de datos que se alimenta de operaciones o transacciones. En estas bases se conservan todos los datos al mayor detalle que respaldan las operaciones de toda la organización.

Las tablas son la forma de representar la información de manera compacta y posible de acceder a la información contenida en dos o más tablas.

Las bases de datos operacionales están constituidas por dos o más tablas que contiene información ordenada de una forma organizada que cumplen con los siguientes puntos:

- Una tabla sólo contiene un número fijo de campos.
- El nombre de los campos de una tabla es distinto.
- Cada registro de la tabla es único.
- El orden de los registros y de los campos no está determinado.
- Para cada campo existe un conjunto de valores posibles.

Por ejemplo, se utilizan para el funcionamiento de los negocios en tiempo real que son llamados "sistemas operacionales" o "sistemas de producción". Manipulan gran número de transacciones simples de lectura/escritura y se basan en datos operacionales o datos actuales del estado de la empresa. Además son importantes ya que juegan un papel fundamental para cualquier organización, pues garantizan la automatización de los procesos y el flujo de la información a través de la misma.

Diferencias entre una base de datos operacional y un almacén de datos

Los sistemas operacionales tradicionales y las aplicaciones de un almacén de datos son opuestos en cuanto a sus requerimientos de diseño y sus características de operación.

En una aplicación la información en la base de datos operacional, se tiene organizada de tal manera que se puede extraer directamente, a diferencia de un almacén de datos donde se tiene una sola base de datos diseñada, estructurada y organizada por áreas temáticas. La ventaja de tener una base de datos por aplicación es el rápido acceso a los datos ya que los mismos están preestablecidos, mientras que en un almacén de datos para tener esta ventaja se necesita que la información esté desnormalizada para evitar que el motor de búsqueda tenga que recorrer toda la base de datos para encontrar lo que necesita.

En la tabla 1.2 se muestran las diferencias entre una base de datos operacional y un almacén de datos.

	Base de datos operacional	Almacén de datos
Propósito	Operaciones diarias. Soporte a las aplicaciones.	Recuperación de información, informes, análisis y minería de datos.
Actividad	Predomina la actualización.	Predomina la consulta.
Contenido de datos	Valores actuales.	Archivado, resumen.
Datos	No redundantes.	Redundantes.
Estructura de datos	Optimizado para las transacciones.	Optimizado para consultas complejas.
Modelo de datos	Datos normalizados.	Datos en estrella, en copo de nieve, parcialmente desnormalizados, multidimensionales.
Explotación	Explotación de la información relacionada con la operativa de cada aplicación.	Explotación de toda la información interna y externa relacionada con el negocio.
Frecuencia de acceso	Alta.	Moderada a escasa.

Granularidad	Datos generales desagregados, al detalle.	Datos en distintos niveles de detalle y agregación.
Horizonte histórico	30 a 90 días.	5 a 10 años.
Organización	Estructura normalmente relacional.	Visión multidimensional.
Rendimiento	Importancia del tiempo de respuesta de la transacción instantánea.	Importancia de la respuesta masiva.
Tipo de acceso	Leer, actualizar, eliminar.	Leer.
Uso	Predecible y repetitivo.	Ad hoc, al azar, heurística.
Tiempo de respuesta	Segundos.	Varios segundos a minutos.
Usuarios	Número grande (cientos/miles: aplicaciones, operarios, administrador de la base de datos).	Número relativamente pequeño (decenas: directores, ejecutivos, analistas).
Volatilidad	Actualizable.	Carga, pero no actualización.
Volumen de datos	Pequeño/Medio. Del orden de MB a GB.	Medio/Grande. Del orden de GB a TB.

Tabla 1.2. Diferencias entre una base de datos operacional y un almacén de datos.

1.4. OLTP

OLTP es la sigla en inglés de Procesamiento de Transacciones En Línea (OnLine Transaction Processing). Es un tipo de procesamiento que facilita y administra aplicaciones transaccionales a través de una red de computadoras. La función principal de los sistemas OLTP es el procesamiento de consultas muy rápidas (operaciones de tipo INSERT, UPDATE Y DELETE) manteniendo la integridad de los datos y es eficaz en términos de números de transacciones por segundos.

A continuación se explican aspectos importantes de un OLTP:

- **Alineación de los datos:** Los OLTP están alineados por aplicación. Diferentes sistemas tienen distintos tipos de datos, los cuales son estructurados por aplicación. Se enfoca en el cumplimiento de los requerimientos de una aplicación en especial o una tarea específica.

- **Integración de datos:** Los datos por lo general no están integrados, son calificados como datos operacionales, que son estructurados independientemente de unos de otros, pudiendo tener diferentes estructuras de claves y convenciones de nombres. Son usualmente almacenados en diferentes formatos de archivos.

- **Historia:** Los OLTP retienen datos de entre 60 a 90 días aproximadamente, después son resguardados por los administradores de base de datos en almacenamientos secundarios. Por ejemplo: cintas o en disco a nivel de back up[1].

- **Perfil de usuario:** Dado que los OLTP tienen como objetivo asistir a aplicaciones específicas y asegurar la integridad de los datos, el perfil de usuario que interactúa con dichos sistemas se encuentra dentro de los empleados operacionales de una organización.

[1] Copia de seguridad o su nombre en Inglés Back up es una copia de seguridad de los datos originales que se realiza con el fin de disponer de un medio de almacenamiento para recuperarlos en caso de pérdida.

1.5. OLAP

OLAP es la sigla en inglés de Procesamiento Analítico En Línea (OnLine Analytical Processing). Es una tecnología que se basa en el análisis multidimensional (o cubos OLAP) que contienen datos resumidos de grandes bases de datos, permitiendo a los usuarios tener una visión más rápida e interactiva de los mismos. La arquitectura del almacén de datos está diseñada para OLAP de consultas de usuarios [6].

La razón de usar OLAP para las consultas es la velocidad de respuesta. Una base de datos relacional almacena entidades en tablas. La estructura es buena en un sistema OLTP pero para las consultas complejas multitabla es muy lenta. Un modelo mejor para búsquedas, aunque peor desde el punto de vista operativo, es una base de datos multidimensional. La principal característica que potencia a OLAP, es lo más rápido a la hora de ejecutar sentencias SQL de tipo SELECT.

Beneficios OLAP

- Aumento de la productividad de los administradores de negocios, ejecutivos y analistas.
- Flexibilidad inherente de los sistemas OLAP significa que los usuarios pueden ser autosuficiente ejecutando sus propios análisis sin asistencia de TI[2].
- Beneficios para los desarrolladores de TI porque están usando software diseñado específicamente para el sistema.
- La autosuficiencia para los usuarios, lo que resulta en la reducción de la cartera.
- Acelerar la entrega de aplicaciones.
- Operaciones más eficientes a través de la reducción del tiempo de ejecución de consulta y tráfico en la red.
- Capacidad para modelar desafíos del mundo real con métricas de negocio y dimensiones.

Modelos OLAP

Una explicación muy simple de las variaciones se refiere a la forma de los datos se almacena para OLAP. El proceso sigue siendo el procesamiento analítico en línea, básicamente, la metodología de almacenamiento es diferente.

Los sistemas OLAP se clasifican en las siguientes categorías:

- **ROLAP:** se refiere al procesamiento analítico en línea relacional. En este caso, el sistema OLAP está construido sobre una base de datos relacional.

- **MOLAP:** se refiere al procesamiento analítico en línea multidimensional. En este caso, el sistema OLAP se implementa a través de una base de datos especializada multidimensional.

- **HOLAP:** se refiere al procesamiento analítico en línea híbrida. Este modelo trata de combinar las fortalezas y características de ROLAP y MOLAP.

[2] Tecnología de la información.

Herramientas y productos OLAP

Muchas herramientas y productos OLAP han aparecido y la mayoría de ellos con bastante éxito. La calidad y la flexibilidad de los productos han mejorado notablemente.

Recomendaciones para elegir herramientas y productos OLAP:

- Dejar que los usuarios manejen la selección de los productos OLAP. No dejarse llevar por la tecnología llamativa.
- El sistema OLAP crecerá en tamaño y en el número de usuarios activos. Determinar la escalabilidad de los productos antes de elegir.
- Considerar la facilidad de administrar el producto OLAP.
- Rendimiento y flexibilidad son elementos clave en el éxito del sistema OLAP.
- A medida que la tecnología avanza, las diferencias en cuanto al fondo entre ROLAP y MOLAP parecen estar un tanto borrosas. No preocuparse demasiado acerca de estos dos métodos.

1.6. EIS

EIS es la sigla en inglés de Sistema de Información Ejecutiva (Executive Information System). Es una herramienta de Inteligencia de negocios (Business Intelligence, BI), orientada a usuarios de nivel gerencial, que permite automatizar la labor de obtener los datos más importantes de una organización, resumirlos y presentarlos de la forma más comprensible posible, provee al ejecutivo acceso fácil a información interna y externa al negocio con el fin de dar seguimiento a los factores críticos del éxito.

1.7. Componentes y estructuras de un almacén de datos

La arquitectura de un almacén de datos es una forma de representar la estructura global de los datos, la comunicación, los procesos y la presentación del usuario final. En la figura 1.5 se muestra la arquitectura de un almacén de datos.

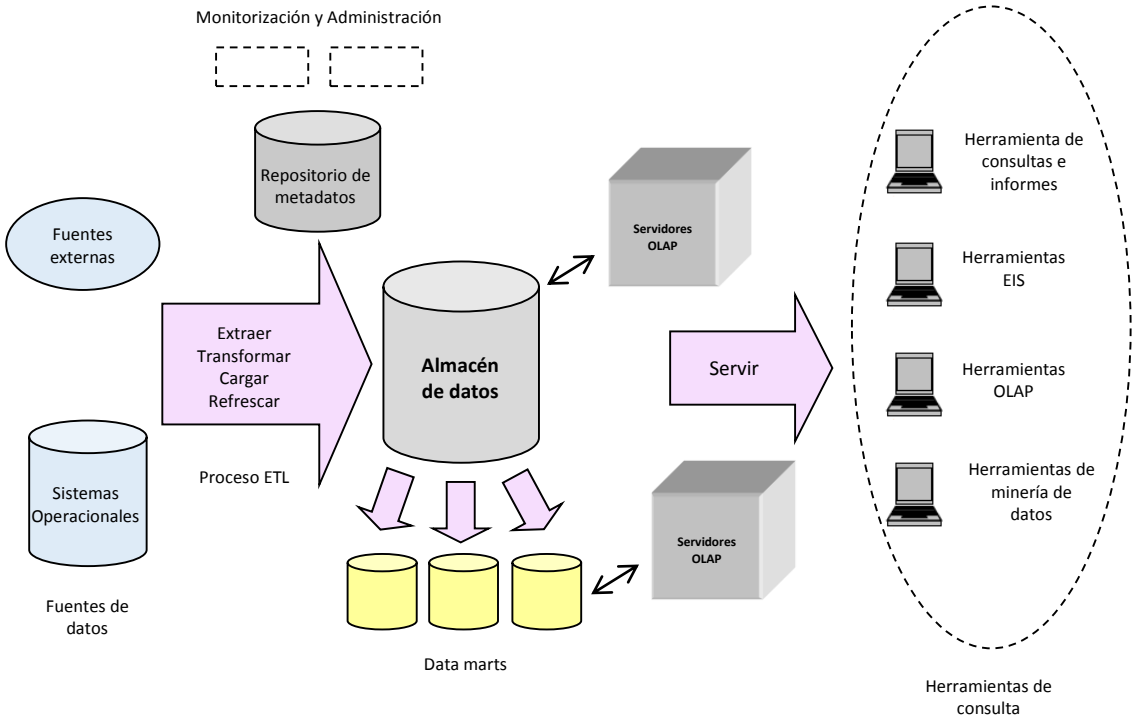

Figura 1.5. Componentes de un almacén de datos. Propuesta por Trujillo, J. C., Mazón, J. N., & Pardillo, J. [41].

A continuación se describe cada componente de un almacén de datos:

1.7.1. Fuente de datos

Una fuente de datos es un componente fundamental en una organización, a partir del cual se realiza la captura de los datos que se contemplará en el almacén de datos. Las fuentes de datos pueden ser bases de datos operacionales, datos históricos (generalmente archivados en cintas), archivos propietarios, estaciones primarias de trabajo e incluso en sistemas externos como: internet, bases de datos comerciales o bases de datos dependientes de los proveedores o clientes de la organización.

1.7.2. Proceso ETL

ETL son las siglas en inglés de Extraer, Transformar y Cargar (Extract, Transform and Load). Es el proceso responsable de la extracción de los datos de varias fuentes, limpieza, personalización e inserción en un almacén de datos [8]. Sin embargo, no es inusual identificar valores nulos en una tabla de hechos de un almacén durante el proceso ETL y esto puede tener un impacto negativamente en la precisión de los resultados del análisis de datos [11].

Una pregunta interesante que hay que tener en cuenta es ¿Por qué no extraer los datos operativos y cargarlos al almacén de datos directamente? Esto a simple vista parece sencillo y hasta lógico,

pero esto no se debe hacer, sino que a partir de los requerimientos que se han planteado desde el inicio del proyecto deben guiarse qué datos se necesitarán extraer y de qué sistemas de origen.

Cada función del proceso ETL cumple un propósito específico, para cambiar los datos en información útil primeramente se deben capturar los datos, después de capturarlos no se deben cargar los datos directamente al almacén de datos, sino que tienen que someterse a todo tipo de transformaciones de manera que se ajusten para ser convertidos en información estratégica y finalmente trasladarlos al almacén.

El proceso ETL es extremadamente complejo, propenso a errores y consume mucho tiempo [9,10].

Algunos desafíos del proceso ETL son:

- Diversos sistemas de origen y dispares.
- La calidad de los datos es dudosa en muchos de los sistemas de origen que han evolucionado con el tiempo.
- Falta de coherencia entre los sistemas de origen.
- A veces el tipo y/o formato de los datos en los sistemas de origen no tienen significado directo bajo los términos del usuario.

Por otro lado, en [9,10] se menciona que el proceso ETL representa el 80% de los recursos de desarrollo de un proyecto de almacén de datos.

Extracción: Extracción es el proceso de obtener los datos de diversas fuentes de origen. Debe realizarse una selección de los campos y registros de los sistemas operacionales, ya que no todos los datos son importantes para un almacén de datos.

Por lo regular los formatos de las fuentes de origen se encuentran en bases de datos relacionales o archivos de texto plano, pero cuando los datos son extraídos de estructuras diferentes, el proceso de extracción convierte a los datos en un formato preparado para iniciar el proceso de transformación.

Aspectos a considerar en el proceso de extracción son los siguientes:

- Identificación de las fuentes de datos.
- Método de extracción para cada fuente de datos.
- Frecuencia de extracción de los datos.
- Espacio de tiempo entre cada extracción de datos de las diferentes fuentes.
- Calendarización de tareas.
- Control de excepciones - Determina cómo manejar los registros de entrada que no se pueden extraer.

Transformación: Transformación es el proceso en donde los datos se preparan de manera adecuada para ser cargados al almacén de datos. El proceso está compuesto por las siguientes actividades:

- **Limpieza de datos:** La limpieza de los datos se refiere a una actividad que determina y detecta datos inconsistentes, valores perdidos y errores; que ayuda a corregir y mejorar la calidad de los datos provenientes de distintas fuentes de origen.

  La presencia de datos sucios en un almacén de datos provoca problemas graves, incluso desconfianza por parte de la organización, ya que los datos se utilizan con el objetivo de

tomar decisiones, planeación estratégica, análisis de las tendencias. Por lo tanto la limpieza de los datos es la única solución que se puede aplicar para evitar los problemas antes mencionados.

Podemos contemplar dos tipos de limpieza:

- **Limpieza moderada:** Identificar que dos palabras significan lo mismo o representan la misma información. Por ejemplo: Dir. y Dirección o eliminar otros signos de puntuación: puntos, comillas, etc.

- **Limpieza intensa:** Consiste en que el usuario proponga reglas para realizar la limpieza de los datos. Por ejemplo: Juan tiene dos casas en diferentes direcciones ¿Debe considerarse cómo el mismo cliente? Está información determina cómo se almacenará la información.

- **Integración de formatos:** Frecuentemente pasa que muchas organizaciones utilizan diferentes formatos. Por ejemplo: La fecha para algunos puede ser almacenada como "ddmmyyyy", "yyyymmdd", o algunos sistemas que almacenan datos de tipo dinero son un problema ya que algunos guardan valores enteros y el sistema agrega el punto decimal, mientras que en otros sistemas el usuario tiene que poner el punto decimal. Por estas razones se busca estandarizar los tipos de formatos.

- **Integración semántica:** Hace referencia al significado de los datos, ya que para algunas personas de la organización los datos pueden significar diferente que para otros miembros. Es por eso que los datos que se inserten al almacén de datos tengan significados precisos y que sean conocidos por todos los usuarios de la organización.

- **Integración de datos:** Es común que datos que se van a insertar en un almacén de datos se construyan a partir de otras fuentes de datos. Este proceso sigue una serie de reglas para garantizar que el dato que se va a cargar al almacén sea correcto.

Carga: Cuando la información ya ha sido extraída de las diferentes fuentes de datos y transformada, los datos ya pueden ser cargados al almacén de datos. Por lo tanto cuando se han cargados los datos, lo siguiente es actualizar el almacén periódicamente.

Tipos de aplicaciones de datos en el almacén de datos:

- Carga inicial: Poblar todas las tablas del almacén de datos por primera vez.
- Carga incremental: Aplicar actualizaciones periódicas como sean necesarias.
- Refrescado completo: Borrar completamente una tabla y recargarla con datos frescos.

Durante las cargas de datos el almacén de datos no está disponible para los usuarios. Es difícil estimar los tiempos de carga.

Se pueden crear programas para realizar la carga. Lo cual implica esfuerzo extra para su creación y el mantenimiento. Pero se recomienda en lo posible el uso de las utilidades que vienen con los SGBD[3].

[3] Sistema de Gestión de Bases de Datos.

1.7.3. Repositorios de metadatos

1.7.3.1. Metadatos

Los metadatos son datos que describen información, contenido, calidad de los datos almacenados en una base de datos que incluyen información cómo:

1. Descripción de tablas y campos del almacén de datos (tipos de datos y rangos de valores).
2. Descripción de tablas y campos de las bases de datos fuentes.
3. Descripción de los datos de cómo han sido transformados (formato, formulas, conversión de valores).

Los metadatos se clasifican en tres categorías [5]:

- Metadatos operacionales.
- Metadatos para extracción y transformación.
- Metadatos para información de usuario final.

Metadatos operacionales: Los datos provienen de varios sistemas operacionales de la empresa, esto implica distintas estructuras de datos, diferentes tipos de datos y longitudes. Algunos de los elementos seleccionados para el almacén de datos deben ser separados, concatenados, codificados, etc.

Los metadatos operacionales contienen toda la información para recuperar los conjuntos de datos fuentes.

Metadatos para extracción y transformación: Estos metadatos contienen la información acerca de la extracción de datos de los sistemas fuentes: frecuencias, métodos y reglas de negocio. También contienen información acerca de la transformación de los datos en la etapa de preparación antes del almacenamiento de los datos.

***Metadatos para información de usuario final*:** Es el mapa de navegación del almacén de datos. Permite al usuario final usar sus propias terminologías del negocio y buscar información en la forma en la que el usuario normalmente piensa en el negocio.

Los metadatos son importantes porque actúan como la unión de todas las partes del almacén de datos, provee información a los desarrolladores acerca de la estructura y el contenido del almacén y finalmente abren la puerta hacia el usuario final y hace que el contenido sea reconocible en sus propios términos.

Los metadatos son necesarios para:

- Usar un almacén de datos: Hay una gran diferencia entre un almacén de datos y cualquier sistema operacional tal como una aplicación de procesamiento de pedidos. La diferencia es el uso y el acceso a la información.

- Construir el almacén de datos: para implementar los componentes de extracción y trasformación de los datos del almacén de datos, el desarrollador necesita metadatos acerca de los sistemas fuentes. Además que el diseño físico de la base de datos del almacén necesitan metadatos del diseño lógico.

- Administrar el almacén de datos: debido al gran tamaño y complejidad, actualmente es absolutamente imposible administrar el almacén de datos sin metadatos fundamentales.

1.7.4. Data mart

Un data mart es una parte del almacén de datos adaptado para un área específica de la empresa u organización. Aun entendiéndose que puede ser desarrollado independientemente del almacén de datos, debe considerarse la necesidad de disponer de los datos de forma totalmente integrada para poder explotar de forma rápida, eficiente y segura con la máxima capacidad de información potencial.

Características importantes de un data mart:

- Usuarios limitados.
- Área específica.
- Tiene un propósito específico
- Tiene una función de apoyo.

Hay que destacar que un data mart se define por el alcance funcional de sus usuarios y no por el tamaño de la base de datos del data mart.

Los data marts surgen por la complejidad y costos elevados asociados a la implementación de un proyecto de almacén de datos.

Diferencias entre un data mart y un almacén de datos

En la tabla 1.3 se muestran las diferencias entre un data mart y un almacén de datos.

Data mart	Almacén de datos
Departamental.	Corporativo.
Alto nivel de granularidad.	Menor nivel de granularidad.
Estructura Star-join[4].	Estructura normalizada.
Menor cantidad de datos históricos.	Robusta cantidad de datos históricos.
Tecnología óptima para accesos y análisis.	Tecnología óptima para mantenimiento, gestión de grandes volúmenes de datos.
Cada departamento tiene una estructura diferente.	Estructura adaptada a la comprensión de los datos corporativos.
Altamente indexada.	Ligeramente indexado.

Tabla 1.3. Diferencias entre un data mart y un almacén de datos.

1.7.5. Servidores OLAP

El servidor OLAP es el encargado de recibir las solicitudes por información que provienen de la capa de presentación, entonces calcula todas las operaciones necesarias con cubos OLAP y envía de regreso a la interfaz. Un servidor OLAP debe ofrecer un lenguaje para consulta y manipulación de cubos, así como la función de traducción entre ese lenguaje y el lenguaje del administrador de datos que se encuentra en la capa de almacenamiento y servicios.

[4] Star-join considera que un conjunto de uniones forman una unión en estrella cuando una tabla de hechos se unen a dos o más tablas de dimensiones.

1.7.6. Herramientas de consulta

Sin las herramientas adecuadas de acceso y análisis un almacén de datos puede no tener utilidad para el usuario final. Es necesario poseer técnicas que capturen los datos importantes de manera rápida y puedan ser analizados desde diferentes puntos de vista. También deben transformar los datos capturados en información útil para el negocio. Actualmente existen herramientas que se les conoce como herramientas OLAP, herramientas de inteligencia de negocios y están situadas conceptualmente sobre un almacén de datos.

Las herramientas OLAP presentan al usuario una visión multidimensional de los datos para cada actividad que es objeto de análisis. Los usuarios pueden formular sus consultas a las herramientas OLAP, seleccionando los atributos del modelo multidimensional sin conocer la estructura interna del almacén de datos. Por ejemplo: las herramientas OLAP generan las consultas correspondientes y las envía al gestor de consultas del sistema (mediante una sentencia SELECT).

Cada usuario debe seleccionar la herramienta que mejor se ajuste a sus necesidades y al almacén de datos que está utilizando. Se incluye tanto hardware como software involucrados en mostrar la información en pantalla, generar reportes, hojas de cálculo, gráficos, diagramas para el análisis y presentación.

1.8. Diseño del almacén de datos

1.8.1. Técnicas de diseño

En cuanto al desarrollo, a la hora de construir un almacén de datos no hay una sola metodología en la que se base el diseño, sino dependiendo de las necesidades en las que se encuentre la empresa y objetivos que se persigan, se puede emplear uno u otro enfoque; “De arriba abajo” o “De abajo arriba” que corresponden con las metodologías propuestas por Inmon y Kimball respectivamente y la metodología propuesta por Bernabeu (HEFESTO), que se entrará en detalle en capítulos posteriores.

1.8.2. Consideraciones de diseño

1.8.2.1. Tamaños

Los tamaños de los almacenes de datos pueden variar dependiendo de la cantidad de datos con que se cuente. Por lo tanto podemos clasificarlo conforme a su tamaño del depósito de datos [7]:

- Personal: si su tamaño es menor a 1 Gigabyte.
  DW < 1 GB
- Pequeño: si su tamaño es mayor a 1 Gigabyte y menor a 50 Gigabyte.
  1 GB < DW < 50 GB
- Mediano: si su tamaño es mayor a 50 Gigabyte y menor a 100 Gigabyte.
  50 GB < DW < 100 GB
- Grande: si su tamaño es mayor a 100 Gigabyte y menor a 1 Terabyte.
  100 GB < DW < 1 TB
- Muy grande: si su tamaño es mayor a 1 Terabyte.
  DW > 1 TB

1.8.2.2. Tiempo de construcción

No existe como tal un tiempo preciso de construcción de un almacén de datos, pero se han considerado tres fases al respecto:

- "El 70% del tiempo total dedicado al proyecto es para definir el problema y en preparar los datos".
- "Se estima el tiempo aproximado de construcción de un almacén de datos, ese tiempo aproximado se multiplica por dos y se agrega una semana de holgura".
- "Regla 90 - 90": El primer 90% de la construcción de un sistema absorbe el 90% del esfuerzo y tiempo invertido. El último 10% se lleva el otro 90% del esfuerzo y del tiempo asignado.

1.8.2.3. Rendimiento

Es importante la configuración del SGBD en la que se almacenará y mantendrá el almacén de datos. Para el mejor desempeño del almacén, lo recomendable es llevar a cabo las siguientes acciones sobre el almacén y la estructura de datos:

- Prestar atención a los tipos de datos utilizados, es decir asignar los tamaños adecuados a los tipos de datos.
- Utilizar llaves sustitutas (llaves subrogadas).
- Utilizar técnicas de partición.
- Crear diferentes niveles de resumen (sumarización).
- Crear vistas materializadas.
- Utilizar técnicas de administración de datos en memoria caché.
- Utilizar técnicas de multiprocesamiento.

1.8.2.4. Mantenimiento

Dar constante mantenimiento al correcto funcionamiento del almacén de datos es sumamente importante, ya que a medida que transcurra el tiempo el almacén crecerá considerablemente y surgirán nuevos cambios tanto en los requerimientos como en las fuentes de datos.

Para la construcción de un almacén de datos es muy importantes que los usuarios participen durante su desarrollo, ya que ellos son los que conocen el negocio y cuáles son sus necesidades.

Un almacén de datos integra fuentes de datos de diversas áreas, dándoles a los usuarios el beneficio de poseer toda esa información en un sólo depósito de datos, esto facilita a que en diferentes áreas compartan los mismos datos lo cual conduce un mayor entendimiento, confianza por parte de los usuarios.

1.8.2.5. Balance de diseño

En la figura 1.6 se muestran tres características importantes que se deben equilibrar al momento de diseñar y construir un almacén de datos.

Figura 1.6. Balance de diseño. Propuesta por Dario, B. [7].

Las tres características están relacionadas entre sí, por lo tanto el valor que adopte cualquiera una de ellas, afectará a las otras significativamente.

Por ejemplo, si se centra en la atención en los requerimientos de usuarios, se obtendrá un almacén de datos completo que cubra todas las necesidades de análisis. Pero el rendimiento de las consultas se verá disminuida y como consecuencia un aumento de mantenimiento de bases de datos.

1.8.2.6. Desnormalización

Un modelo de datos completamente normalizado, no contiene redundancia, reduce problemas de integridad y optimiza las actualizaciones, quizás con el costo del tiempo de recuperación, pero cuando se pretende evitar esta demora, resultado de la combinación de muchas tablas entonces se puede utilizar la desnormalización.

La desnormalización es el proceso de poner la misma información en varios lugares que procura optimizar el desempeño de la base de datos.

Dependiendo de la aplicación se debe tener en cuenta en qué casos desnormalizar, evaluando los siguientes factores:

- Tamaño de la aplicación.
- Número de acceso concurrente sobre dichas tablas.
- Velocidad de respuestas de las consultas.

La responsabilidad del diseñador es asegurarse que la base de datos desnormalizada no llegue ser inconsistente.

Hay que tener en cuenta que un modelo de datos desnormalizado no es lo mismo que un modelo de datos que no ha sido desnormalizado, por lo tanto la desnormalización debe tomar lugar solamente después de que se haya llevado a cabo un nivel de normalización satisfactorio y hayan sido creadas las restricciones requeridas.

1.8.2.7. Granularidad

El aspecto más importante del diseño de un almacén de datos es el tema de la granularidad [1]. La granularidad se refiere al nivel de detalle de las unidades de datos en un almacén de datos.

Entre mayor sea el nivel de detalle de los datos, se podrá hacer un mejor análisis, ya que estos podrán ser resumidos hasta obtener una granularidad media o gruesa. Pero no sucede lo mismo en sentido contrario.

Por ejemplo, si los datos almacenados con granularidad media podrán ser resumidos pero no podrán ser analizados a nivel de detalle. O si la granularidad con la que se guardan los datos a nivel de día, estos datos podrán ser resumidos por semanas, meses, semestres, años. Pero si estos registros se hubieran almacenados a nivel de mes podrían resumirse por semestres y años, pero no por días y semanas.

La granularidad es un problema si no es diseñado correctamente, ya que esto afecta al volumen de los datos en el almacén de datos y el tipo de consultas que pueden ser contestadas.

Beneficios de la granularidad:

1. Capacidad de ajustar los datos si es necesario.
2. Flexibilidad.
3. Contiene historial de actividades y eventos a través de la corporación.
4. Suficientemente detallada.
5. Nuevos requerimientos pueden ser acomodados.

Con muy bajo nivel de granularidad se puede responder prácticamente cualquier tipo de consulta, sin embargo un alto nivel de granularidad limita el número de preguntas que los datos pueden manejar. Por lo tanto debe existir un equilibrio entre responder a cualquier tipo de pregunta y el tiempo en responderla.

Ejemplo:

Registro de llamadas realizadas por un cliente en un lapso de un tiempo determinado.

En la figura 1.7 se muestra el ejemplo de los problemas de granularidad. El lado izquierdo muestra un bajo nivel de granularidad. Cada actividad para este caso una llamada telefónica se registró en detalle. Al final del mes, cada cliente tiene en promedio 200 registros de llamadas que requieren más espacio en disco.

El lado derecho de la figura 1.7 se muestra un alto nivel de granularidad. Es decir representa la información en resumen. Cada registro resume la actividad de un mes para un cliente, que requiere menos espacio en disco.

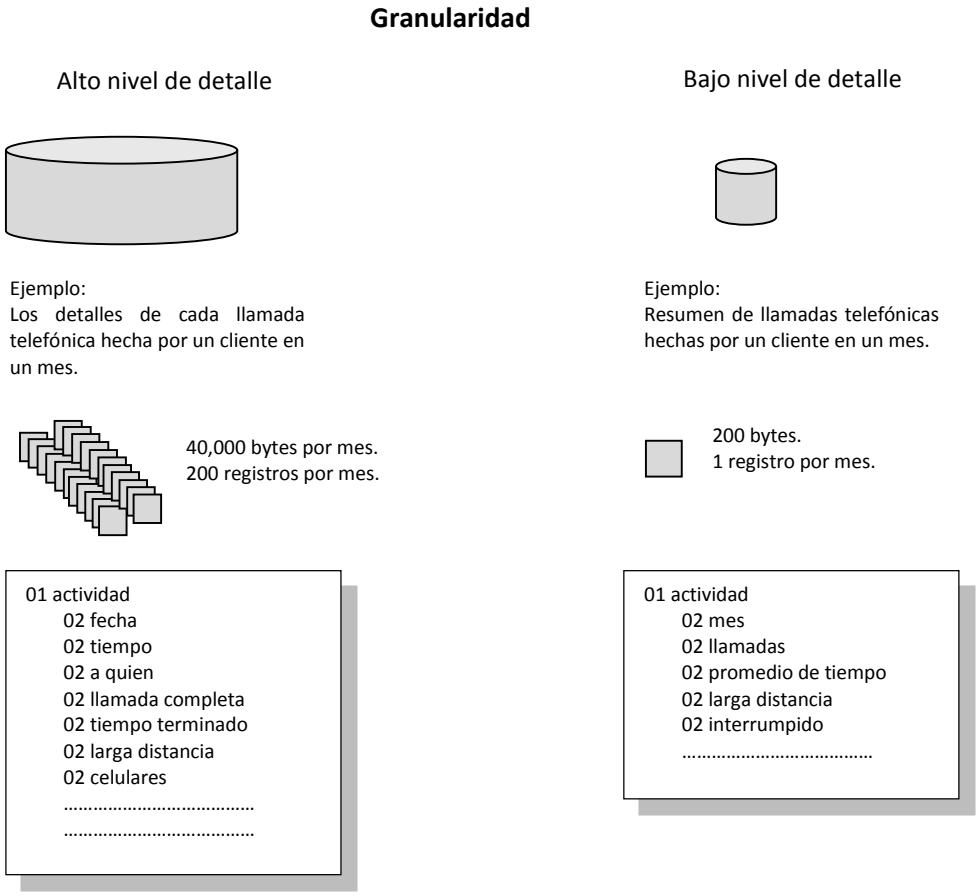

Figura 1.7. Determinar el nivel de granularidad es el problema de diseño más importante en un almacén de datos. Propuesta por H Inmon, W. [1].

1.8.2.8. Particionamiento de los datos

En un almacén de datos el particionamiento se utiliza para dividir una tabla de hechos en varias tablas más pequeñas, a través de ciertos criterios preestablecidos.

Razones principales del por qué particionar los datos:

- Fácil optimización en el mantenimiento del almacén de datos y de su correspondiente proceso ETL.
- Aumentar el rendimiento de las consultas.

Tareas que no se pueden llevar a cabo cuando los datos están en unidades físicas grandes:

- Reestructuración.
- Indexación.

- Recuperación.
- Monitoreo.

Los datos se pueden dividir por diferentes criterios:

- Por dato.
- Por línea de negocio.
- Por Geografía.
- Por unidad organizacional.

Ejemplo:

En la figura 1.8 se muestra la organización física de una tabla particionada en comparación con una tabla no particionada.

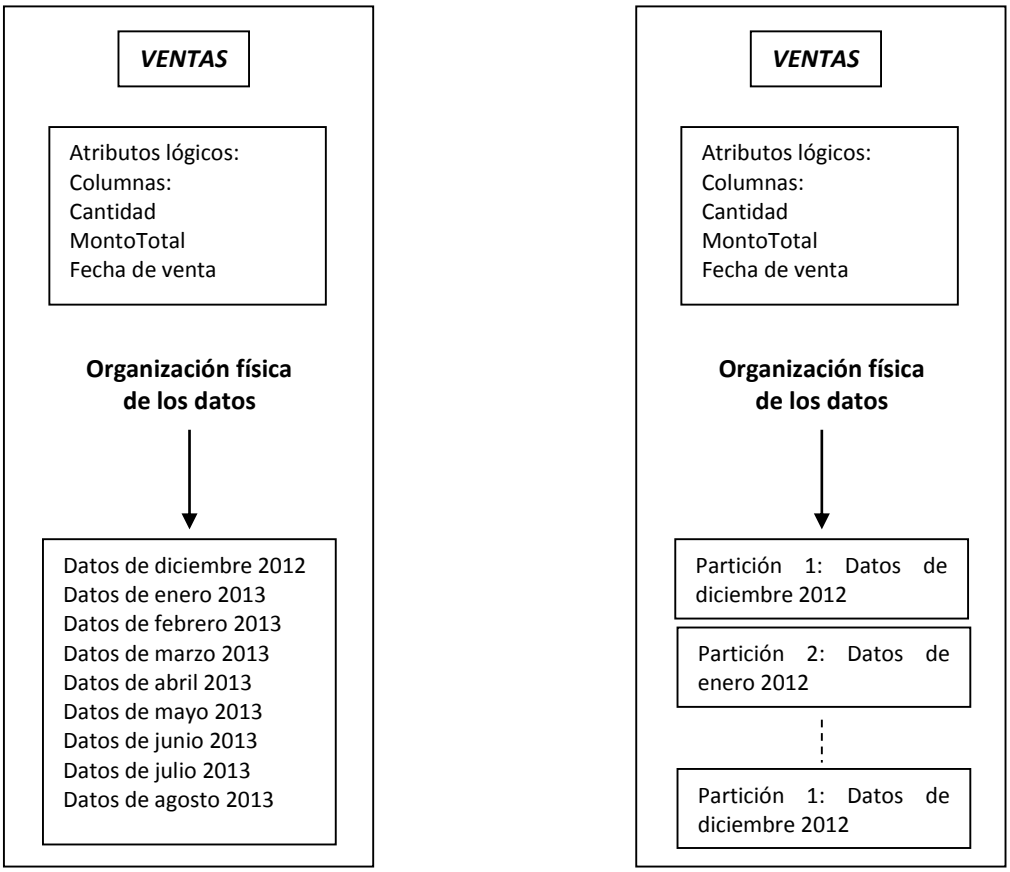

Figura 1.8. Comparación entre una tabla particionada y no particionada.

1.8.2.9. Llave sustituta (llave subrogada)

La llave sustituta es un concepto muy utilizado en bases de datos pero principalmente en entorno de almacén de datos e inteligencia de negocios. Una llave sustituta es un identificador único definido artificialmente que se le asigna a cada registro de una tabla de dimensión. Está llave generalmente no tiene ningún sentido específico de negocio. Son de tipo numérico secuencial y no tienen relación directa con ningún dato.

En un sistema operacional utilizan sus propias llaves, aunque suelen ser de tipo caracter y tienen sentidos específicos para los empleados de las compañías. Por ejemplo: La CURP de cada empleado puede ser llave única de la tabla empleados.

La llave sustituta suele utilizarse especialmente en tablas de dimensiones versionadas o históricas, conocidas como Slowly Changing Dimension (SCD) de tipo 2, es decir, tablas dimensiones que almacenan datos actuales (versión actual) como datos históricos (versiones antiguas).

Entonces porqué crear llaves sustitutas como nuevos identificadores. A continuación se explican algunas razones:

- **Fuentes heterogéneas:** Un almacén de datos se alimenta de varias fuentes de datos, cada una de ellas con sus propias llaves, por lo tanto es arriesgado asumir un código de alguna aplicación en particular. ¿Qué ocurrirá si posteriormente se añade información de una aplicación que tiene su propia llave? seguro que esto generará un problema.

- **Cambios en las aplicaciones de orígenes:** Puede ocurrir que cambie la lógica operacional de alguna llave que hubiésemos supuesto única. Por ejemplo ¿Qué pasará si un empleado llega sin CURP? Lo mejor es crear nuestras propias llaves sustitutas desde el principio del proyecto.

- **Rendimiento:** Ocupan menos espacio y dan más rendimiento que las tradicionales llaves naturales y más si estas últimas son de tipo texto. Por ejemplo: identificar una ciudad con 5 bytes o una persona de 9 bytes es un desperdicio considerable en espacio, pero no el espacio es el principal problema, sino el tiempo que se pierde en leer el dato.

1.8.2.10. Dimensiones lentamente cambiantes

Dimensiones lentamente cambiantes o SCD (Slowly Changing Dimensions). Habitualmente se encuentran tablas de bases de datos, las cuales es posible que produzcan modificaciones sobre sus datos a lo largo del tiempo y además es necesario poder realizar un seguimiento en sus cambios a lo largo del tiempo. En entorno de almacén de datos e inteligencia de negocios, este tipo de tablas suelen tratarse de tablas de dimensiones las cuales se denominan Slowly Changing Dimensions o simplemente tablas SCD.

Cuando ocurren estos cambios, se puede optar por seguir alguna de estas dos grandes estrategias:

- Registrar el historial de cambios.
- Reemplazar los valores que sean necesarios.

Tipos de estrategias SCD:

- SCD Tipo 1: Sobrescribir.

- SCD Tipo 2: Añadir fila.
- SCD Tipo 3: Añadir columna.
- SCD Tipo 4: Tabla de Historia separada.
- SCD Tipo 6: Híbrido.

1.9. Modelo dimensional

Modelo dimensional es el nombre que se le da a una técnica de diseño lógico que se utiliza para la construcción de almacenes de datos. Contiene la misma información que un modelo ER, pero principalmente se centra en que debe ser entendible, el acceso a la información debe ser rápido e intuitivo para el usuario.

Se basa en el diseño y la construcción de una estructura lógica y física denominada diagrama estrella, que es utilizada para representar a los data marts. Además ayuda al rendimiento de las consultas a causa de la desnormalización.

1.9.1. Modelos básicos dimensionales

- Modelo estrella.
- Modelo copo de nieve.

1.9.1.1. Modelo estrella

Un modelo estrella o en inglés conocido como Start Schema es un modelo de datos donde tiene una tabla principal llamada hechos que contiene datos de análisis y está rodeada de tablas de dimensiones.

En la figura 1.9 se muestra el modelo estrella.

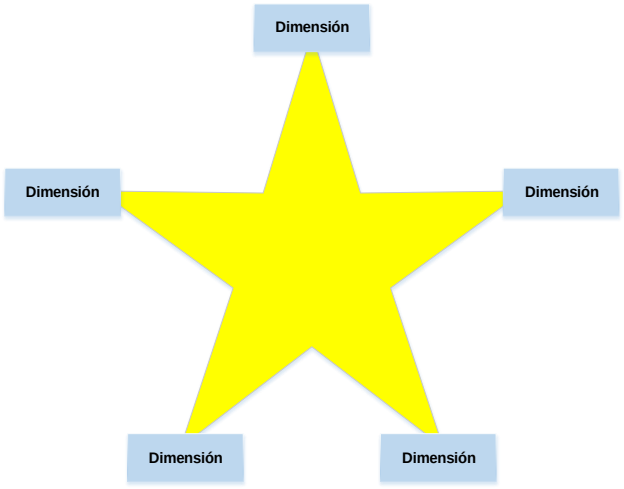

Figura 1.9. Modelo estrella.

Los principales componentes de un modelo estrella son:

- Hechos.
- Dimensiones.
- Medidas o métricas.

Hechos: Una tabla de hechos es la tabla principal del modelo estrella, donde se almacenan las medidas o métricas las cuáles describen eventos específicos de negocios. En otro concepto podemos decir que una tabla de hechos es una representación de un proceso de negocio.

La tabla de hechos contiene dos o más llaves foráneas unidas por sus respectivas tablas de dimensiones.

Dimensiones: Una tabla de dimensión es la tabla que acompaña a la tabla de hechos. Cada dimensión es definida por su llave primaria que sirve como base de integridad referencial con la tabla de hechos con la que es unida.

Características:

- Describen las entidades dentro del negocio.
- Contiene atributos que proporcionan contexto necesario para los datos numéricos.
- La tabla de dimensión por lo general tiene una llave que se auto-genera.

Dimensiones más habituales:

- Tiempo.
- Geografía.
- Cliente.
- Vendedor.

Medidas o métricas: Las medidas o métricas representan valores que son analizados. Por ejemplo: unidades de ventas o números de empleados. Por lo general son numéricas, puesto que estos valores son la base desde la cual pueden efectuarse cálculos.

1.9.1.2. Modelo copo de nieve

El modelo copo de nieve o en inglés conocido como Snowflake Schema es una variación del modelo estrella, en el cuál las jerarquías existentes en una tabla de dimensión son almacenadas en múltiples tablas de dimensiones.

En la figura 1.10 se muestra el modelo copo de nieve.

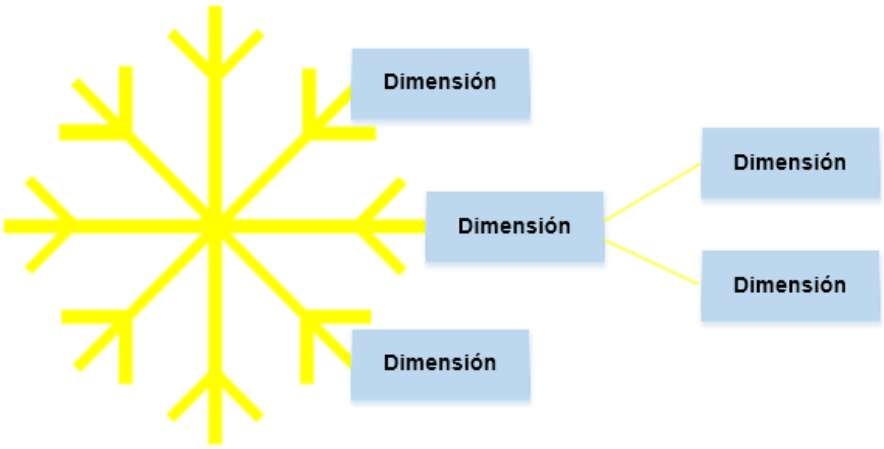

Figura 1.10. Modelo copo de nieve.

1.9.2. Diferencias entre el modelo estrella y copo de nieve

En la tabla 1.4 se muestran las diferencias entre el modelo estrella y copo de nieve.

Estrella	Copo de nieve
Desnormalizado.	Forma normalizada de tablas de dimensiones (solo las dimensiones primarias están enlazadas con la tabla de hechos).
Modelo simple.	Modelo complejo.
Habilidad para el análisis dimensional.	Rompe el esquema del análisis dimensional.
Se usa con una cantidad de datos moderada.	Se usa con gran cantidad de datos.
Optimiza el tiempo de respuesta de la base de datos.	Las consultas se realizan en mayor tiempo, ya que hace referencia a más tablas.
Sentencia SQL simple.	Sentencia SQL más compleja.

Tabla 1.4. Diferencias entre el modelo estrella y copo de nieve.

1.9.3. Ventajas del modelo dimensional

- **El modelo dimensional presenta importantes ventajas que el modelo relacional carece:** uno de los puntos fuertes del modelo dimensional, es que el marco predecible del modelo estrella resiste a los cambios inesperados en el comportamiento del usuario:
    - Dimensiones que cambian lentamente.
- **Flexibilidad:** todas las tablas pueden modificarse simplemente agregando nuevos registros de datos o se pueden incluir nuevas dimensiones al modelo:
    - Agregar atributos a las dimensiones.
    - Agregar nuevas dimensiones, siempre que exista un único valor de dicha dimensión definido para cada registro de la tabla de hechos.
    - Agregar medidas a la tabla de hechos, siempre que sean consistentes con el mayor nivel de detalle de las dimensiones.

1.10. Modelo multidimensional

En un modelo multidimensional, la información se representa como matrices multidimensionales, cuadros de múltiples entradas. El cubo queda determinado por un conjunto de datos para cada eje y un conjunto de datos para la matriz. A los ejes se les llaman dimensiones y a los datos que representan la matriz, se les llaman medidas. A los elementos del producto cartesiano de los ejes (dimensiones) se les llaman coordenadas.

Se utilizan principalmente para crear cubos OLAP. Permiten procesar gran volumen de información, acceso inmediato a los datos para su consulta y posteriormente para su análisis.

Cuando una base de datos puede ser visualizada como un cubo de tres o más dimensiones es mucho más fácil para el usuario organizar la información e imaginarse una rebanada de cada cubo a través de sus dimensiones para buscar la información deseada.

En la figura 1.11 se muestra la representación matricial de un cubo multidimensional, en donde las variables asociadas existen a lo largo de varios ejes o dimensiones y la intersección de la misma representa el hecho que se está evaluando.

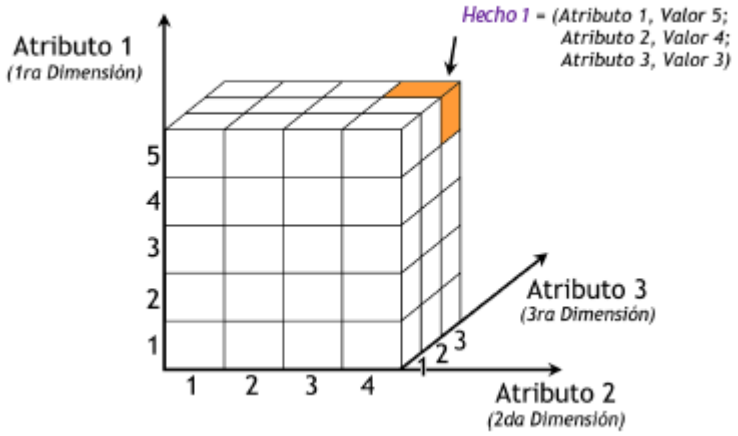

Figura 1.11. Cubo multidimensional. Propuesta por Dario, B. [7].

“Si no conozco una cosa,
la investigaré.”

- Louis Pasteur (1822-1895); Químico y microbiólogo francés.

Capítulo 2

Metodología de William H. Inmon

En este capítulo se presentan conceptos y definiciones generales sobre almacenes de datos, que son utilizados en capítulos posteriores.

Los temas desarrollados en este capítulo permiten establecer un marco teórico acerca de cómo diseñar y construir un almacén de datos.

2.1. Arquitectura de Fábrica de Información Corporativa

La Fábrica de Información Corporativa o Corporate Information Factory (de aquí en adelante lo llamaremos como CIF por sus siglas en inglés) fue presentada por primera vez por W.H. Inmon a principios de 1980 [17]. La Fábrica de Información Corporativa, es una arquitectura conceptual, que proporciona inteligencia y administración de negocios, impulsados por las operaciones de negocio. Los datos se organizan mediante un modelo ER y son el centro que proporciona datos para los data marts.

Los componentes de esta arquitectura están divididos en dos grupos de componentes con sus respectivos procesos como se muestra en la figura 2.1.

El primer grupo, denominado como Getting data in, contiene los procesos y bases de datos necesarias para la adquisición de los datos de los sistemas operacionales. En este grupo, los datos se integran, limpian y se almacenan en una base de datos para su fácil utilización.

El segundo grupo, denominado Getting information out, contiene los procesos y bases de datos necesarias para entregar la información a los usuarios y analistas finales.

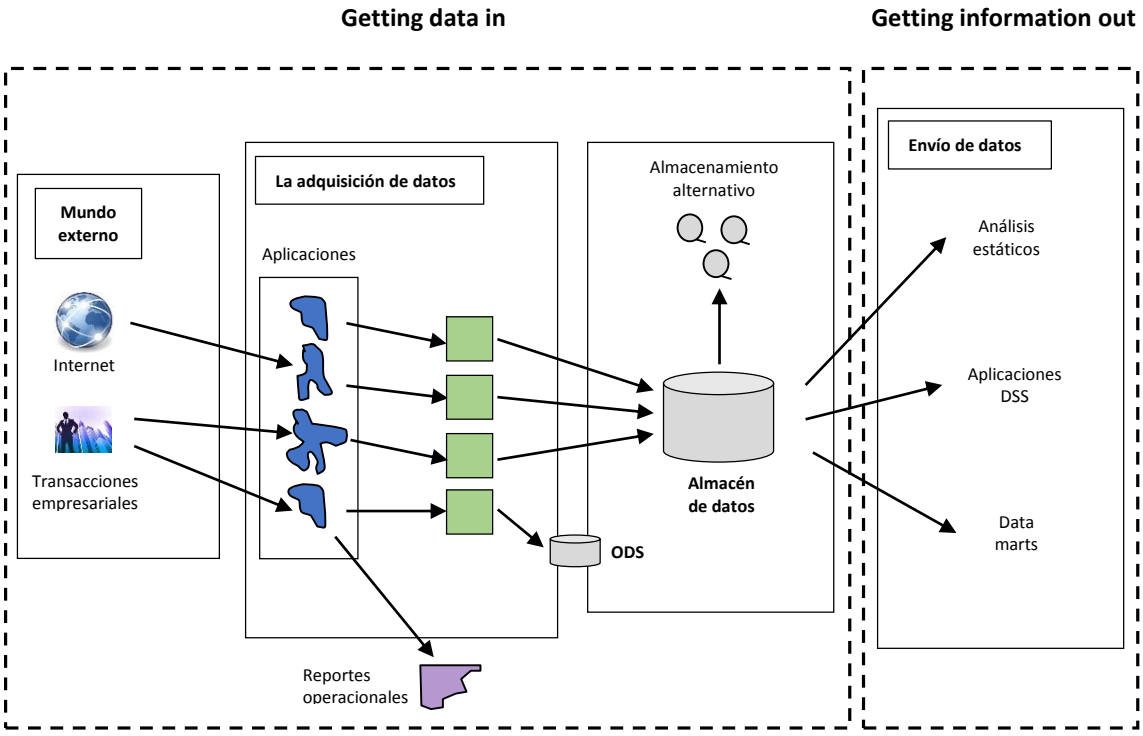

Figura 2.1. Arquitectura básica de la Fábrica de Información Corporativa (CIF). Propuesta por H Inmon, W., Imhoff, C., y Sousa, R. [17].

Antes de empezar con el desarrollo de un almacén de datos, se debe poner mucha atención en aspectos fundamentales que se muestran a continuación, ya que son la base del modelo del almacén de datos.

2.2. Modelo de desarrollo

2.2.1. Descripción del modelo de negocio

El modelo de negocio es la base de todos los modelos de sistemas y por lo tanto del desarrollo del modelo del almacén de datos.

Construir el modelo del almacén, consiste en la trasformación del modelo de negocio con ocho pasos definidos que posteriormente se detallará.

2.2.1.1. Escenario de negocio

Se utiliza el escenario de negocio de un fabricante de artículos o productos de limpieza para desarrollar el modelo de tema, modelo de datos de negocio y el modelo de datos del almacén de datos.

Tras la descripción de la situación empresarial tomando de [7], entraremos con el desarrollo del modelado de tema.

Descripción de la empresa comercial

La descripción de la empresa comercial servirá para desarrollar el ejemplo práctico de cada metodología en capítulos posteriores.

La empresa desarrolla actividades comerciales de mayorista y minorista de artículos de limpieza, en un ambiente geográfico de alcance nacional. De acuerdo a su volumen de operaciones, se le puede considerar de tamaño mediano.

Objetivos

El objetivo principal de la empresa comercial es el de maximizar sus ganancias. Pero también, se puede adicionar el objetivo de expandirse a un nuevo nivel de mercado, con el fin de conseguir una mayor cantidad de clientes y posicionarse competitivamente sobre sus rivales.

Políticas

La empresa comercial posee escasos clientes con un gran poder adquisitivo y son precisamente estos, los que adquieren el volumen de los productos que se comercializan.

Debido a ello, la política que se utiliza para cubrir los objetivos antes mencionados, es la de satisfacer ampliamente las necesidades de sus clientes, brindándoles confianza y promoviendo un ambiente familiar entre los mismos. Esta acción se realiza con el fin de mantener los clientes actuales y para que nuevos se interesen en su forma de operar.

Existe otra política que es implícita, por lo cual, no está definida tan estrictamente como la anterior y es la de mejorar continuamente, con el objetivo de calmar las exigencias y cambios en el mercado en el que actúa y para conseguir una mejor posición respecto a sus competidores.

Estrategias

Dentro de las estrategias existentes, se han destacado dos por considerarse más significativas, que son:

- Expandir el ámbito geográfico, creando varias sucursales en puntos estratégicos del país.
- Añadir nuevos rubros a su actividad de comercialización.

Organigrama

El Organigrama de la empresa comercial está representado como se muestra en la figura 2.2.

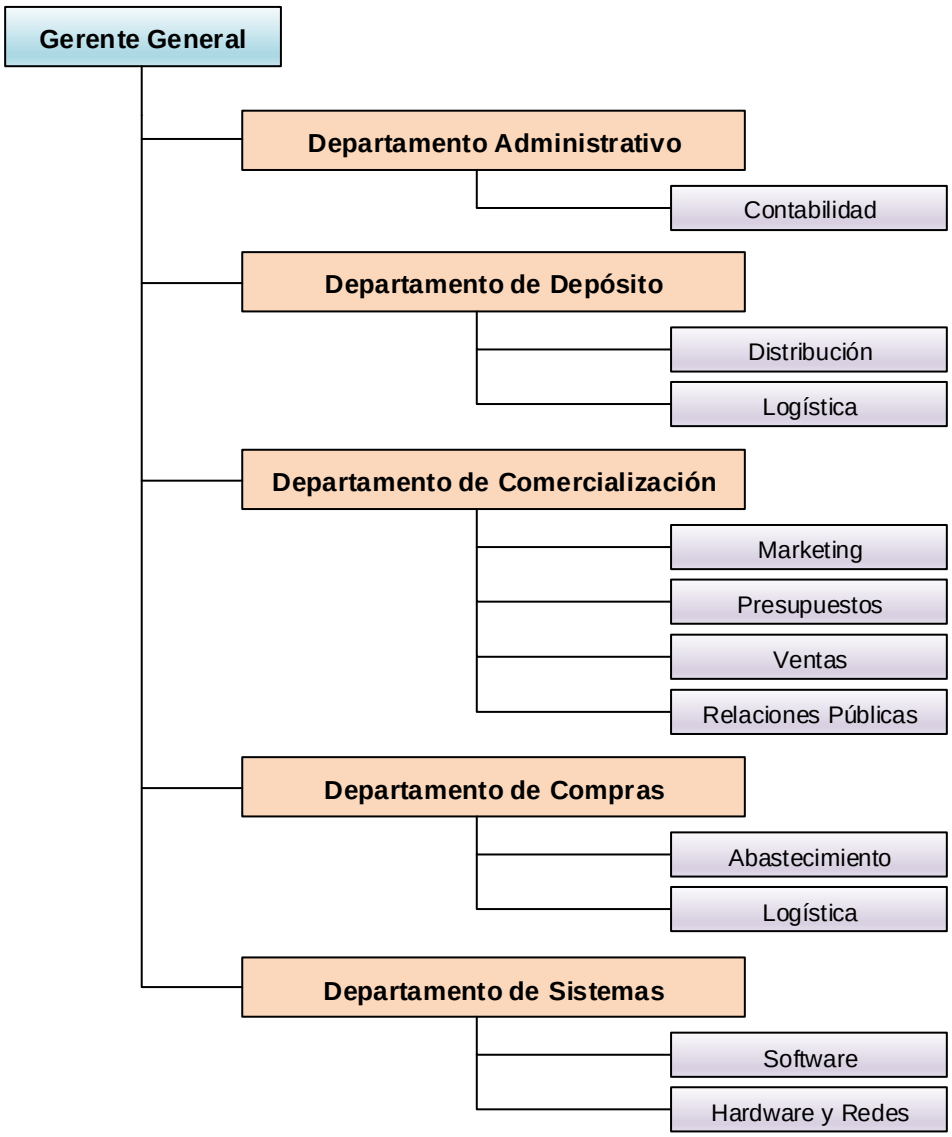

Figura 2.2. Organigrama de la empresa comercial. Propuesta por Dario, B. [7].

Relación de las metas de la organización con las del almacén de datos

El almacén de datos coincide con la metas de la empresa comercial, ya que esta necesita mejorar su eficiencia en la toma de decisiones y contar con información detallada a tal fin. Esto es vital, ya que es muy importante para procurar una mayor ventaja competitiva, conocer cuáles son los factores que inciden directamente sobre su rentabilidad, como así también, analizar su relación con otros factores y sus respectivos por qué.

El almacén de datos aportará un gran valor a la empresa comercial, entre las principales ventajas e inconvenientes que solucionará se pueden mencionar las siguientes:

- Permitirá a los usuarios tener una visión general del negocio.
- Transformará datos operativos en información analítica, enfocada a la toma de decisiones.
- Se podrán generar reportes dinámicos, ya que actualmente son estáticos y no ofrecen ninguna facilidad de análisis.
- Soportará la estrategia de la empresa comercial.
- Aportará a la mejora continua de la estructura de la empresa comercial.

Procesos

Los principales procesos que se llevan a cabo son los siguientes:

- Ventas:
    - Minorista: es la que se le realiza a los clientes particulares que se acercan hasta la empresa comercial para adquirir los productos que requieren.
    - Mayorista: es la que se le efectúa a los grandes clientes, ya sea por medio de comunicaciones telefónicas, o a través de visitas o reuniones.
    - Al realizarse una venta, el departamento de depósito se encarga de controlar el stock, realizar encargos de mercadería en caso de no cubrir lo solicitado, armar el pedido y enviarlo por medio de transporte propio o de terceros al destino correspondiente.
- Compras:
    - El departamento de compras, al recibir del departamento de depósito las necesidades de mercadería, realiza una comparación de los productos ofrecidos por sus diferentes proveedores en cuestión de precio, calidad y confianza. Posteriormente, se efectúa el pedido correspondiente.

Caso práctico: El proceso de negocio elegido es Ventas.

Las preguntas de negocio son:

- Se desea conocer cuántas unidades de cada producto fueron vendidas a sus clientes en un periodo determinado.
- Se desea conocer cuál fue el monto total de ventas de productos a cada cliente en un periodo determinado.

La dimensión tiempo es fundamental en el almacén de datos.

En el OLTP de la empresa analizada, el proceso de negocio Ventas está representado por el modelo ER como se muestra en la figura 2.3.

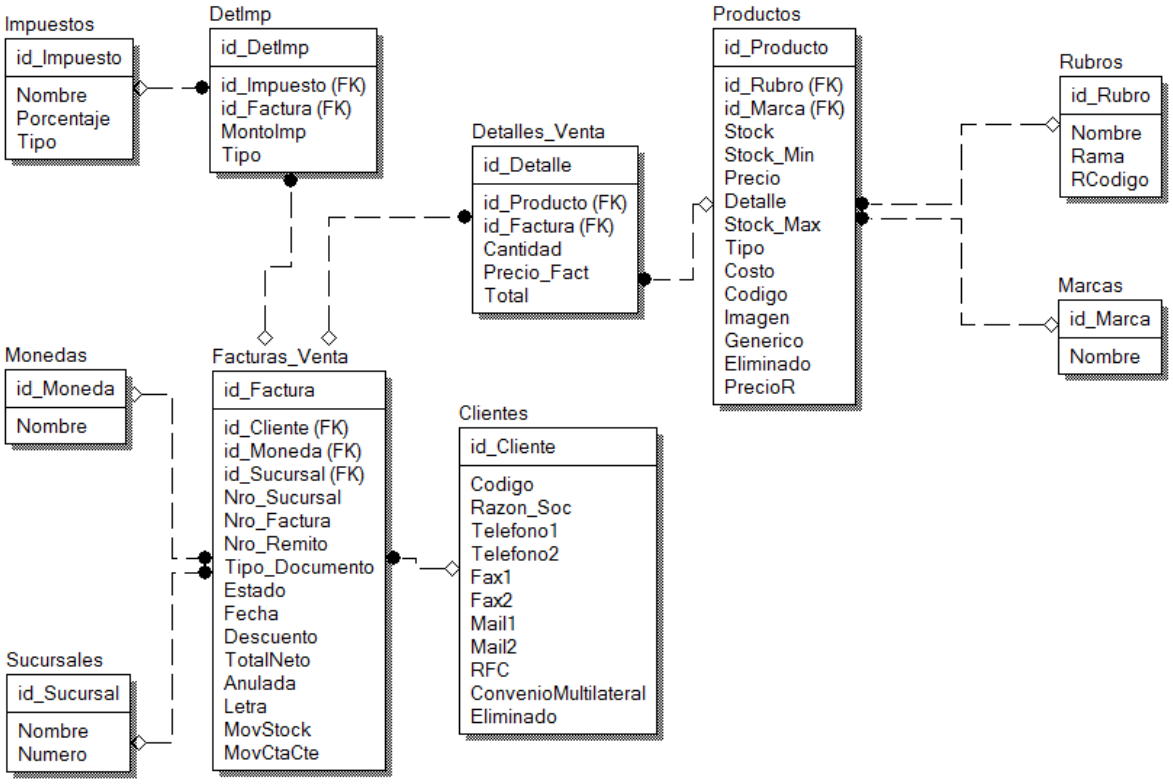

Figura 2.3. Empresa comercial, Modelo Entidad Relación. Propuesta por Dario, B. [7].

2.2.1.2. Modelado de tema

Un almacén de datos está organizado por temas (los temas son importantes agrupaciones de elementos físicos, conceptos, personas, lugares y eventos de interés para la empresa), por lo tanto para el modelo de datos de un almacén comenzar con el modelado de tema es normal. En un sistema operacional tradicional este paso a menudo se omite, ya que está orientado hacia las funciones y procesos de negocios específicos y la eficiencia con la que puedan procesar las transacciones. Entonces en un almacén de datos la orientación temática es el centro del diseño físico de la base de datos. Los principales procesos de una organización están representados en las fuentes de los sistemas operacionales.

Una empresa que desarrolla primero el modelo de tema, puede beneficiar el trabajo de los demás desarrolladores, ya que no tendrán que empezar de cero.

En las empresas u organizaciones, existen muchos temas en común, ya que prácticamente todas las empresas, tienen clientes, proveedores, productos, servicios, etc., por lo tanto estos son candidatos para empezar a identificar los temas. En la tabla 2.1 se muestran temas para un modelo de datos genérico y algunos temas serán tomados para la empresa comercial.

Temas	Definición	Ejemplos	Observación
Clientes	Las personas o empresas que adquieren los artículos o productos.	• Clientes. • Consumidores.	Que tipos de productos utilizan más frecuente.
Geografía	Área geográfica.	• Países. • Estados. • Ciudades. • Barrios.	Esto puede incluir otros tipos de localizaciones como: correo electrónico, número de teléfono, etc.
Productos	Los artículos o productos relacionados con la empresa se ponen a disposición a los clientes.	• Productos. • Servicios.	Se hace un seguimiento de los artículos de la empresa, para apoyar a decisiones futuras
Ventas	Operaciones que cambian el control de un producto de la empresa a un cliente.	• Transacciones de ventas. • Detalles de las transacciones de ventas.	Las ventas son las intersecciones del cliente, producto, etc.
Proveedores	Personas que se encargan de proveer bienes y servicios de la empresa.	• Agentes. • Fabricantes. • Proveedores.	En el caso de un contratista, la persona que realiza el trabajo se incluye en el área de recursos humanos y la empresa que proporciona a esa persona está incluida en proveedores.
Finanzas	Información sobre dinero que se recibe, retiene, gasta o seguido por la empresa.	• Dinero. • Cobrar. • Pagar.	

Tabla 2.1. Posibles temas para la empresa comercial. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Proceso de desarrollo del modelado de tema

Existen tres métodos principales de desarrollo para el modelado de tema y la elección dependerá de los involucrados. A continuación se describe cada método:

1. Habitación cerrada.
2. Entrevistas.
3. Sesiones facilitadas.

Habitación cerrada

Este tipo de desarrollo en habitación cerrada, los modeladores de temas trabajan con poca o ninguna participación con los representantes de la empresas, ya que se supone que el modelador entiende el negocio y desarrolla el tema a partir de sus percepciones sobre la empresa. No es muy recomendable este método ya que el modelador rara vez entiende completamente el negocio.

Entrevistas

Las entrevistas es una buena forma de obtener información de los representantes de las empresas y de los representantes principales de los departamentos de las empresas. Cada representante se les debe pedir una descripción del flujo de trabajo de su área. Con esta información el modelador debe identificar los principales grupos de información de interés para cada persona.

Sesiones facilitadas

Este método es considerado por muchos autores como el más eficiente. Ya que incluye además de las entrevistas, la interacción entre las personas involucradas. Se pueden hacer una o dos sesiones facilitadas. El modelador debe explicar qué es un tema, como se deben identificar y definirlas, para que los involucrados participen en una sesión de lluvias de ideas y luego identificar los temas.

No es raro ver que las personas involucradas identifiquen informes, procesos, funciones, atributos, entidades, organizaciones, etc.

Cada uno de los métodos tiene sus ventajas y desventajas, como se muestra en la tabla 2.2.

Método	Descripción	Ventajas	Desventajas
Habitación cerrada	El modelador de los datos de los modelos de tema, con base a la información que tiene lo desarrolla y se encarga de presentarlo para su aprobación.	• El modelador entiende los procesos. • Un modelo puede desarrollarse rápidamente.	• El modelador puede no tener suficiente conocimiento del negocio.
Entrevistas	El modelador hace entrevistas individuales y utiliza la información para desarrollar los modelos de tema.	• Cada persona de los entrevistados tiene la oportunidad de contribuir. • Los entrevistados poseen conocimiento del negocio.	• Las entrevistas individuales requieren más tiempo. • Mientras se obtiene el conocimiento, no se pueden construir los modelos de tema.
Sesiones facilitadas	Un facilitador lidera a un grupo de personas de la empresa para desarrollar los modelos de tema.	• Los colaboradores poseen conocimiento del negocio. • Se desarrolla a través de la interacción.	• La programación de los participantes puede ser difícil. Se desarrolla a través de la interacción.

Tabla 2.2. Métodos diferentes para el desarrollo de tema. Propuesta por Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Independientemente del método seleccionado anteriormente, se deben identificar diversos temas para su análisis en diferentes sesiones. Para la empresa comercial, se identificaron los posibles temas como se muestra en cada página de cada uno de los documentos de la figura 2.4.

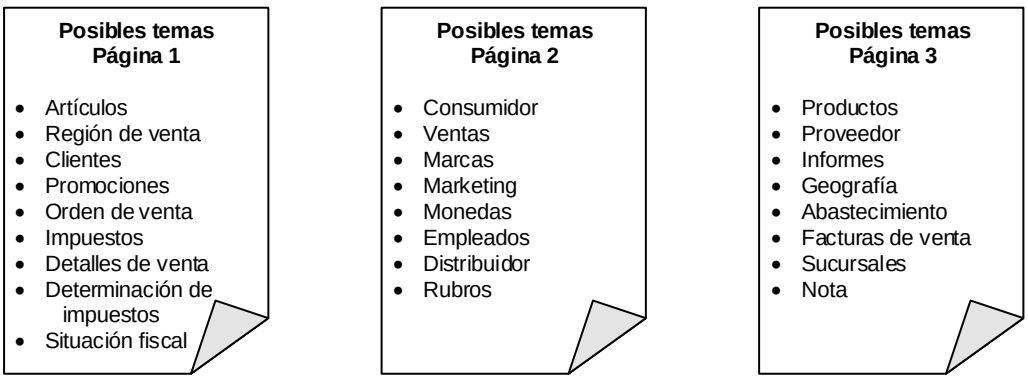

Figura 2.4. Identificación de temas, primera sesión. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

El siguiente paso es excluir elementos que no son posibles temas. Cuando este proceso termine, deberá haber menos temas.

Algunas transformaciones que se llevaron a cabo son las siguientes:

- Artículos y Productos son lo mismo, es decir que se utiliza para todos los productos.
- Clientes y Consumidor están decididos a ser lo mismo. Cliente fue seleccionado como el término que se utilizará.
- Geografía y Región de venta, no se mencionan como parte para responder a las preguntas de negocio.
- Marketing se determinó que es una función y por lo mismo es eliminado.

Los posibles temas principales se han identificado como se muestra en la figura 2.5. Pero existen temas que pueden ser más significativos que otros, pero finalmente quedan como se muestra en la figura 2.6.

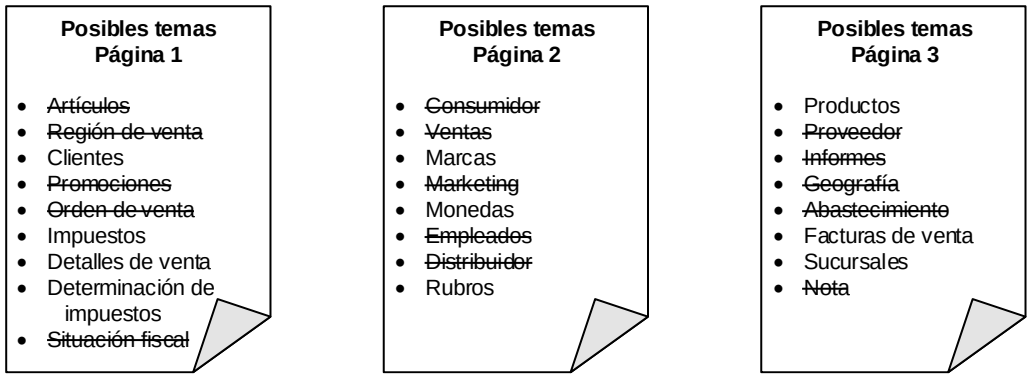

Figura 2.5. Redefiniendo temas. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

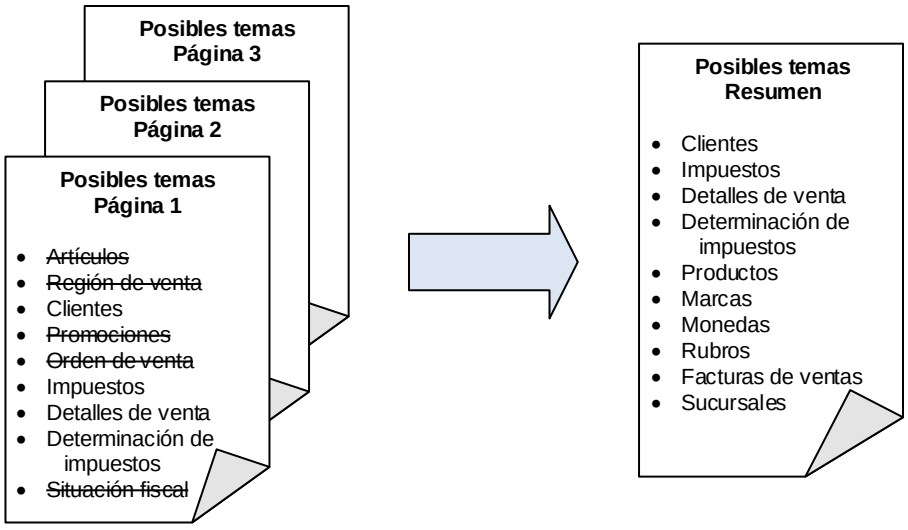

Figura 2.6. Resultado de reducción de temas. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Modelo de tema para la empresa comercial

En la figura 2.7 se muestra el modelo de tema para este ejemplo con los posibles temas.

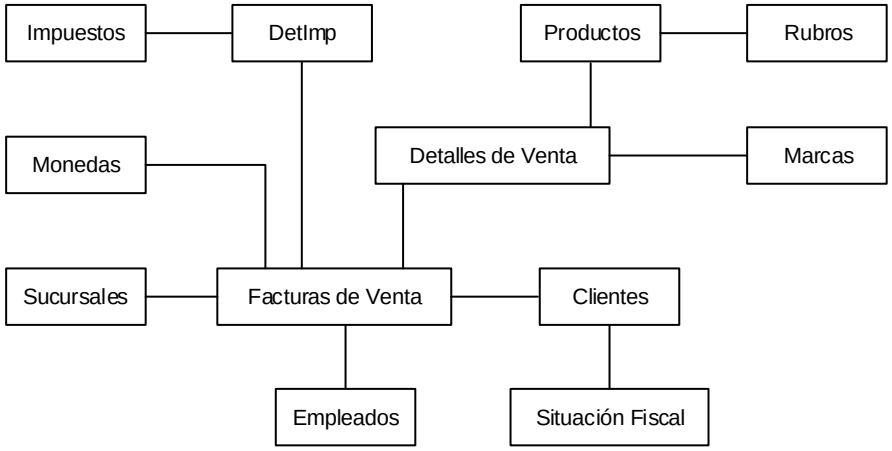

Figura 2.7. Posibles temas. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

En la figura 2.8 se muestra el modelo de tema con los temas reducidos.

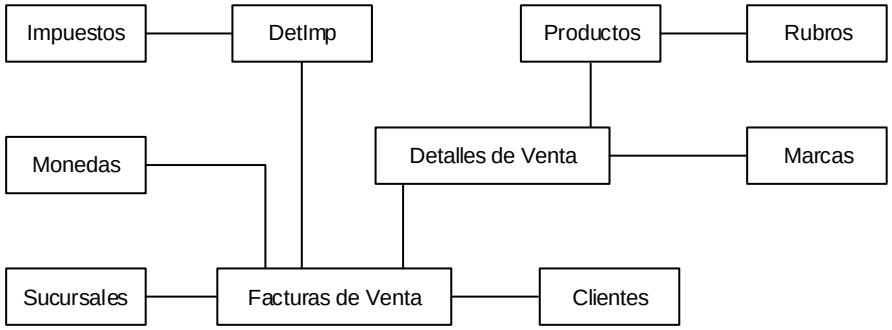

Figura 2.8. Temas reducidos. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

2.3. Enfoque de William H. Inmon

Inmon presenta dos caminos al momento de construir un almacén de datos. Uno se conoce como plan de migración y el otro una metodología.

El plan de migración describe actividades generales, mientras tanto la metodología describe las actividades específicas. Para este ejemplo, nos centraremos en el plan de migración en detalle.

2.3.1. Plan de migración

El plan de migración es un modelo de datos de negocio o modelo de datos corporativo que representa las necesidades de la organización, no necesariamente lo que realmente tiene. Además se construye sin consideración de alguna tecnología para su desarrollo.

Modelo de datos de negocio

Primeramente hay que definir qué es un modelo de datos. Un modelo de datos es una colección de herramientas conceptuales para la descripción de datos, relaciones entre datos, semántica de los datos y restricciones de consistencia.

El modelo de datos de negocio es un tipo de modelo y representación de los datos en un entorno de negocio. Ayuda a las personas imaginar cómo la información de la empresa se relaciona con la información del negocio. El modelo de datos de negocio es para los sistemas operacionales, almacenes de datos y data marts. El modelo de tema que se explicó anteriormente es la base para el desarrollo del modelo de datos de negocio.

El modelo de datos de negocio puede ser construido internamente o haber sido generada a partir de un modelo de datos genérico. Si el modelo de datos de negocio no existe, debe desarrollarse antes de entrar de lleno en el desarrollo del modelado de datos del almacén.

Proceso de desarrollo del modelado de datos de negocio

El desarrollo de un modelo de datos de negocio puede tomar de 6 a 12 meses, sin entregable del negocio tangible [16]. Es recomendable seguir los siguientes pasos:

- Identificar los temas relevantes de la empresa.
- Identificar las principales entidades y establecer identificadores.
- Determinar las relaciones entre pares de entidades.
- Añadir atributos.
- Confirmar la estructura del modelo.
- Confirmar el contenido del modelo.

Ya que se tiene en modelo de datos de negocio, es necesario definir el sistema de registro, el cual define los sistemas que ya existen en la empresa. El sistema de registro es la identificación de la mejor información que tiene la organización en función del modelo de datos.

La determinación de la mejor fuente de datos existente utiliza los siguientes criterios:

- Datos más completos.
- Datos más oportunos.
- Datos más precisos.
- Datos más cercanos a la fuente de entrada.
- Datos más compatibles a la estructura del modelo de datos.

Si la actividad del modelo de datos de negocio se ha realizado de manera correcta, el diseño del almacén datos es más fácil. Sólo unos pocos elementos del modelo de datos de negocio necesitan ser cambiados para convertir en diseño de almacén.

Identificar los temas relevantes de la empresa

Los temas con la información necesaria para responder a las preguntas de negocio, planteadas en el escenario de negocio que se mostró en la figura 2.8 son: Productos, Impuestos, Monedas, Sucursales, DetImp, Facturas de Venta, Detalles de Venta, Clientes, Rubros, Marcas. Existen otros temas que son importantes, pero estos son los necesarios por lo menos para este ejemplo. Ahora si podemos ver el alcance de nuestro almacén de datos.

El modelo de tema que se creó anteriormente, se puede reducir a sólo responder preguntas más específicas con excluir más temas.

Identificar las principales entidades y establecer identificadores

Una entidad es un tipo de objeto sobre el que se recoge información: Persona, lugar, cosa, evento o en el concepto de interés para la empresa. Una forma de identificar las entidades es examinando las especificaciones de requerimientos de usuarios. Conforme se van identificando las entidades, se les dan nombres que tengan un significado y que sean obvias para el usuario.

Por ejemplo, Clientes es una entidad porque los clientes existen, sepamos o no sus nombres, direcciones y teléfonos. Siempre que sea posible, el usuario debe colaborar en la identificación de las entidades. Una entidad en particular tendrá un valor para cada uno de sus atributos.

Finalmente el modelo se trasformará en un modelo de datos físico. Cada tabla de la base de datos requiere una llave que lo identifique. Por lo tanto cada entidad a modelar debe tener su identificador.

Cada entidad tiene al menos un identificador. En este paso, se trata de encontrar todos los identificadores de cada una de las entidades. Los identificadores pueden ser simples o compuestos. Cabe resaltar que los temas no son las entidades, sino que los temas contienen a las entidades y en este caso se puede ver que los nombres de los temas y entidades son iguales.

Definición de Entidades:

Entidad	Definición
Productos	Artículos o productos que se ponen a disposición del cliente.
Impuestos	Impuesto al valor agregado.
Monedas	Tipo de moneda con la que se realiza la transacción de venta.
Sucursales	Sucursal en donde se realiza la venta.
DetImp	Determina el monto total de impuesto.
Facturas_Venta	Documento mercantil que refleja toda la información de una venta.
Detalles_Venta	Contiene información acerca de los detalles de las ventas realizadas.
Clientes	Los clientes son personas y organizaciones que adquieren un producto de la empresa comercial.
Rubros	Rubro al que pertenece el producto.
Marcas	Marca a la que pertenece un producto.

Tabla 2.3. Tablas de Entidades. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

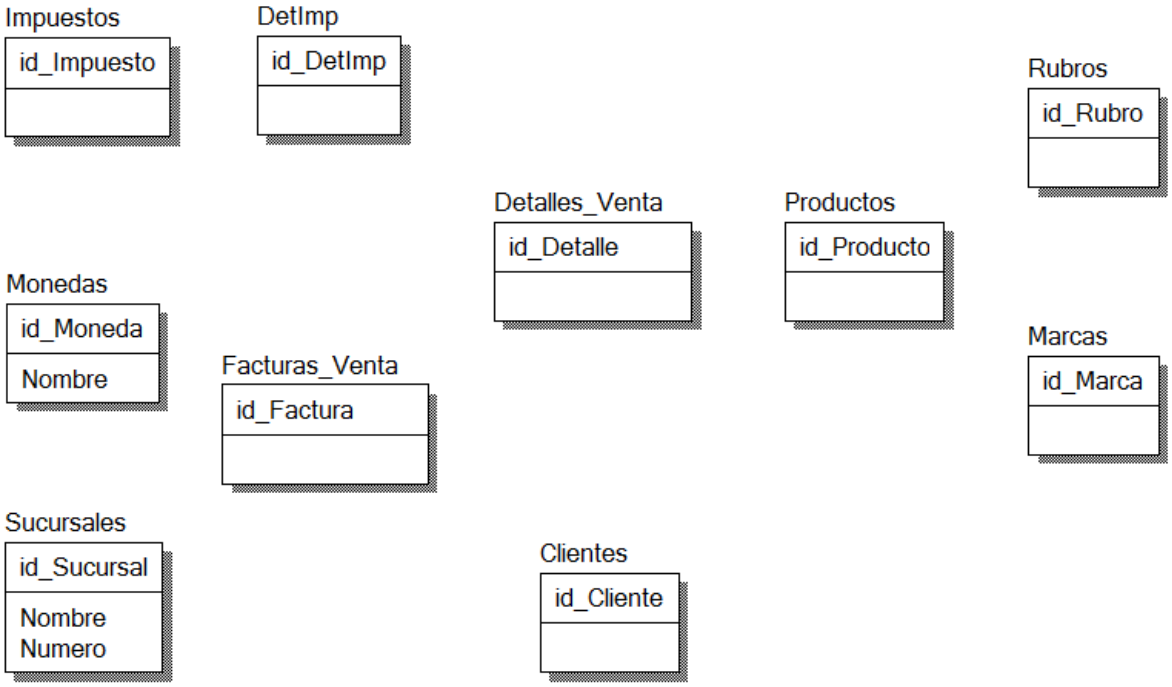

Figura 2.9. Entidades con sus respectivos indicadores (llaves primarias). Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Determinar las relaciones entre pares de entidades

Las relaciones representan esquemáticamente las reglas de negocio que deben ser reflejados en el modelo de datos de negocio. A continuación se muestran algunas reglas de negocio en el modelo de datos:

- Un Producto pertenece a un Rubro.
- Cada Rubro puede tener muchos Productos.
- Un Producto pertenece a una Marca.
- Una Marca tiene muchos Productos.
- Un Cliente puede generar muchas Facturas.
- Una Factura es generada sólo por un Cliente.

El siguiente paso en el proceso es definir las relaciones entre pares de entidades, como se muestra en la figura 2.10.

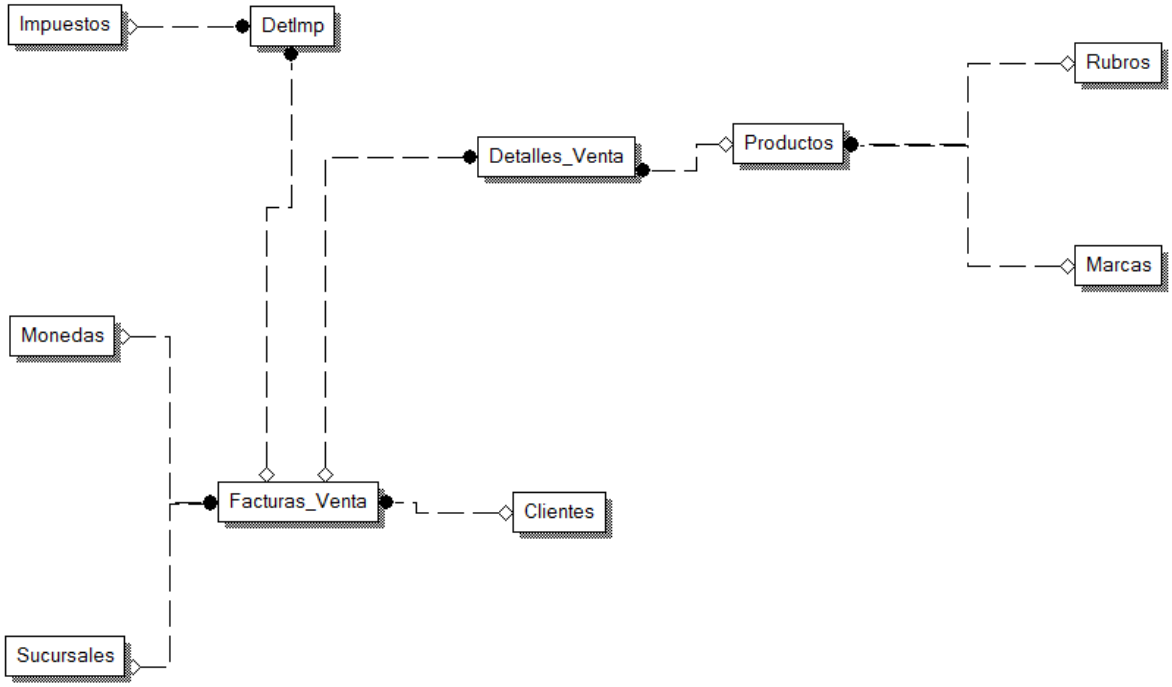

Figura 2.10. Modelo Entidad Relación - Entidades. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Añadir atributos

Un atributo es un hecho de información relacionada con la entidad. Al identificar los atributos, hay que tener en cuenta si son simples o compuestos. El escoger entre atributo simple o compuesto depende de los requerimientos de usuarios. Además se deben identificar los atributos derivados o calculados, que son aquellos cuyo valor se puede calcular a partir de los valores de otros atributos. Algunos diseñadores no representan los atributos derivados en los modelos conceptuales. Los atributos que ya se incluyeron en modelo anterior son los identificadores, que son atributos necesarios para responder a las preguntas de negocio.

En la figura 2.11 se muestra la ampliación del modelo de datos con los atributos incluidos.

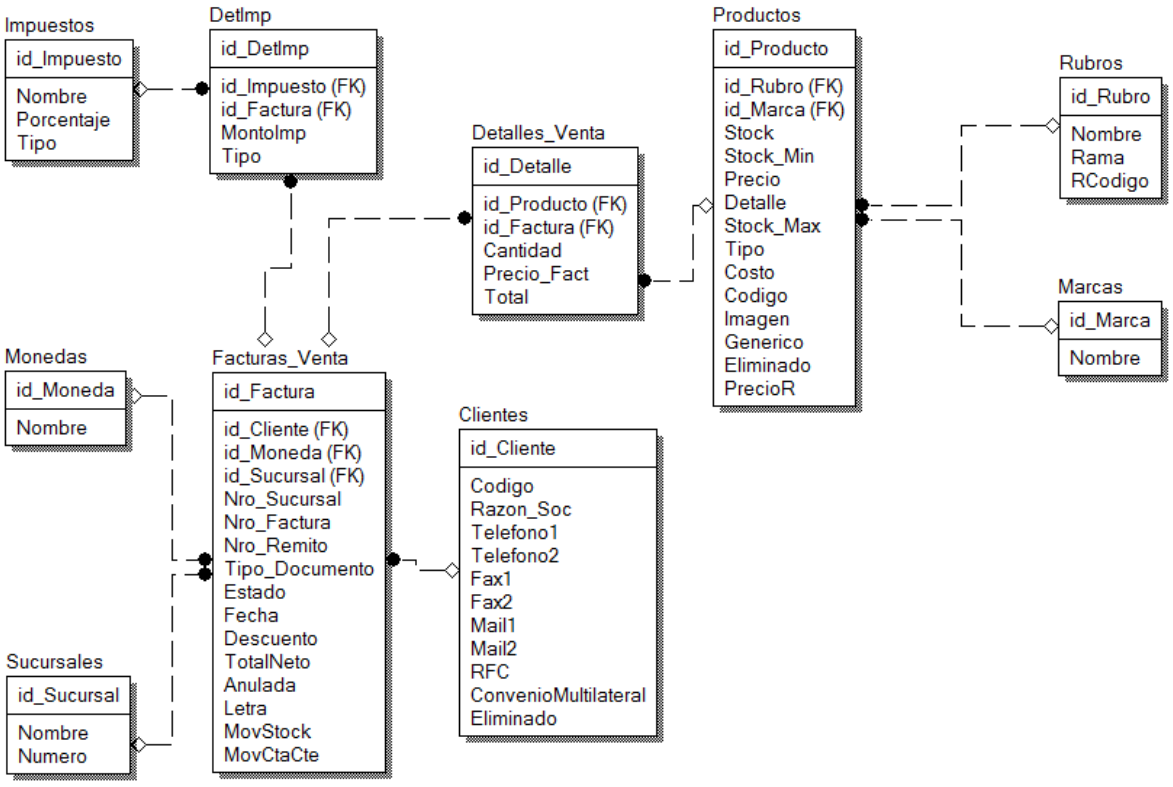

Figura 2.11. Modelo Entidad Relación - Entidades y Atributos. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Confirmar la estructura del modelo

El modelo de datos de negocio debe ser representado en un modelo en tercera forma normal (como se vio en el paso anterior), ya que proporciona el mayor grado de flexibilidad y estabilidad y garantiza el mayor grado de coherencia, en este caso es el Modelo Entidad Relación.

Confirmar el contenido del modelo

Antes de dar por finalizado el modelo de datos de negocio, debe revisarse el contenido del modelo con los representantes de las empresas. Este modelo está formado por el Modelo Entidad Relación y toda la documentación que describe el modelo. Si se encuentra alguna anomalía, hay que corregirla haciendo los cambios oportunos, por lo que posiblemente haya que repetir algunos de los pasos anteriores. Este proceso debe repetirse hasta que se está seguro de que el modelo de datos de negocio cumpla con los objetivos.

2.4. El desarrollo de la transformación del modelo de datos de negocio

Dado que uno de los objetivos del almacén de datos es proporcionar una visión coherente de los hechos y cifras de la empresa, es importante comenzar con un modelo que cumpla con ciertos criterios, el modelo de dato de negocio en tercera forma normal lo cumple para su trasformación, ya que proporciona consistencia en los datos y restringe su redundancia.

Sin embargo el modelo de datos de negocio en la tercera forma normal planteada al principio de este capítulo y desarrollada, no es la mejor estructura para un almacén de datos, es por ello que existe una metodología de ocho pasos para transformar el modelo de datos de negocio.

Metodología

Este proceso comienza con el modelo de datos de negocio aplicándole ocho pasos, para crear un modelo de datos optimizado y cumplir con los objetivos del almacén de datos. A continuación se detalla cada uno de los pasos de la metodología:

1. Seleccionar los datos de interés.
2. Añadir la dimensión de tiempo a las llaves.
3. Añadir datos derivados.
4. Determinar el nivel de granularidad.
5. Resumir los datos.
6. Unir entidades.
7. Crear arreglos.
8. Separar los datos.

Estos ocho pasos se pueden agrupar en dos categorías. Los primeros cuatro pasos pertenecen a los temas relacionados con la empresa y a la creación del almacén de datos y los siguientes cuatro pasos sirven para mejorar el rendimiento y optimizar el tiempo en respuesta del almacén de datos.

Paso 1. Seleccionar los datos de interés

El primer paso en el desarrollo del modelo del almacén de datos, es seleccionar los datos de interés. Existen dos razones principales para hacer esto. Primeramente el propósito y objetivos de negocio del almacén y en segundo lugar, limitar el alcance del modelo del almacén para no saturarlo con datos innecesarios.

Entradas

El modelo de negocio es solo una de las entradas para el paso 1. Pero existen otras entradas que incluyen el alcance del proyecto, requerimientos de información, prototipos, reportes y consultas existentes.

Una de las ventajas del modelo relacional para el almacén de datos, es de separar los data marts del almacén, ya que esto facilita la incorporación de nuevos requerimientos de información descubiertos, sin afectar a los data marts existentes.

Modelo de datos de negocio

El modelo de datos de negocio contiene cientos de elementos de datos. El equipo encargado del desarrollo, posiblemente tendrá que sacrificar elementos que no sean necesarios y no deban ser incluidos al almacén de datos. El equipo de desarrollo se debe encargar de crear el modelo de datos cuando este no exista para el alcance del proyecto. En la figura 2.12 se muestra el modelo de datos de negocio para este ejemplo a nivel de entidad.

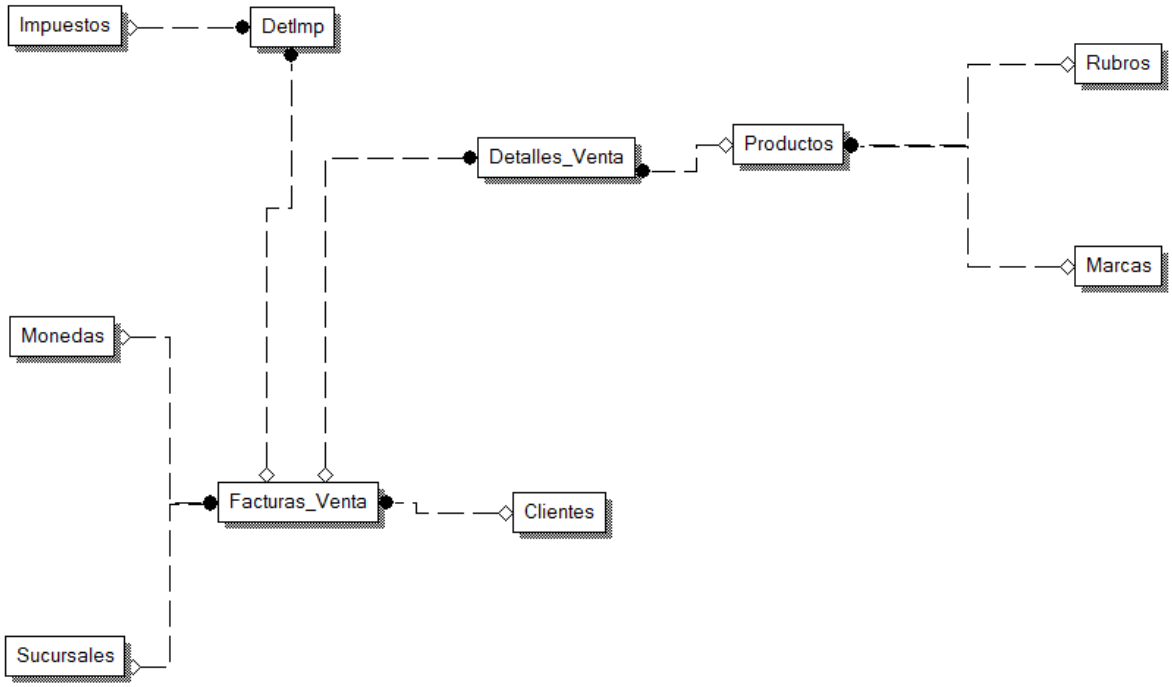

Figura 2.12. Modelo de datos de negocio de la empresa comercial. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Documento de alcance

Este documento de alcance debe establecer cuáles son las expectativas que se espera del almacén de datos: Identificar datos a ser incluidos al almacén. Las preguntas de negocio planteadas para este ejemplo en este capítulo debe proporcionar la información necesaria para el alcance del contenido del almacén de datos. En la figura 2.12 se muestra cómo el documento de alcance se puede utilizar para reducir las entidades que ya están contempladas en el proyecto.

Requerimientos de información

Los requerimientos de información es parte de las entradas. Un almacén de datos debe estar alineado con los objetivos de negocio, documentos de planeación de la empresa, sesiones de entrevistas con los ejecutivos, analistas y usuarios finales. Estas sesiones están diseñadas para identificar las cuestiones específicas de negocio que deben ser respondidas y elementos de datos que se necesitan para responder a esas preguntas.

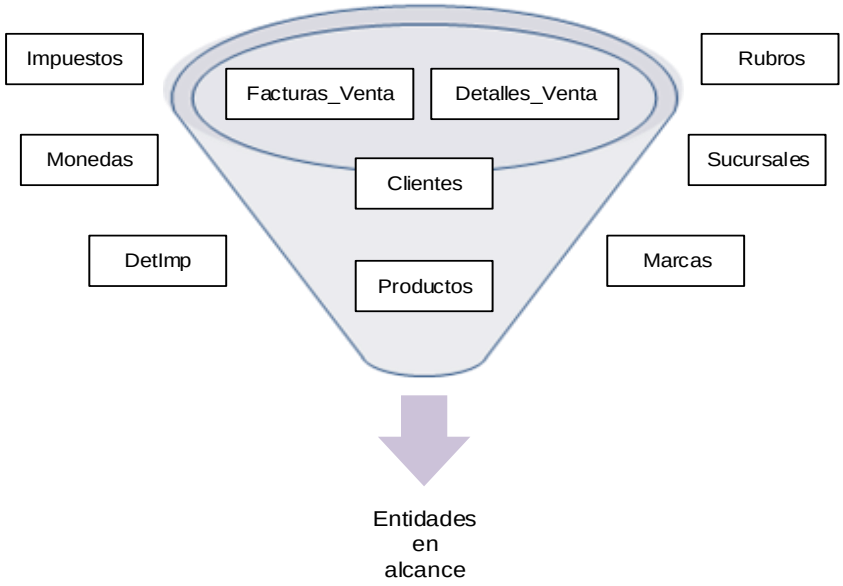

Figura 2.13. Reducción de entidades. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

El nivel de atributos del modelo de datos de negocio para este ejemplo se muestra en la figura 2.14. En las tablas 2.4, 2.5, 2.6, y 2.7 se listan los atributos disponibles dentro de los temas y la decisión de exclusión con respeto a la inclusión dentro del almacén de datos y solamente están basadas en las preguntas de negocio conocidas.

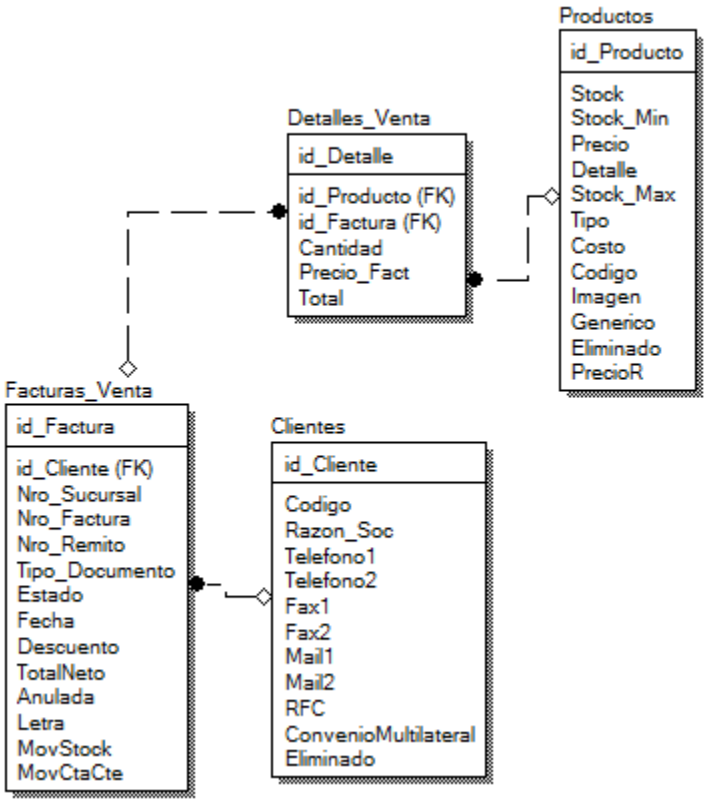

Figura 2.14. Temas de la empresa comercial. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Entidad	Atributo	Decisión	Justificación
Clientes	id_Cliente	Conservar	Necesario como llave primaria para establecer relaciones.
Clientes	Codigo	Omitir	No se necesita para responder alguna pregunta.
Clientes	Razon_Soc	Conservar	Necesario para saber el nombre del cliente.
Clientes	Telefono1	Omitir	No se necesita para responder alguna pregunta.
Clientes	Telefono2	Omitir	No se necesita para responder alguna pregunta.
Clientes	Fax1	Omitir	No se necesita para responder alguna pregunta.
Clientes	Fax2	Omitir	No se necesita para responder alguna pregunta.
Clientes	Mail1	Omitir	No se necesita para responder alguna pregunta.
Clientes	Mail2	Omitir	No se necesita para responder alguna pregunta.
Clientes	RFC	Omitir	No se necesita para responder alguna pregunta.
Clientes	ConvenioMultilateral	Omitir	No se necesita para responder alguna pregunta.
Clientes	Eliminado	Omitir	No se necesita para responder alguna pregunta.

Tabla 2.4. Exclusión de atributos de Clientes. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Entidad	Atributo	Decisión	Justificación
Productos	id_Producto	Conservar	Necesario como llave primaria para establecer relaciones.
Productos	Stock	Omitir	No se necesita para responder alguna pregunta.
Productos	Stock_Min	Omitir	No se necesita para responder alguna pregunta.
Productos	Precio	Omitir	No se necesita para responder alguna pregunta.
Productos	Detalle	Omitir	No se necesita para responder alguna pregunta.
Productos	Stock_Max	Omitir	No se necesita para responder alguna pregunta.
Productos	Tipo	Omitir	No se necesita para responder alguna pregunta.
Productos	Costo	Omitir	No se necesita para responder alguna pregunta.
Productos	Codigo	Omitir	No se necesita para responder alguna pregunta.
Productos	Imagen	Omitir	No se necesita para responder alguna pregunta.
Productos	Generico	Omitir	No se necesita para responder alguna pregunta.
Productos	Eliminado	Omitir	No se necesita para responder alguna pregunta.
Productos	PrecioR	Omitir	No se necesita para responder alguna pregunta.

Tabla 2.5. Exclusión de atributos de Productos. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Entidad	Atributo	Decisión	Justificación
Detalles_Venta	id_Detalle	Conservar	Necesario como llave primaria para establecer relaciones.
Detalles_Venta	Cantidad	Conservar	Necesario para calcular las unidades vendidas.
Detalles_Venta	Precio_Fact	Conservar	Necesario para calcular el monto total de ventas.
Detalles_Venta	Total	Omitir	No se necesita para responder alguna pregunta.

Tabla 2.6. Exclusión de atributos de Detalles_Venta. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Entidad	Atributo	Decisión	Justificación
Facturas_Venta	id_Factura	Conservar	Necesario como llave primaria para establecer relaciones.
Facturas_Venta	Nro_Sucursal	Omitir	No se necesita para responder alguna pregunta.
Facturas_Venta	Nro_Factura	Omitir	No se necesita para responder alguna pregunta.
Facturas_Venta	Nro_Remito	Omitir	No se necesita para responder alguna pregunta.
Facturas_Venta	Tipo_Documento	Omitir	No se necesita para responder alguna pregunta.
Facturas_Venta	Estado	Omitir	No se necesita para responder alguna pregunta.
Facturas_Venta	Fecha	Conservar	Necesario para responder alguna (s) pregunta (s).
Facturas_Venta	Descuento	Omitir	No se necesita para responder alguna pregunta.
Facturas_Venta	TotalNeto	Omitir	No se necesita para responder alguna pregunta.
Facturas_Venta	Anulada	Omitir	No se necesita para responder alguna pregunta.
Facturas_Venta	Letra	Omitir	No se necesita para responder alguna pregunta.
Facturas_Venta	MovStock	Omitir	No se necesita para responder alguna pregunta.
Facturas_Venta	MovCtaCte	Omitir	No se necesita para responder alguna pregunta.

Tabla 2.7. Exclusión de atributos de Facturas_Venta. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Cabe resaltar que la mayoría de los atributos se excluyen, ya que no responden a las preguntas del proceso de negocio.

Informes y consultas existentes

Los informes y consultas existentes proporcionan una fuente más de información de tal manera que deben ser utilizados con los cuidados pertinentes. Ya que pueden existir informes que fueron creados para satisfacer una necesidad especifica que ya no existe y aunque se utilice puede contener algunos datos que no se utilicen.

Los datos pueden ser utilizados ya sea porque se incluyó “Por si acaso”, pero en realidad nunca se necesite o en su momento se utilizaron para crear informes, pero la circunstancias cambiaron.

Proceso de selección

La selección de los elementos de datos para su inclusión al almacén de datos no es tan simple como parece.

Para este ejemplo se consideran las siguientes preguntas de negocio:

- Se desea conocer cuántas unidades de cada producto fueron vendidas a sus clientes en un periodo determinado.

- Se desea conocer cuál fue el monto total de ventas de productos a cada cliente en un periodo determinado.

Paso 2. Añadir la dimensión de tiempo a la llave

El modelo de datos del almacén de datos es un modelo sobre el tiempo. Un modelo sobre el tiempo retrata a una empresa con una perspectiva histórica. Este tipo de modelo es el adecuado para un almacén de datos.

El segundo paso en el desarrollo del modelo de un almacén de datos, consiste en agregar el componente de tiempo o fecha a la llave de cada entidad de interés, para proporcionar una perspectiva histórica, como se muestra en la figura 2.15. Para este ejemplo se opta por tomar snapshot mensuales y por lo tanto la llave es mes año.

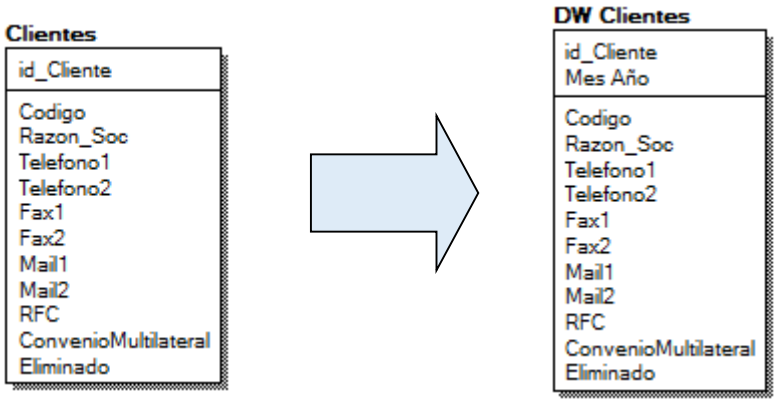

Figura 2.15. Adición del componente tiempo en la llave. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

El ejemplo empresa comercial planteado en este capítulo, no introduce la entidad de Territorio, Región y Área (en algunos ejemplos posteriores se utilizarán estas entidades), ya que el proceso de negocio Ventas no lo requiere para responder a las preguntas de negocio. Pero para que quede claro que algunas relaciones de uno a muchos o de muchos a muchos pueden cambiar. A continuación se explica el siguiente ejemplo representado en la figura 2.16:

Ejemplo:

En la parte izquierda de la figura 2.16 se representa la regla de negocio, en un momento del tiempo una región de ventas tiene muchas áreas de ventas y el área de ventas pertenece a una región. Entonces para cada una de las entidades se le añade el componente de tiempo. En un periodo de tiempo en las dos entidades no hay cambios, pero durante el cual un área de ventas se transfiere de una región de ventas a otra. La relación de muchos a muchos que se genera se resuelve con una entidad asociativa que contiene la fecha como parte de su llave.

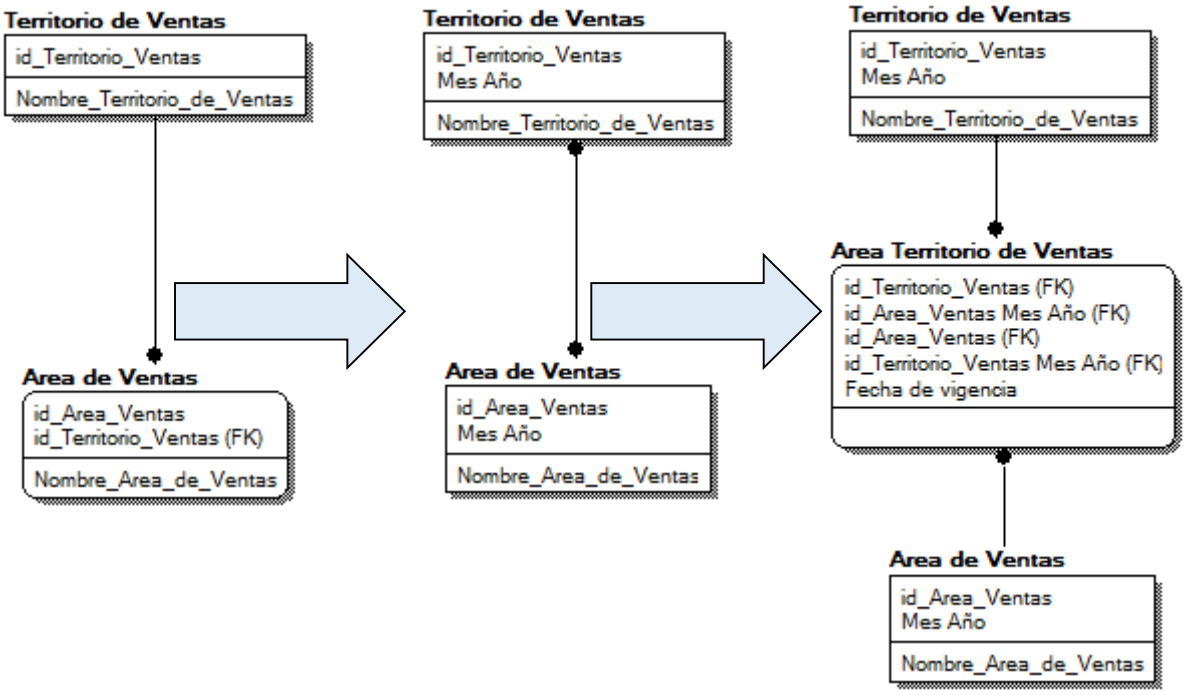

Figura 2.16. Transformación en las relaciones. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Paso 3. Añadir datos derivados

El tercer paso en el desarrollo del modelo del almacén de datos, es agregar los datos derivados. Los datos derivados son datos que resultan de una operación matemática de uno o más datos. Es necesario incluirlos al almacén por dos razones; para asegurar la consistencia y la optimización de los datos. Los elementos que se pueden incluir o excluir en la definición serían: descuentos especiales, descuentos de clientes, impuestos sobre las ventas.

Para este ejemplo, nos interesa saber las unidades vendidas de cada producto a cada cliente en un tiempo determinado y el monto total de ventas de cada producto a cada cliente en un tiempo determinado. Mediante la adición de atributos como Cantidad y MontoTotal, todas las personas involucradas en el almacén de datos van a utilizar la misma definición. Las posibles consideraciones:

Cantidad:

- ¿Se excluyen los domingos?
- ¿Se excluyen vacaciones?

MontoTotal:

- ¿Se excluye el primer día de mes?
- ¿Se excluye el último día de mes?

La creación del campo derivado no ahorra espacio en disco desde cada uno de los componentes utilizados. El uso de los datos derivados mejora el rendimiento de la entrega de los datos a costa del rendimiento de la carga.

Paso 4. Determinar el nivel de granularidad

El cuarto paso en el desarrollo del modelo del almacén de datos, es para el nivel de detalle del almacén. El nivel de granularidad es muy importante ya que de eso depende el poder responder a las preguntas de negocio. Desde el punto de vista técnico es uno de los principales determinante del tamaño del almacén. Desde el punto de vista del proyecto, el nivel de granularidad afectará la cantidad de trabajo que se tendrá en el equipo. Si por lo consiguiente tenemos un almacén de datos pequeño y el nivel de granularidad aumenta, el almacén puede llegar a ser muy grande y requerirá consideraciones adicionales.

Existen diversos factores que afectan el nivel de granularidad en un almacén de datos:

- Necesidad del negocio actual.
- Necesidad de la empresa prevista.
- Necesidad de negocio extendido.
- Necesidad de minería de datos.
- Necesidad de datos derivados.
- Granularidad de sistemas operacionales.
- Rendimiento de adquisición de datos.
- Costo de almacenamiento.
- Administración.

Este paso debe ir de la mano con el paso 1. El primer paso se vuelve importante cuando se necesita un mayor nivel de granularidad. El paso 4 es el último para asegurar que los datos satisfacen las necesidades de negocio. Los pasos restantes no son obligatorios, ya que si no se realizan, el almacén de datos debe ser capaz de satisfacer las necesidades de negocio.

Para este ejemplo tenemos la granularidad de las siguientes tablas; Clientes, Productos, Detalles_Venta y Facturas_Venta:

Clientes

Clientes contiene información acerca del cliente que realiza la compra de productos que ofrece la empresa, como se muestra en la tabla 2.8.

Clientes
id_Cliente
Codigo
Razon_Soc
Telefono1
Telefono2
Fax1
Fax2
Mail2
Mail2
RFC
ConvenioMultilateral
Eliminado

Tabla 2.8. Clientes.

Productos

Productos contiene información acerca de los productos que maneja la empresa y permite el que se puedan obtener reportes a diferentes niveles de detalles respecto a la venta del producto, como se muestra en la tabla 2.9.

Productos
id_Producto
Stock
Stock_Min
Precio
Detalle
Stock_Max
Tipo
Costo
Codigo
Imagen
Generico
Eliminado
PrecioR

Tabla 2.9. Productos.

Detalles_Venta

Detalles_Venta contiene información acerca de los detalles de las ventas realizadas, como se muestra en la tabla 2.10.

Detalles_Venta
id_Detalle
Cantidad
Precio_Fact
Total

Tabla 2.10. Detalles_Venta.

Facturas_Venta

Facturas_Venta contiene información acerca de los detalles de las facturas de ventas realizadas, como se muestra en la tabla 2.11.

Facturas_Venta
id_Factura
Nro_Sucursal
Nro_Factura
Nro_Remito
Tipo_Documento
Estado
Fecha
Descuento
TotalNeto
Anulada
Letra
MovStock
MovCtaCte

Tabla 2.11. Facturas_Venta.

Estimación bruta del almacén de datos

Por ejemplo, en la empresa comercial, si hay 10.000 filas de datos de las fechas en que fueron generadas facturas a los clientes, casi cualquier nivel de granularidad se hace. Si hay 10 millones de filas, posiblemente un bajo nivel de granularidad. Pero si hay 10 mil millones de filas, es necesario un alto nivel de granularidad.

A continuación se muestra una ruta de acceso algorítmica para calcular el espacio ocupado por un almacén de datos [14]:

- Como primer paso es identificar todas las tablas que se hayan construido.
- Como regla general, habrá una o dos tablas muy grandes y otras muy pequeñas.
- Se estima el tamaño de la fila de cada tabla (puede ser que no se conozca el tamaño exacto).

Estimación de filas/espacio para el entorno de almacenamientos de datos:

1. Para cada tabla conocida:
   ¿Qué tan grande es una fila (en bytes)?
   - Estimación mayor.
   - Estimación menor.

   Para la estimación en un año:
   ¿Cuál será el número máximo de filas posibles?
   ¿Cuál será el número mínimo de filas posibles?

   Para la estimación en cinco años:
   ¿Cuál será el número máximo de filas posibles?
   ¿Cuál será el número mínimo de filas posibles?

   Para cada llave de la tabla:
   ¿Cuál es el tamaño de la llave (en bytes)?
   Total máximo - espacio en un año = fila mayor X 1 año máximo de filas
   Total mínimo - espacio en un año = fila menor X 1 año mínimo de filas

   Además de espacio de índice.

2. Repetir paso 1 para todas las tablas conocidas.
   Tenga en cuenta que en algunos casos, el almacenamiento en disco se utiliza para crear copias de seguridad y de recuperación.

Por desgracia, es casi imposible calcular con precisión el volumen de datos en un horizonte de tiempo determinado.

Nivel dual de granularidad para este ejemplo

Cuando una empresa tiene millones de datos en el almacén de datos, tiene sentido considerar crear dos o más niveles de granularidad, veamos porqué.

Por ejemplo, en la parte izquierda de la figura 2.17 se muestran los datos operacionales, donde se encuentran los detalles de transacción de ventas a un cliente, estos datos se almacenan en el entorno operacional durante 60 días. Mientras en la parte derecha de la figura se muestran los datos en funcionamiento hasta 10 años de historia de actividades.

El nivel de archivo requiere la creación de un conjunto de datos basado en disco, que es un medio adecuado para la gestión masiva de datos. Por ejemplo, la empresa comercial puede almacenar 30 días de actividades en línea, al final de los 30 días, los datos se envían a cintas magnéticas y se ponen a disposición durante los próximos 30 días de datos de archivos, todo esto para el almacenamiento de desbordamiento en el almacén de datos. El tipo de almacenamiento en disco es barato para la empresa, pero caro en el sentido de acceso a los datos.

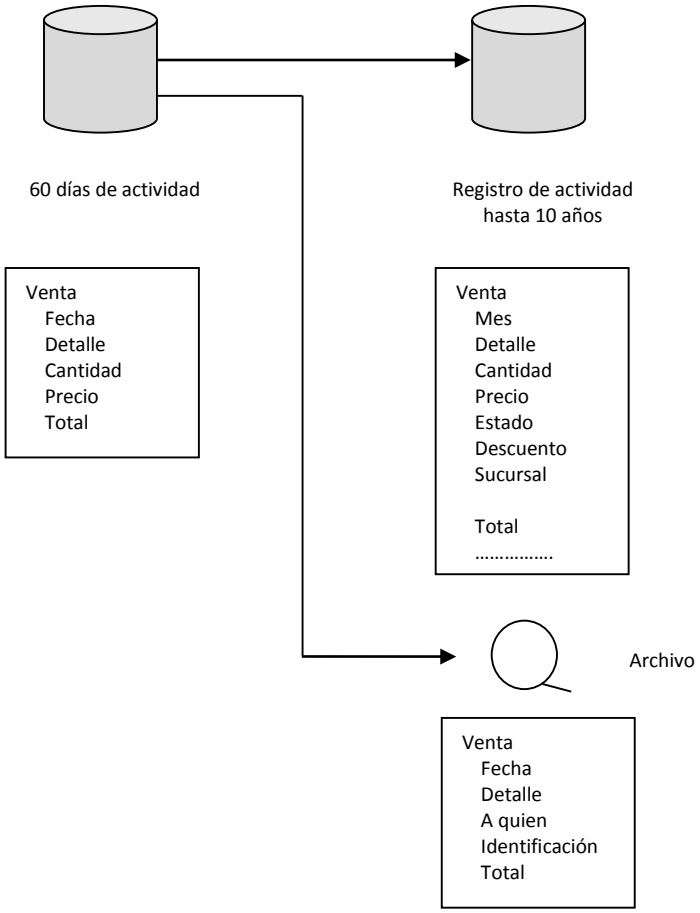

Figura 2.17. Nivel dual de granularidad para la empresa comercial. Basado en H Inmon, W. [1].

Como se mencionó anteriormente el paso 4 es el último para asegurar que los datos satisfagan las necesidades de negocio. Por lo tanto el diseño final del modelo del almacén de datos, queda como se muestra en la figura 2.18.

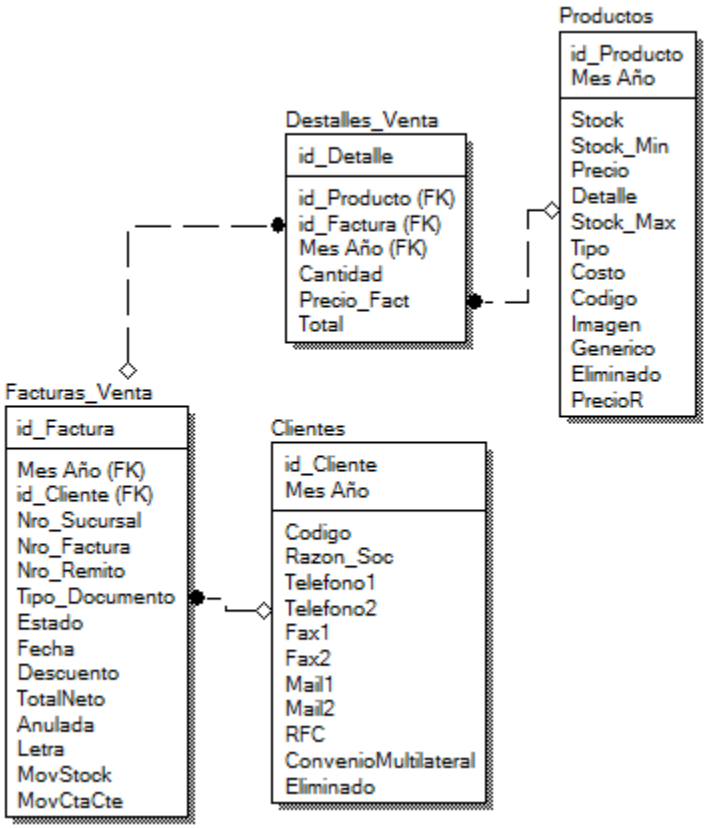

Figura 2.18. Modelo de datos del almacén de datos.

Paso 5. Resumir los datos

El quinto paso en el desarrollo del modelo del almacén de datos, es la creación de los datos resumidos. Hay que tener en cuenta que los datos resumidos no pueden ahorrar espacio en disco, posiblemente los datos que se utilicen se seguirán manteniendo. Sin embargo mejora el rendimiento del proceso de entrega de datos. El tiempo es el criterio más común, ya que los datos en el almacén se representan en algún punto del tiempo. Por ejemplo, la cantidad de productos vendidos en un día o el número de productos en inventario al final del día. Existen cinco tipos de resúmenes: acumulaciones simples, resumen rolling, archivos directos simples, archivos continuos y resumen vertical. A continuación se explican dos tipos de resúmenes:

Resumen de datos sobre periodo de tiempo

Las acumulaciones sencillas y resumen rolling se aplican a los datos que pertenecen a un periodo de tiempo. Acumulación simple representan la suma de los datos a través de unos de sus atributos, como el tiempo. Por ejemplo, resumen de ventas hechas anuales, se ofrece un resumen de todas las ventas hechas del año, si para este ejemplo a menudo se desea obtener la cantidad de unidades vendidas de cada producto a cada cliente en un tiempo determinado y el monto total de ventas de cada producto a cada cliente en un tiempo determinado, el resumen que se muestra en la tabla 2.12 podría desahogar la carga de procesamiento en el proceso de entrega de datos.

Transacción Ventas

Tiempo Mes	Producto Detalle	Ventas Cantidad	Ventas $
enero	Producto1	6	$ 3.00
enero	Producto1	7	$ 7.00
febrero	Producto1	5	$ 4.00
febrero	Producto2	5	$ 3.00
febrero	Producto1	2	$ 5.00
marzo	Producto1	6	$ 4.00
marzo	Producto2	9	$ 9.00
abril	Producto2	6	$ 2.00
abril	Producto2	1	$ 5.00
mayo	Producto1	7	$ 3.00
mayo	Producto2	3	$ 8.00
mayo	Producto2	2	$ 9.00
junio	Producto1	2	$ 7.00
junio	Producto2	2	$ 7.00
julio	Producto1	1	$ 4.00
agosto	Producto2	4	$ 4.00
agosto	Producto2	4	$ 2.00
agosto	Producto2	2	$ 3.00
agosto	Producto2	5	$ 7.00
septiembre	Producto2	7	$ 3.00
octubre	Producto2	8	$ 4.00
octubre	Producto1	8	$ 4.00
noviembre	Producto1	5	$ 5.00
diciembre	Producto1	8	$ 4.00
diciembre	Producto1	2	$ 2.00

Ventas Mensuales

Tiempo Mes	Producto Detalle	Ventas Cantidad	Ventas $
enero	Producto1	13	$ 10.00
febrero	Producto1	7	$ 9.00
febrero	Producto2	5	$ 3.00
marzo	Producto1	6	$ 4.00
marzo	Producto2	9	$ 9.00
abril	Producto2	7	$ 7.00
mayo	Producto1	7	$ 3.00
mayo	Producto2	5	$ 17.00
junio	Producto1	2	$ 7.00
junio	Producto2	2	$ 7.00
julio	Producto1	1	$ 4.00
agosto	Producto2	15	$ 16.00
septiembre	Producto2	7	$ 3.00
octubre	Producto1	8	$ 4.00
octubre	Producto2	8	$ 4.00
noviembre	Producto1	5	$ 5.00
diciembre	Producto1	10	$ 6.00

Tabla 2.12. Acumulación simple. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Por ejemplo, un resumen rolling mensual, proporciona información de ventas trimestrales como se muestra en la tabla 2.13.

Ventas Mensuales

Tiempo Mes	Producto Detalle	Ventas Cantidad	Ventas $
enero	Producto1	13	$ 10.00
febrero	Producto1	7	$ 9.00
febrero	Producto2	5	$ 3.00
marzo	Producto1	6	$ 4.00
marzo	Producto2	9	$ 9.00
abril	Producto2	7	$ 7.00
mayo	Producto1	7	$ 3.00
mayo	Producto2	5	$ 17.00
junio	Producto1	2	$ 7.00
junio	Producto2	2	$ 7.00
julio	Producto1	1	$ 4.00
agosto	Producto2	15	$ 16.00
septiembre	Producto2	7	$ 3.00
octubre	Producto1	8	$ 4.00
octubre	Producto2	8	$ 4.00
noviembre	Producto1	5	$ 5.00
diciembre	Producto1	10	$ 6.00

Resumen Rolling Trimestral

Tiempo de inicio	Tiempo de finalización	Producto Detalle	Ventas Cantidad	Ventas $
1 enero	31 marzo	Producto1	26	$ 23.00
1 enero	31 marzo	Producto2	14	$ 12.00
1 abril	30 junio	Producto1	9	$ 10.00
1 abril	30 junio	Producto2	14	$ 31.00
1 julio	30 septiembre	Producto1	1	$ 4.00
1 julio	30 septiembre	Producto2	22	$ 19.00
1 octubre	31 diciembre	Producto1	23	$ 15.00
1 octubre	31 diciembre	Producto2	8	$ 4.00

Tabla 2.13. Resumen Rolling. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Paso 6. Unir entidades

El sexto paso en el desarrollo del modelo del almacén de datos consiste en unir o combinar entidades en una sola, es decir, desnormalización de los datos. La unión de las entidades mejora el rendimiento del proceso de entrega de los datos, al reducir el número de combinaciones implica la creación de dimensiones compatible para su posterior uso en los data marts. A continuación se presentan criterios antes de unir las entidades: Las entidades comparten una llave en común, los datos de las entidades combinadas se utilizan frecuentemente.

Primera condición: Si los datos no se pueden unir a la misma llave, no se podrán unir a una entidad común, ya que en un modelo ER, los datos dentro de una entidad dependen de la llave.

Segunda condición: Esta condición es la razón por la que los datos se combinaron en primer lugar, ya que tener datos en la misma entidad, evita la unión durante la entrega de los datos al data mart.

Tercera condición: Se refiere a la condición de rendimiento y de almacenamiento de carga. Cuando los datos se combinan en una sola entidad, toda vez que haya un cambio en un atributo, se generará una nueva fila.

Las dimensiones conformadas son un tipo de dimensiones combinadas como se muestra en la figura 2.19 donde las llaves de las dimensiones Región y Territorio no se optó por llevarlas a la nueva dimensión conformada.

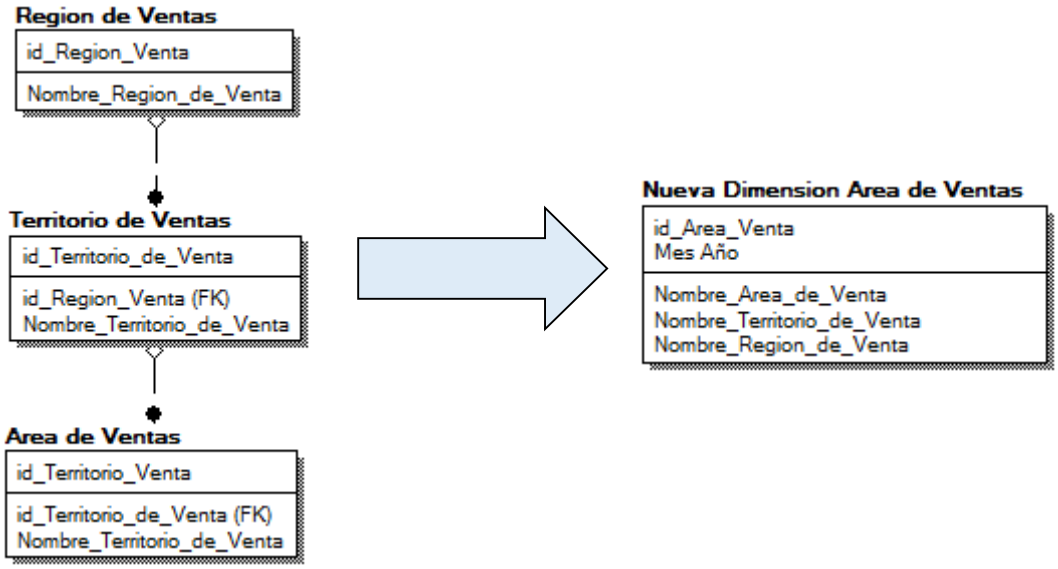

Figura 2.19. Dimensión conformada. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Paso 7. Crear arreglos

El séptimo paso en el desarrollo del modelo del almacén de datos consiste en crear arreglos. Por lo general este paso se utiliza muy poco, salvo que sea necesario mejorar la población de los data marts.

Por ejemplo, información para las cuentas por cobrar, si la información se captura en cada uno de los cinco grupos (corriente, 1-30 días de demora, 31-60 días de demora, 61-90 días de demora y más de 90 días de atraso), esta es una entidad atributiva. Esto también podría ser representado como una matriz, como se muestra en la figura 2.20. Dado que el objetivo del almacén de datos es mejorar la entrega de datos, este enfoque sólo tiene sentido si el data mart contiene una matriz.

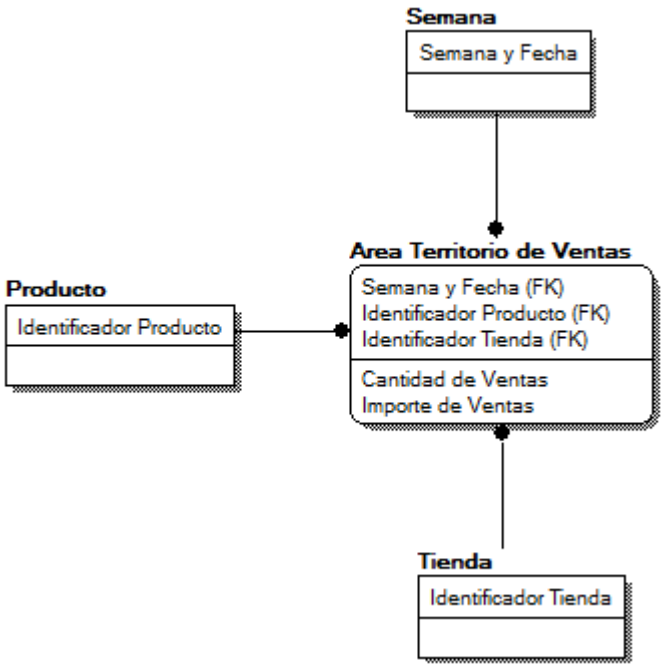

Figura 2.20. Arreglo. Basado en Imhoff, C., Galemmo, N., & Geiger, J. G. [16].

Paso 8. Separar los datos

El octavo paso en el desarrollo del modelo del almacén de datos consiste en separar los datos en base a su estabilidad y uso. Por ejemplo, si datos que se utilizan constantemente en un conjunto de tablas separadas, el acceso a los datos genera un join entre tablas que contienen los elementos necesarios y por consecuencia sufre una reducción de rendimiento en la recuperación de los datos.

"Un científico debe tomarse la libertad de plantear cualquier cuestión, de dudar de cualquier afirmación, de corregir errores."

- Robert Oppenheimer (1904-1967); Físico estadounidense.

Capítulo 3

Metodología de Ralph Kimball

Ralph Kimball es un autor reconocido a nivel mundial sobre almacenamientos de datos e inteligencia de negocios, es considerado como uno de los arquitectos originales del almacenamiento de datos que considera que un almacén de datos debe ser diseñado para ser entendible y rápido. Kimball define una metodología descendente "Bottom-up" a la hora de diseñar un almacén de datos y llama la arquitectura de un almacén de datos como arquitectura MultiDimensional (MD).

En la publicación del primer libro de Ralph Kimball *"The Data Warehouse Toolkit, segunda edición"* se hace referencia de cómo utilizar el modelo dimensional para el diseño de un almacén de datos. Sin embargo el segundo libro que publica *"The Data Warehouse Lifecycle Toolkit"* describe una metodología que utilice estas técnicas para construir data marts y almacenes de datos completos. El ciclo de vida se refiere a los pasos del desarrollo de software tradicional, pero el ciclo de vida de Kimball es una metodología de cómo diseñar, construir y desarrollar paso a paso data marts y almacenes de datos.

En este capítulo se explicarán las etapas y fases que conforman el ciclo de vida según Kimball y en concreto se desarrollará el modelo dimensional de la descripción de la empresa comercial planteada en el capítulo 3, mediante la aplicación de cuatro pasos para su diseño.

3.1. Ciclo de vida de negocio según Kimball

Antes de enfocarnos a las fases del ciclo de vida de un almacén de datos se deben contemplar cuatro principios básicos:

- **Centrarse en el negocio:** En un almacén de datos centrarse en la definición de los requerimientos de negocio constituye la fuerza más importante, ya que la definición de requerimientos contiene información valiosa para apoyar a todas las fases del proceso de desarrollo del ciclo de vida.

- **Construir una infraestructura de información adecuada:** Diseñar una base sólida donde la información sea integra, fácil de usar, donde se reflejen los requerimientos de la empresa.

- **Realizar entregas en incrementos significativos:** Crear el almacén de datos en partes entregables por plazos de tiempo ya sea de 6 a 12 meses, de esta manera se verá el avance que se va teniendo en la construcción del almacén, semejante a las metodologías de construcción de software.

- **Ofrecer la solución completa:** Proporcionar al usuario un almacén de datos sólido, fácil de usar, con la mejor calidad, además de todas las facilidades y elementos necesarios para la explotación correcta del almacén, como herramientas para hacer consultas, dar soporte, capacitación y documentación necesaria.

La construcción de un almacén de datos es compleja, requiere de mucho esfuerzo y tiempo. En la figura 3.1 se muestra el ciclo de vida dimensional de negocio según Kimball.

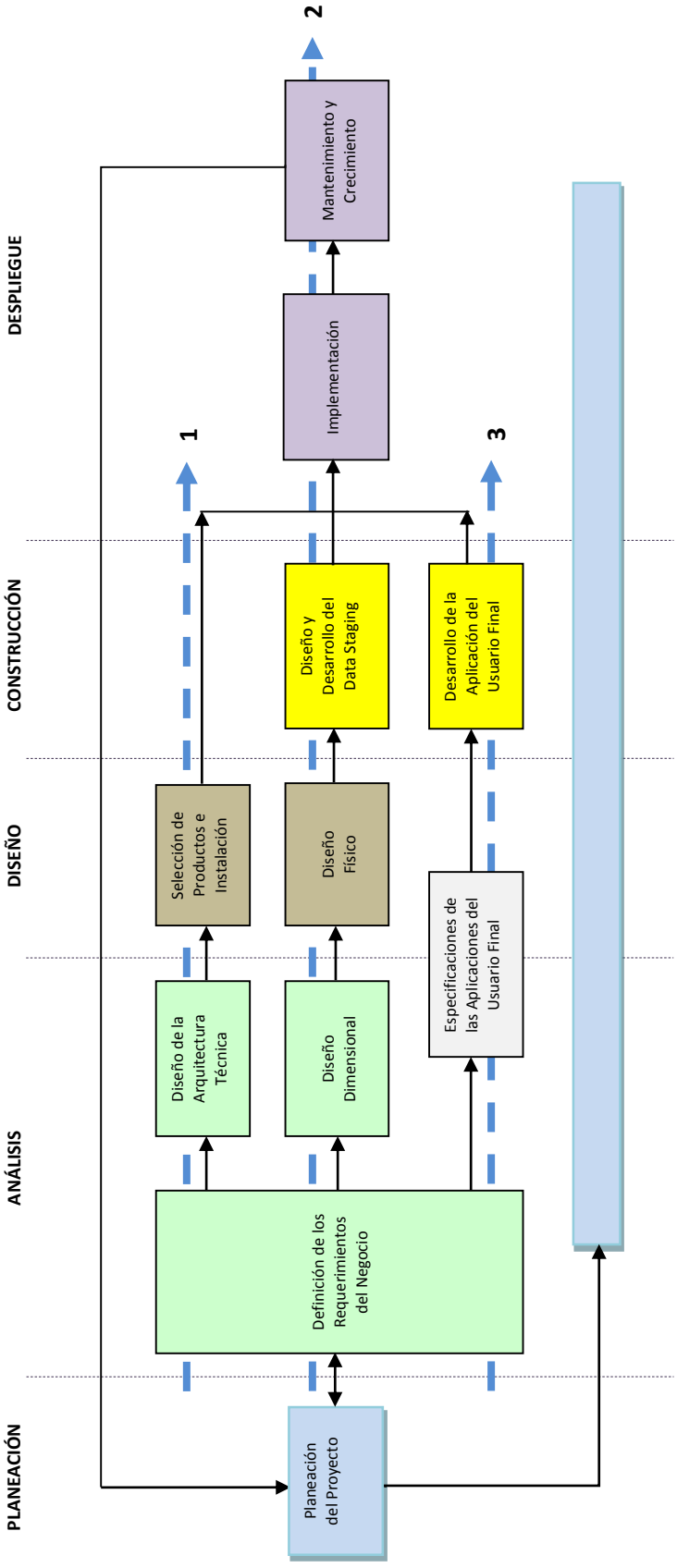

Figura 3.1. Ciclo de vida Dimensional de Negocio. Propuesta por Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. [4].

La definición de requerimientos de negocio es la parte central de la figura 3.1. Ya que es el sustento de las demás tareas, pero nótese que la planeación del proyecto tiene influencia en los requerimientos por esta razón tiene doble flecha.

Los flujos de actividades están representados por flechas punteadas de color azul y las dependencias entre tareas están indicadas por flechas continuas de color negro de manera horizontal y vertical.

Las tres flechas punteadas de color azul de manera horizontal como se observó en la figura 3.1 de arriba hacia abajo tienen sus significados con sus respectivos números:

1. **Tecnología:** Diseño de la arquitectura técnica, además indica el uso, selección e instalación de software específico. Por ejemplo: SQL Server Analysis Services[5].
2. **Datos:** Diseñar e implementar el modelo dimensional, diseño físico y se desarrolla el proceso ETL.
3. **Aplicaciones de inteligencia de negocios:** Diseñar y desarrollar aplicaciones de inteligencia de negocios que los usuarios finales van a utilizar para consulta de información y toma de decisiones.

En la parte de abajo de la figura 3.1 se muestra la actividad general de la administración del proyecto de un almacén de datos. Por lo general abarca todas las fases del ciclo de vida de Kimball.

La metodología del ciclo de vida del almacén de datos de Kimball, ayuda a simplificar la complejidad de construcción del almacén. A continuación se explica cada fase del ciclo de vida del almacén de datos:

3.2. Planeación

3.2.1. Planeación y administración del proyecto

Es la primera fase que debe realizarse, ya que en ella se tienen que plantear los siguientes objetivos:

- **Alcance:** Se definen los límites del proyecto de negocio partiendo de los requerimientos para su mejor desarrollo.

- **Justificación:** Se define la justificación del proyecto de negocio, es decir que lo respalde y hacer las evaluaciones de factibilidad.

- **Planeación del proyecto:** Se determina cuál es el propósito del proyecto del almacén de datos, se identifica el tipo de personal (usuarios, gerentes del proyecto, equipo del proyecto, desarrolladores) y se le da el seguimiento al desarrollo del plan de proyecto.

  A continuación se mencionan algunas actividades importantes en la planeación del proyecto:

  - Identificar las actividades.
  - Definir el alcance del negocio.
  - Programar las actividades a realizar.
  - Control del uso de los recursos.

[5] Ofrece funciones de procesamiento analítico en línea (OLAP) y minería de datos para aplicaciones de Business Intelligence.

- Asignar las actividades.
- Plan de manejo de riesgos.
- Costo estimado.
- Elaborar un documento que avale el plan del proyecto.
- Protocolo de entrega al cliente.

- **Administración del proyecto**

  - Se monitorean las actividades planeadas del proyecto con la situación actual, está revisión se hace con los mismos involucrados.
  - El avance del proyecto se mide por periodos específicos de tiempo.
  - Se revisa el plan de riesgo.

3.3. Análisis

3.3.1. Definición de los requerimientos de negocio

La fase de requerimientos de la implementación del almacén de datos, es una especificación precisa de las funciones que se obtendrán del almacén. Cada tarea que se realiza en cada fase del desarrollo del almacén de los datos es determinada por los requerimientos.

Cada organización es diferente y única. De tal manera que es casi imposible conocer los requerimientos desde un principio, por tal motivo se debe recolectar información mediante entrevistas o sesiones con los encargados de más alto nivel del negocio: ¿Cuál es el tipo de información que manejan para tomar decisiones?, ¿En qué consisten los trabajos que realizan?, ¿Qué tipo de información frecuentan más?, ya que proporcionará una visión más clara al momento de determinar y analizar los requerimientos para el diseño del almacén de datos.

A continuación se describen las actividades que deben llevarse a cabo durante y después de la entrevista:

- **Preparar la entrevista:** Para llevar a cabo el proceso de entrevista, se debe asignar a un equipo entrevistador el cuál debe estar conformado por una persona que entreviste y otra que tome nota.

- **Investigar la organización:** Antes de empezar con el levantamiento de requerimientos se debe hacer un estudio en base a qué tipos de documentos, reportes o información toman sus decisiones, además de conocer cuáles son las fortalezas y debilidades de la organización.

- **Seleccionar a los entrevistados:** Para recolectar la información, es necesario identificar las personas representativas de cada área o departamento de la organización.

  Por lo consiguiente identificar los directivos dentro de la organización es vital, ya que ellos son las personas que cuentan con el mayor conocimiento del negocio dentro de la organización.

- **Hacer las entrevistas:** Se empiezan con las preguntas más generales hasta las más específicas, de tal manera que se pueda recaudar información suficiente del negocio,

objetivo y visión. Además que las entrevistas se deben hacer en el departamento de tecnología para saber con cuáles recursos de desarrollo cuentan.

- **Revisar los informes y análisis existentes:** Tener en cuenta que para levantar los requerimientos también es aconsejable analizar los reportes que maneja la empresa, fuentes de datos, conocer la manera de cómo se lleva a cabo en análisis de la información y en base a qué tipos de parámetros toman las decisiones. Ya que proporciona una guía de ayuda para la elaboración del modelo dimensional de datos, que se necesitará para el diseño del almacén de datos.

3.3.2. Marco de referencia de análisis de requerimientos

El marco de referencia de Zachman[6] es una forma de ver a un sistema de información desde varias perspectivas de usuarios ver [18]. Cada usuario tiene distintas expectativas y requerimientos del sistema.

Se busca dar respuesta a las siguientes preguntas:

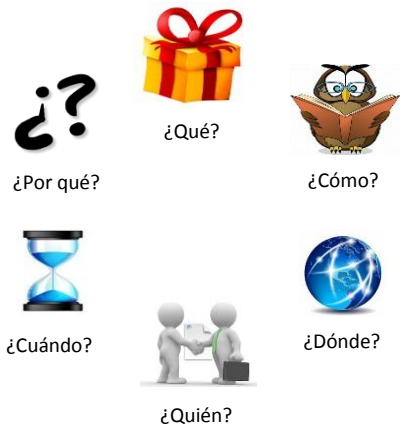

Figura 3.2. Preguntas.

Las dimensiones se caracterizan por:

Dimensión	Pregunta	Ejemplo
Entidades	¿Qué?	Cliente
Actividades	¿Cómo?	Conocer al cliente
Lugares	¿Dónde?	Cada tienda
Personas	¿Quién?	Marketing
Tiempo	¿Cuándo?	Semanal
Motivaciones	¿Por qué?	Mejorar servicio

Tabla 3.1. Dimensiones.

[6] Autor del Zachman Framework y también es ampliamente considerado como el fundador del espacio de Arquitectura Empresarial.

En la figura 3.3 se muestran los pasos del análisis de requerimientos.

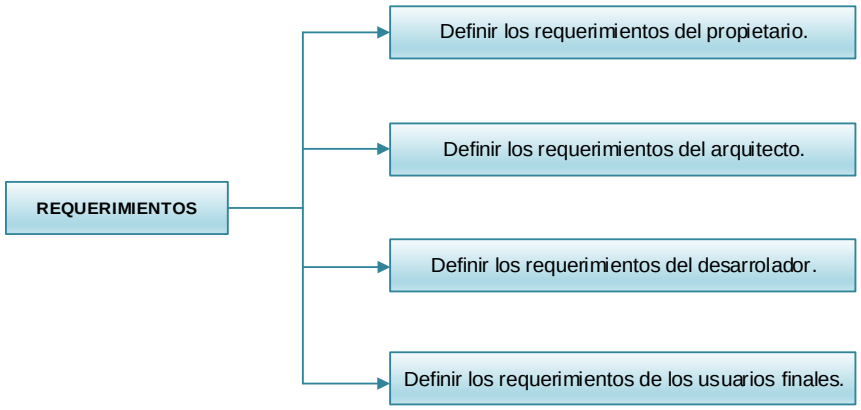

Figura 3.3. Pasos del análisis de requerimientos.

A continuación se explica cada uno de los pasos del análisis de requerimientos:

- **Definir los requerimientos del propietario:** Cuestiones que proponen los propietarios de almacenes de datos:

  1. ¿Por qué construir un almacén de datos?, ¿Qué problema empresarial abordará?
  2. ¿Cuáles son los objetivos de la empresa?
  3. ¿Quién es el cliente?, ¿Quién es el patrocinador?
  4. ¿Cuánto costará?
  5. ¿Cuándo estará listo?
  6. ¿Cuál es el impacto sobre la gente?, ¿Sobre las habilidades?, ¿Sobre la organización?
  7. ¿Cómo afectan nuestras inversiones actuales en cómputo?
  8. ¿Tenemos la capacidad para hacerlo?
  9. ¿Cuáles son los riesgos?

  El levantamiento de requerimientos debe responder a todas estas preguntas formuladas, ya que formarán criterios de aceptación desde la perspectiva del propietario.

- **Definir los requerimientos del arquitecto:** El arquitecto es la persona que se encargará de diseñar los componentes del almacén de datos que se necesiten y los que a futuro se puedan necesitar.

  La calidad de la arquitectura determina lo siguiente:

  - Características y funciones a ofrecer.
  - Plataformas necesarias para su implementación.
  - Interfaces abiertas.
  - Flexibilidad de incorporar mejoras a futuro.

  El arquitecto debe recaudar una serie de requerimientos que deben coincidir con la visión del propietario.

- **Definir los requerimientos del desarrollador:** Los requerimientos del desarrollador es menos abstracto que la del arquitecto, el desarrollador quiere que la arquitectura de datos y aplicación que formuló el arquitecto se desglosen más: interfaces, aplicaciones, computadoras, base de datos, pantallas de usuario, por lo tanto los requerimientos del desarrollador es un refinamiento de los del arquitecto**.** También los requerimientos del desarrollador se relaciona con la arquitectura de tecnología para especificar ciertos elementos como leguaje de programación, SGBD, etc.

- **Definir los requerimientos de los usuarios finales:** El usuario final ve al almacén de datos como algo abstracto, solamente tiene acceso mediante aplicaciones y herramientas de consulta. Los requerimientos de los usuarios se pueden ubicar en algunas de las siguientes categorías:

    - Flujo de trabajo.
    - Requerimiento de reportes.
    - Satisfacer la necesidad de los usuarios.
    - Fácil acceso.
    - Capacitación y soporte.
    - Permitir que los usuarios creen sus propias consultas.
    - Manejar el hardware y software, así como el SGBD del almacén de datos actual.
    - Visualización de los datos.

3.3.3. Arquitectura de Kimball

Kimball hace una analogía y dice “Oye, tengo algo de madera y algo de concreto, vamos a construir una casa” [4]. Por lo tanto no basta con tener todos los componentes para empezar a construir, hay que tener un plan antes de comenzar, esto servirá de comunicación entre los clientes y el arquitecto, además un conjunto de planos servirá para remodelar o hacer incorporaciones de la manera más adecuada en un futuro.

3.3.3.1. El valor de la arquitectura

Entonces ¿Por qué es importante la arquitectura de un almacén de datos? La arquitectura es de gran importancia ya que ayuda a la construcción del proyecto cómo hacerlo, mejor comunicación y planeación, además de que el sistema sea más flexible y mejoré la productividad.

A continuación se explican algunos aspectos importantes dentro de la organización:

- **Comunicación:** La comunicación entre distintos niveles de usuarios es vital, ya que ayuda entender la magnitud y complejidad del proyecto. Este tipo de comunicación es importante cuando se trabaja con otros grupos ya sea en un almacén de datos o en un data mart.

- **Planeación:** La arquitectura proporciona una verificación para el plan del proyecto. Si no se tiene cuidado los detalles arquitectónicos pueden terminar esparcidos y terminar por dejar el proyecto.

- **Flexibilidad y mantenimiento:** Significa que el almacén de datos sea más flexible y más fácil de mantener. Por ejemplo, se pueden utilizar herramientas y metadatos para agregar rápidamente nuevas fuentes de datos.

- **Aprendizaje:** La arquitectura juega un papel importante en el aprendizaje por parte de los miembros del equipo, ya que ayuda a ponerlos al día en los componentes, contenidos y conexiones. La alternativa es convertir a los miembros a crear sus propios mapas mentales a través de los ensayos de prueba y error.

- **Productividad y reutilización:** La arquitectura nos ayuda a elegir las herramientas necesarias para automatizar el almacén de datos en vez de construir capas de código a mano. La reutilización de procesos de almacén es más fácil para un desarrollador adoptarlo ya que no es lo mismo empezar desde cero.

3.3.3.2. Arquitectura de almacén de datos

Se hace el desglose de cómo Ralph Kimball divide la arquitectura de un almacén de datos:

- Diseño de la arquitectura técnica.
- Infraestructura.
- Metadatos.

3.3.3.3. Diseño de la arquitectura técnica

La arquitectura técnica se refiere la forma en cómo se van a representar los datos, comunicación entre procesos y la presentación de la información para los usuarios finales. Además cómo los datos sufrirán transformaciones a lo largo de su flujo.

Los almacenes de datos requieren de integración de ciertas tecnologías. Se deben tener en cuenta los siguientes factores: Requerimientos de negocio, el ambiente técnico actual y estrategias futuras planeadas, de esta manera establecer el diseño de la arquitectura técnica en el contexto de almacén de datos.

El diseño de la arquitectura técnica se divide en dos conjuntos [4]:

- Back room[7].
- Front room[8].

Cada uno con sus componentes, requerimientos y servicios como se muestran en las figuras 3.4 y 3.5 respectivamente.

[7] También conocido como cuarto de atrás.
[8] También conocido como cuarto de enfrente.

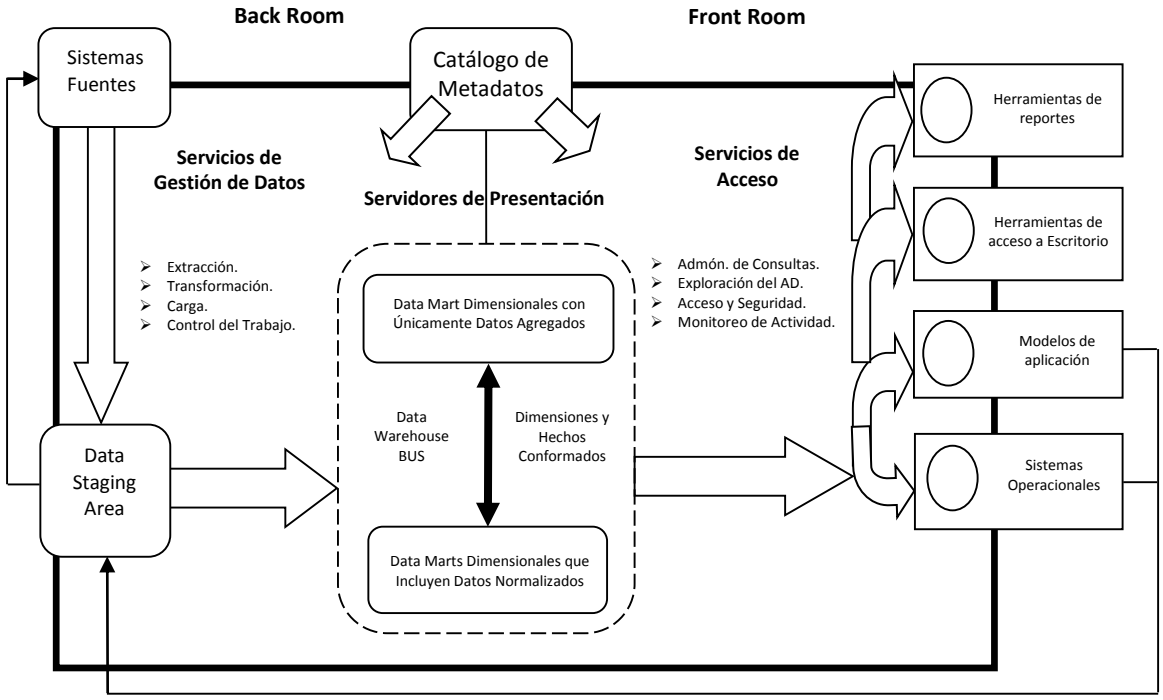

Figura 3.4. Modelo de Arquitectura Técnica de alto nivel. Propuesta por Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. [4].

El Back Room y Front Room deben estar ligados, ya que ayuda a identificar qué es lo que necesitamos hacer y cómo haremos que funcione.

A continuación se explica cada uno de los conjuntos antes mencionados:

Back Room

El Back Room es también conocido como adquisición de la información, pero lo manejaremos como Back Room de aquí en adelante.

El Back Room es donde se dá el proceso de preparación de los datos, su principal objetivo es resolver los problemas específicos de migración de datos desde el punto A al punto B con las transformaciones adecuadas, en el momento apropiado [4].

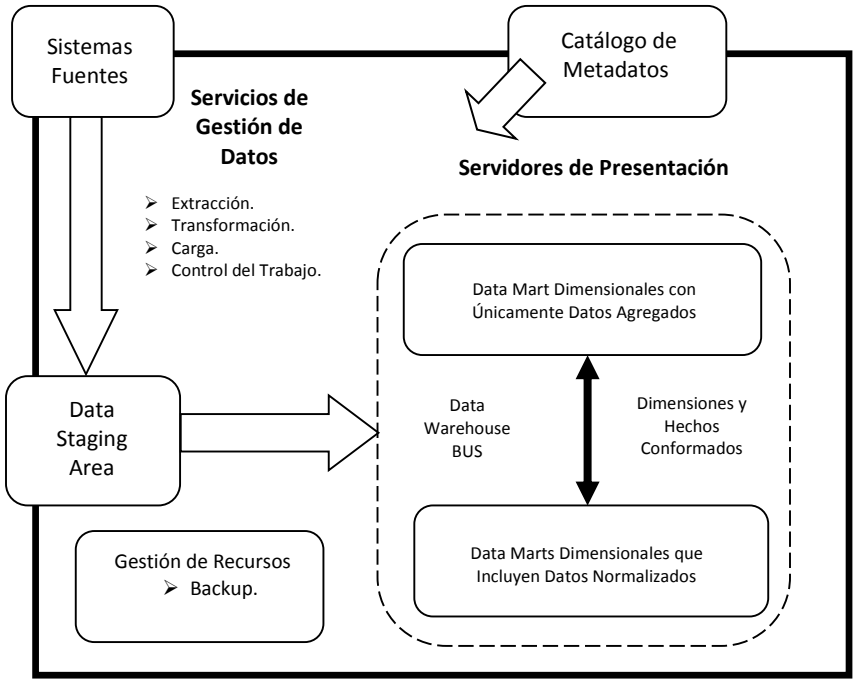

Figura 3.5. Arquitectura Técnica del Back Room. Propuesta por Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. [4].

Front Room

El Front Room es también conocido como acceso a la información, pero lo manejaremos como Front Room de aquí en adelante.

El Front Room es la parte que el usuario ve sin importarle todo lo de atrás, en otro contexto es la presentación del almacén de datos, el usuario debe ser capaz de explotar y sacarle el mejor provecho posible a las funciones.

Su objetivo principal es ocultar las complejidades de la plataforma técnica y hacer fácil lo que buscan a los usuarios.

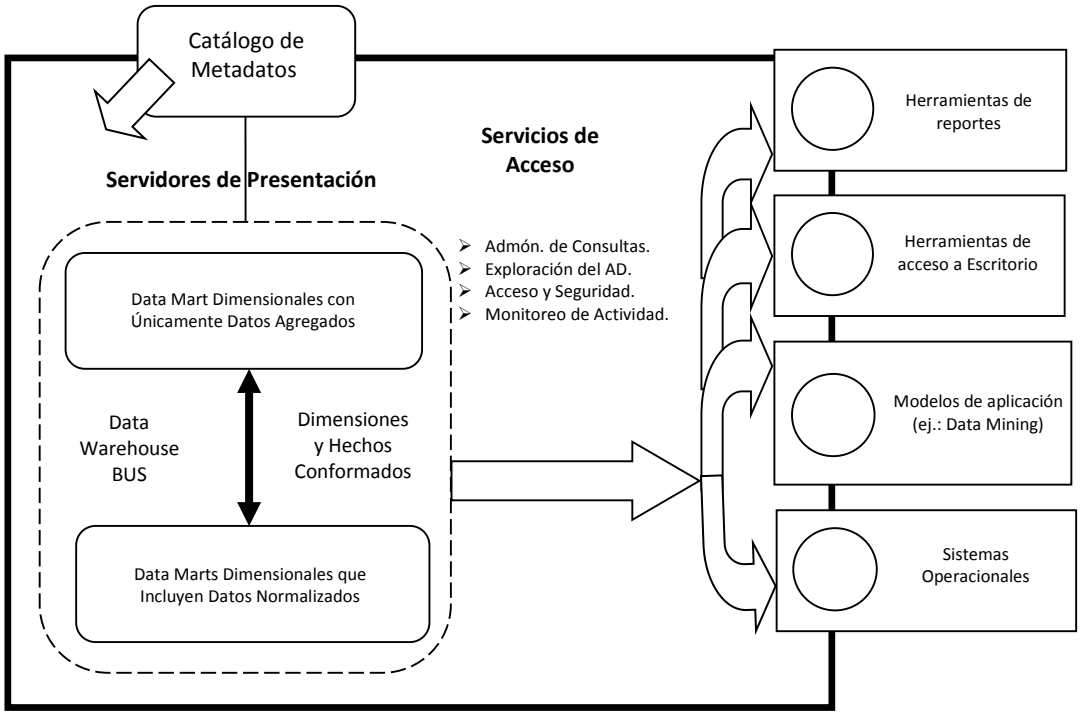

Figura 3.6. Arquitectura Técnica del Front Room. Propuesta por Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. [4].

3.3.3.3.1. Infraestructura

La infraestructura se refiere a la parte física del almacén de datos como base donde se describe el tipo de hardware y software que se utilizarán. Antes de tomar una decisión sobre el tipo de hardware y software, es necesario ver los requerimientos que se plantearon desde un principio para de allí partir.

No hay que confundir arquitectura técnica con la infraestructura, ya que está última se refiere al hardware, redes y funciones de bajo nivel como la seguridad.

3.3.3.3.2. Metadatos

Hay que tener claro que existen metadatos del Back Room donde se encuentra el proceso ETL y metadatos del Front Room hace que el uso de las herramientas de consulta sean fáciles de manejar.

3.3.4. Modelado dimensional

Unas de las principales formas de explotar los datos del almacén de datos, es con la tecnología OLAP, para lo cual es necesario realizar el análisis y diseño dimensional.

Así como los sistemas operacionales, el modelado de las bases de datos se puede hacer con el modelo relacional u orientado a objetos. Para este ejemplo el modelo dimensional debe tener un impacto que pueda responder a todas las preguntas del negocio.

El diseño de una base de datos dimensional consiste en cuatros pasos en un orden en particular como se muestra en la figura 3.7. A continuación se detalla cada paso:

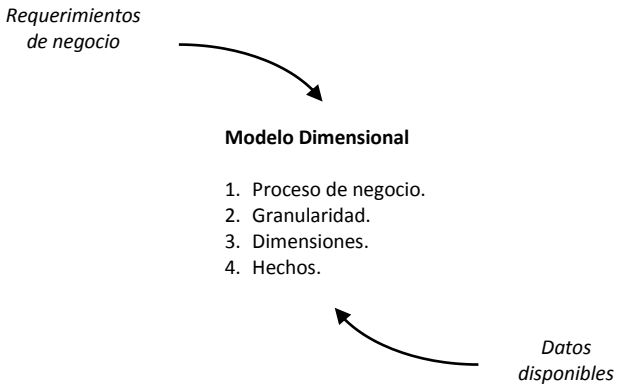

Figura 3.7. Proceso de diseño dimensional de cuatro pasos. Propuesta por Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. [4].

3.3.4.1. Pasos para el diseño del modelo dimensional

- Seleccionar el proceso de negocio.
- Declarar la granularidad.
- Identificar las dimensiones.
- Identificar los hechos.

3.3.4.1.1. Seleccionar el proceso de negocio

El proceso de negocio se llama a las actividades que se realizan dentro de la empresa con el fin de generar resultados para alcanzar los objetivos del negocio.

El primer paso, es decidir qué proceso de negocio se va a modelar, dependiendo de los requerimientos de negocio y los datos disponibles.

Algunos ejemplos de procesos de negocio:

- Contabilidad y Finanzas: Administrar cuentas, crear reportes financieros.
- Ventas y Marketing: Identificar a los clientes, ventas.
- Producción: Producir lista de materiales, verificar calidad.

Retomando el ejemplo planteado en el capítulo 2, la empresa comercial se dedica a la venta de artículos de limpieza, a mayoristas y minoristas. Se considera de tamaño mediana. Su objetivo es maximizar sus ganancias. Los principales procesos de negocio de la empresa son: Proceso de Ventas y Proceso de Compras. Para este ejemplo se utilizará sólo el proceso de Ventas.

Aunque todos los departamentos son importantes para la empresa, para este ejemplo el departamento de comercialización se considera de los más importantes, ya que se quiere entender las ventas realizadas. Así que el proceso de negocio a modelar es el de Ventas, como se muestra en la figura 3.8 y luego se procede a identificar qué es lo que interesa conocer acerca de este proceso de negocio.

Figura 3.8. Proceso de negocio Ventas.

Proceso de negocio: Ventas

El proceso de negocio Ventas permite optimizar las ganancias dentro de la empresa. La información con la que se cuenta para dar soporte a la toma de decisiones y analizarlas son las siguientes:

Métricas:

- Cuántas unidades de cada producto fueron vendidas a sus clientes en un periodo determinado.
- Monto total de ventas de productos a cada cliente en un periodo determinado.

La información almacenada en el almacén de datos debe proporcionar respuestas a las preguntas del proceso de negocio.

3.3.4.1.2. Declarar la granularidad

Ya que se ha identificado el proceso de negocio, se procede a declarar la granularidad del mismo, es decir hasta qué nivel de detalle de los datos se requiere alcanzar en el modelo dimensional. En un contexto más específico declarar la granularidad significa exactamente lo que una fila individual de la tabla de hechos representa.

Ejemplos de declaración de granularidad:

- Un una fila por cada cuenta mensual para cada mes.
- Un fila por cada replica diaria de los niveles de inventario para cada producto en un almacén.
- Una fila por cada pase de abordar escaneada en la entrada de un aeropuerto.
- Una medición de inventario realizado cada mes para cada producto en un almacén.

- Una medición de cantidad de ventas a un cliente de un producto en una tienda.

La granularidad es fundamental ya que de ella depende el volumen de datos que se almacenarán al almacén de datos.

Kimball sugiere no omitir este paso, ya que desarrolladores de almacenes de datos pasan por alto la granularidad. Ya que al omitirlo es casi imposible poder continuar con los siguientes pasos.

Granularidad de Ventas

La granularidad del proceso de negocio Ventas corresponde a un registro de venta de un producto, a un cliente en un tiempo determinado.

3.3.4.1.3. Identificar las dimensiones

Si se tiene claro la granularidad del negocio, las dimensiones se pueden identificar fácilmente. Ya que representan el qué, cómo, dónde, quién, cuándo y por qué asociado al evento. Las dimensiones a definir soportarán los requerimientos establecidos con la granularidad especificada para cada proceso de negocio. Con la elección de cada dimensión ya es posible enumerar los atributos en cada tabla de dimensión.

Antes de elegir y diseñar las tablas de dimensiones hay que tener en cuenta los siguientes puntos:

- Elegir el nombre que identifique a la tabla de dimensión.
- Añadir un campo que represente la llave primaria de la tabla de dimensión.
- Si es necesario redefinir los nombres de los campos.

Las dimensiones son el complemento de información necesaria para la presentación de los datos hacia los usuarios. Es decir, información complementaria a cada uno de los registros de la tabla de hechos. Para Kimball, las tablas de dimensiones son vitales en un almacén de datos, siempre y cuando se escojan los atributos de las dimensiones de manera correcta. Cuanto más tiempo se destine a rellenar los valores de los atributos de una columna con la mejor calidad posible, mejor será el almacén de datos.

Una vez que se ha escogido el grano, analizamos qué dimensiones tienen asociadas. Para este ejemplo las dimensiones más apropiadas son: Fecha, Clientes, Productos. La dimensión Fecha es de las primeras en definirse, ya que es casi imposible que se omita y por lo general es la primera dimensión en cargarse al almacén de datos. A continuación se detallan las dimensiones:

Dimensión Fecha

La dimensión Fecha contiene información acerca del horizonte de tiempo de las actividades de la empresa. Se define que es necesario saber: el día de la semana, el mes, el trimestre, como se muestra en la tabla 3.2.

Dimensión Fecha		
Atributo	**Tipo**	**Descripción**
id_Fecha	DATE	Llave primaria de la dimensión.
Año	CHAR (5)	Año actual de la fecha.
Trimestre	CHAR (5)	Trimestre del año (1...4).
Mes	CHAR (10)	Nombre del mes (enero… diciembre).
Dia	CHAR (10)	Nombre del día de la semana.

Tabla 3.2. Dimensión Fecha.

Dimensión Clientes

La dimensión Clientes contiene información acerca del cliente que realiza la compra de productos que ofrece la empresa, como se muestra en la tabla 3.3.

Dimensión Clientes		
Atributo	**Tipo**	**Descripción**
id_Cliente	INT (10)	Llave primaria de la dimensión.
Codigo	INT (10)	Representa el código del cliente.
Razon_Soc	VARCHAR (45)	Nombre o razón social del cliente.
Telefono1	VARCHAR (15)	Número de teléfono del cliente.
Telefono2	VARCHAR (15)	Segundo número de teléfono del cliente.
Fax1	VARCHAR (15)	Número de fax del cliente.
Fax2	VARCHAR (15)	Segundo número de fax del cliente.
Mail2	VARCHAR (15)	Dirección de correo electrónico del cliente.
Mail2	VARCHAR (15)	Segunda dirección de correo electrónico del cliente.
RFC	VARCHAR (10)	Número de RFC del cliente.
ConvenioMultilateral	VARCHAR (45)	Indica si el cliente posee o no convenio multilateral.
Eliminado	CHAR (5)	Indica si el cliente fue eliminado o no.

Tabla 3.3. Dimensión Clientes.

Dimensión Productos

La dimensión Productos contiene información acerca de los productos que maneja la empresa y permite el que se puedan obtener reportes a diferentes niveles de detalle respecto a la venta del producto, como se muestra en la tabla 3.4.

Dimensión Productos		
Atributo	**Tipo**	**Descripción**
id_Producto	INT (10)	Llave primaria de la dimensión.
id_Rubro	INT (10)	Relación con Rubros.
id_Marca	INT (10)	Relación con Marcas.
Stock	INT (10)	Stock actual del producto.
Stock_Min	INT (10)	Stock mínimo del producto.
Precio	INT (10)	Precio de venta del producto.
Detalle	VARCHAR (45)	Nombre o descripción del producto.
Stock_Max	INT (10)	Stock máximo del producto.
Tipo	VARCHAR (15)	Clasificación del producto.
Costo	INT (10)	Precio de costo del producto.
Codigo	INT (10)	Código del producto.
Imagen	VARCHAR (45)	Dibujo que representa el dibujo.
Generico	INT (5)	Indica si el producto es genérico o no.
Eliminado	INT (5)	Indica si el producto fue eliminado o no.
PrecioR	INT (10)	Precio de lista del producto.

Tabla 3.4. Dimensión Productos.

3.3.4.1.4. Identificar los hechos

El último paso en el diseño es determinar cuáles serán los hechos que aparecerán en la tabla de hechos. Los hechos son determinantes para responder a la pregunta ¿Qué se está midiendo?

Los datos obtenidos a partir del proceso de negocio Ventas incluyen la cantidad de unidades vendidas y monto total de ventas. Donde las unidades vendidas representan la sumatoria de las unidades que se han vendido de un producto en particular y el monto total de ventas representa la sumatoria del monto total que se ha vendido de cada producto y se obtiene al multiplicar las unidades vendidas, por su precio unitario. Por lo tanto se definen dos hechos, que corresponden con sus métricas como se muestra a continuación:

“Cantidad”:

- Hechos: Unidades vendidas.
- Función de sumarización: SUM.

“MontoTotal”:

- Hechos: (Unidades vendidas) * (Precio de venta).
- Función de sumarización: SUM.

Tabla de hechos Ventas

La tabla de hechos Ventas contiene los campos de las llaves de cada tabla de dimensión y los campos de los hechos, como se muestra en la tabla 3.5.

Tabla de Hechos Ventas		
Atributo	**Tipo**	**Descripción**
id_Fecha	DATE	Referencia a la dimensión Fecha.
id_Cliente	INT (10)	Referencia a la dimensión Clientes.
id_Producto	INT (10)	Referencia a la dimensión Productos.
Hechos		
Atributo	**Tipo**	**Descripción**
Cantidad	VARCHAR (15)	Unidades vendidas de cada producto.
MontoTotal	VARCHAR (15)	Monto total de ventas de cada producto.

Tabla 3.5. Tabla de hechos Ventas.

3.3.4.2. Formas de representar el modelo dimensional

La idea central de estos modelos es representar los datos de una organización a través de un cubo. Para este ejemplo los modelos finales del diseño del almacén de datos quedan como se muestran a continuación.

Modelos básicos dimensionales

Modelo Estrella: Como se explicó en el capítulo 1, un modelo estrella consta de una tabla central llamada tabla de hechos, que contienen la llave principal de las tablas de dimensiones y con sus hechos. Alrededor de la tabla de hechos se colocan las tablas de dimensiones, las cuales contienen su llave principal y sus atributos necesarios.

Modelo estrella para obtener:

- "Unidades vendidas de cada producto a cada cliente en un tiempo determinado".
- "Monto total de ventas de cada producto a cada cliente en un tiempo determinado".

Este modelo, consta de una tabla de hechos (Ventas) y tres tablas de dimensiones (Fecha, Clientes y Productos), como se muestra en la figura 3.9.

La tabla de hechos Ventas está formada por las llaves foráneas de las tablas de dimensiones y con sus respectivos hechos (Cantidad y MontoTotal).

La dimensión Fecha permite saber el tiempo en que fue vendido el producto. La dimensión Clientes permite saber toda la información referente al cliente. La dimensión Producto permite manejar información del nombre del producto y la marca.

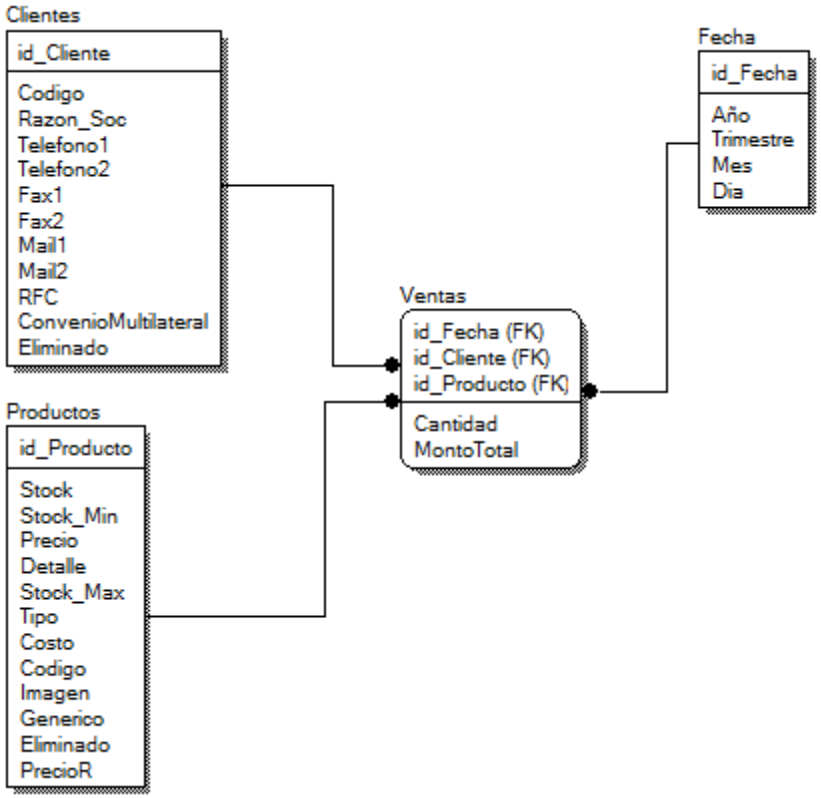

Figura 3.9. Modelo estrella.

Modelo Copo de Nieve: Como se explicó en el capítulo 1, un modelo copo de nieve al igual que el modelo estrella, también contienen tablas de hechos y de dimensiones, solo que las tablas de dimensiones están normalizadas. Las tablas de dimensiones contienen únicamente llave principal de la tabla y la llave foránea del nivel más cercano.

Modelo copo de nieve para obtener:

- "Unidades vendidas de cada producto a cada cliente en un tiempo determinado".
- "Monto total de ventas de cada producto a cada cliente en un tiempo determinado".

Este modelo, consta de una tabla de hechos y de tablas normalizadas en la que cada tabla dimensional contiene solo un nivel de detalle y la llave foránea del nivel más cercano. Como se muestra en la figura 3.10.

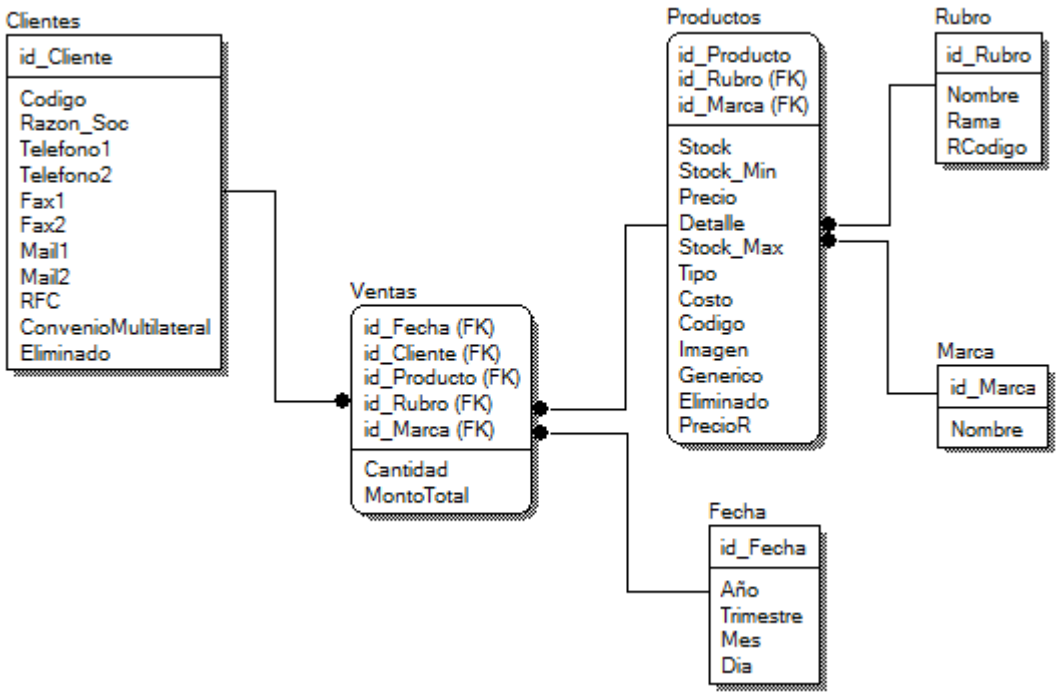

Figura 3.10. Modelo copo de nieve.

Como se pudo observar el modelo anterior, a diferencia del modelo estrella, la tabla de dimensión Productos está normalizada. Es decir la tabla dimensión Productos ahora está conformada por las dimensiones Rubro y Marca.

3.3.5. Especificaciones de las aplicaciones del usuario final

En esta fase se definen los permisos, control de acceso y roles para cada tipo de usuario de forma más estructurada y fácil de acceder al almacén de datos.

3.4. Diseño

3.4.1. Selección de productos e instalación

Sin las herramientas adecuadas el acceso y proceso de análisis, él almacén de datos podría no tener ninguna utilidad [13].

Es necesario avaluar y seleccionar los componentes necesarios utilizando el diseño de la arquitectura técnica.

Una vez que se evaluaron y seleccionaron los componentes, se instalan y realizan las pruebas necesarias en el ambiente de almacén de datos.

Ralph Kimball recomienda algunos puntos importantes [4]:

- **Comprender el proceso de compras corporativas:** Conocer cuáles son las herramientas internas con las que se cuenta (software y hardware) y así cómo los procesos de aprobación de compras de otras herramientas por parte de la organización.

- **Elaborar una matriz de evaluación del producto:** Crear una matriz de evaluación del producto, de esta manera se identificarán las funcionalidades del producto, arquitectura técnica, características del software, impacto en la infraestructura y viabilidad de los proveedores.

- **Realizar investigación de mercados:** Investigar los mercados que ofrecen productos de interés, cuál es el mejor vendedor y sus ofertas.

- **Filtrar opciones y realizar evaluaciones más detalladas:** A pesar de la gran cantidad de proveedores que existen en el mercado, sólo una parte pueden satisfacer nuestras necesidades. De esa poca porción de proveedores seleccionados hacer una evaluación detallada para poder tomar decisión de cuál elegir.

- **Manejar prototipo:** Cuando se tiene un producto seleccionado mediante la experiencia o por recomendaciones que funciona correctamente no es necesario hacer un prototipo. Si no existe una elección clara del producto lo más factible es solicitar a los proveedores de software que proporcionen una solución a un pequeño conjunto de datos de muestra.

- **Seleccionar el producto, instalación y negociación:** Antes de seleccionar el producto final es necesario pedir un periodo de prueba con el proveedor, ya que se tiene la oportunidad de evaluar si cubre con las necesidades.

3.4.2. Diseño físico

Esta fase se refiere a la estructura física que incluye tareas cómo configuración de la base de datos: Nombre de columna, tipos de datos y constraints (integridad referencial). Creación de secuencias para el proceso ETL.

Durante el proceso de diseño físico se trasladan los modelos lógicos previos a las estructuras reales. Hay que transformar las entidades en tablas con sus respetivas relaciones entre dimensiones.

También es necesario plantearse algunas preguntas en esta fase:

- ¿Cómo puede determinar que grande será el sistema de almacén de datos?
- ¿Cuáles son los factores de uso que llevarán a una configuración más grande y más compleja?
- ¿Cómo se debe configurar el sistema?
- ¿Cuánta memoria y servidores se necesitan?, ¿Qué tipo de almacenamiento y procesadores?
- ¿Cómo instalar el software en los servidores de desarrollo, prueba y producción?
- ¿Qué necesitan instalar los diferentes miembros del equipo de almacén de datos en sus estaciones de trabajo?
- ¿Cómo convertir el modelo de datos lógico en un modelo de datos físicos en la base de datos relacional?
- ¿Cómo conseguir un plan de indexación inicial?

- ¿Debe usarse la partición en las tablas relacionales?

3.5. Construcción

3.5.1. Diseño y desarrollo del Data Staging

Para Kimball [9] el proceso ETL es el fundamento de los almacenes de datos.

En esta fase se lleva a cabo el proceso ETL (ver capítulo 1):

- Extracción.
- Transformación.
- Carga.

3.5.2. Desarrollo de la aplicación del usuario final

Los usuarios acceden al almacén de datos por medio de herramientas gráficas, donde contienen información de cada área del negocio, se despliegan diferentes tipos de informes, vistas de análisis, herramientas de análisis, reportes, etc.

3.6. Despliegue

3.6.1. Implementación

En esta fase se unen todas las tareas que se realizaron en fases anteriores con el fin de que el almacén de datos sea accesible para los usuarios. Podemos decir que acá es donde se unen todas las piezas y que encajen bien.

El soporte, la capacitación con los usuarios se debe dar antes de que el almacén de datos se ponga en funcionamiento para el usuario final.

3.6.2. Mantenimiento y crecimiento

Después de la implementación de un almacén de datos, se llevan a cabo las actualizaciones constantes y correspondientes para dar un mejor funcionamiento al producto final.

El almacén de datos estará destinado a seguir creciendo, por lo tanto es recomendable establecer prioridades para poder manejar los nuevos requerimientos de los usuarios y de esa manera poder crecer.

"Excelente maestro es aquel que, enseñando poco, hace nacer en el alumno un deseo grande de aprender."

- Arturo Graf (1848-1913); Poeta italiano.

Capítulo 4

Hefesto: Metodología para la Construcción de un Almacén de Datos

Bernabeu Ricardo Dario es Ingeniero de Sistemas para el Instituto Universitario Aeronáutico (Instituto Universitario Aeronáutico) - UA. Cofundador de eGluBI (http://www.eglubi.com.ar). Se especializa en el desarrollo e implementación de soluciones OSBI[9], gestión de proyectos, análisis de requerimientos o necesidades, la implementación y configuración de soluciones de inteligencia de negocios, el diseño de los procesos de integración de datos, modelado de depósito de datos, diseño de cubos multidimensionales y modelos de negocios, el desarrollo de informes especiales, informes avanzados, análisis, cuadros de mando interactivos, etc. Es profesor, investigador y un entusiasta del software libre, sus publicaciones más notables son "Data Warehousing: Investigación y Concepto Sistematización - Hefesto: Metodología para la construcción de un DW". Coordinador de la red social Red de BI (Open http://www.redopenbi.com), también hace muchas contribuciones a diversos foros, wikis, blogs, etc. [15].

HEFESTO es una metodología para la construcción de un almacén de datos, que inicia con la recolección de requerimientos y necesidades de los usuarios finales y termina con la entrega de la información. Lo que permite la construcción de un almacén de datos de manera sencilla, ordenada e intuitiva.

Precisamente, la inteligencia de negocios, permite que el proceso de toma de decisiones esté fundamentado sobre un amplio conocimiento de sí mismo y del entorno, minimizando de esta manera el riesgo y la incertidumbre [7].

El sentido de esta metodología es que simples datos, pasen a convertirse en información importante en la toma de decisiones. Cuyo objetivo va más allá de responder preguntas de negocio, sino ver las tendencias de lo ya sucedió, está sucediendo para la construcción de modelos, los cuales puedan predecir eventos futuros.

La metodología HEFESTO se puede acoplar a cualquier ciclo de vida. Para la comprensión de cada paso en este capítulo se desarrollará el ejemplo de la empresa comercial planteada en el capítulo 2.

[9] Por sus siglas en Inglés Open Source Business Intelligence.

4.1. Descripción

Pasos que considera la metodología HEFESTO:

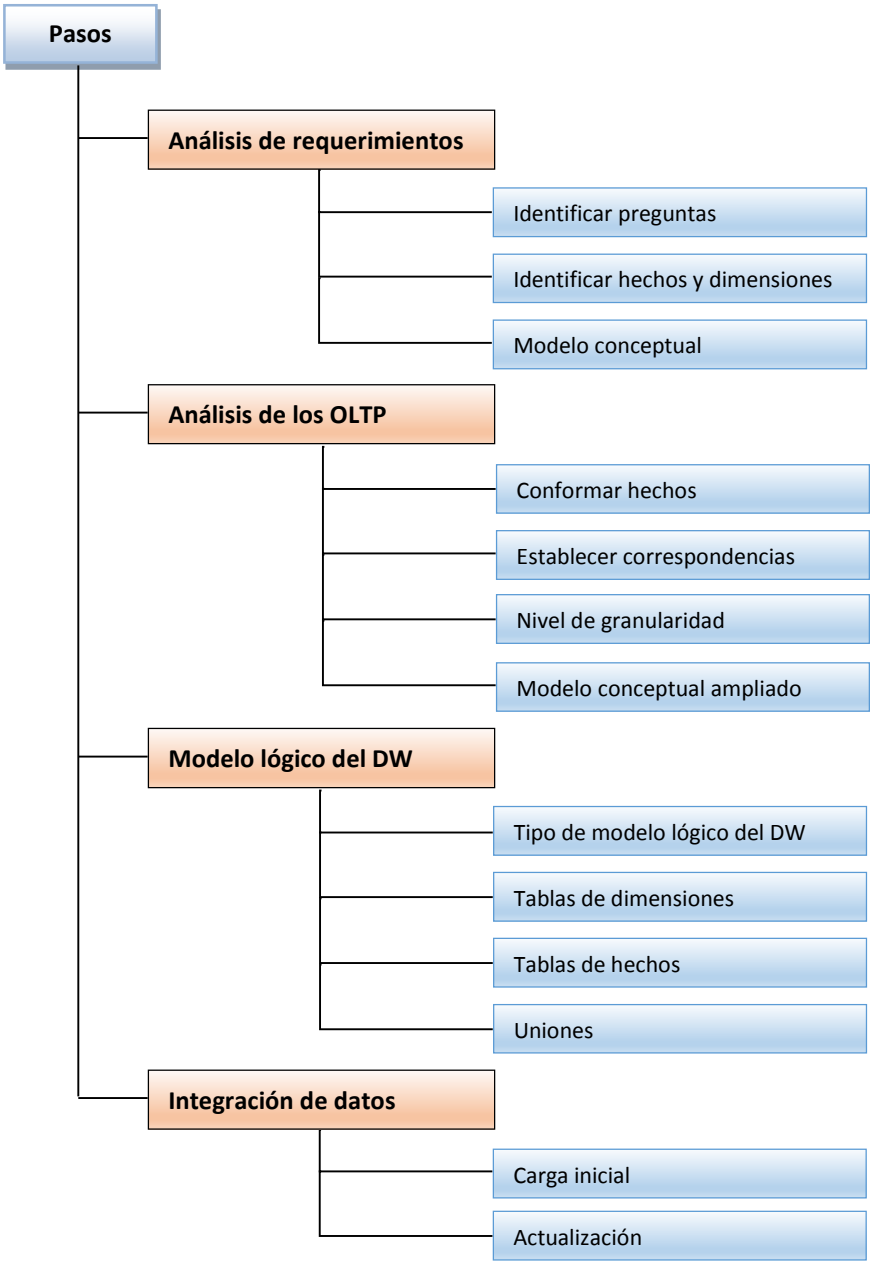

Figura 4.1. Pasos de la metodología HEFESTO. Propuesta por Dario, B. [7].

4.2. Características de esta metodología

- Cada fase es fácil de entender.
- Se basa en los requerimientos de los usuarios, por esta razón es fácil de adaptarse ante cambios inesperados del negocio.
- Reduce la resistencia al cambio, ya que se involucra a los usuarios finales en cada etapa.
- Es fácil de interpretar y analizar los modelos conceptuales y lógicos.
- Es independiente de cualquier ciclo de vida que se emplee.
- Es independiente del tipo de herramienta para su implementación.
- Es independiente de las estructuras físicas que contenga el almacén de datos.
- Cuando se termina una fase, los resultados generados dan principio de partida para la siguiente fase.
- Las mismas fases se aplican tanto para el almacén de datos como para el data mart.

4.3. Pasos y aplicación metodológica

4.3.1. Análisis de requerimientos

Primeramente se empieza con las preguntas hechas a los usuarios, esto con el fin de identificar los requerimientos de los mismos, luego estas preguntas serán analizadas para identificar cuáles serán los hechos y las dimensiones que se tomarán en cuenta al momento de construir el almacén de datos. Y finalmente se diseñará el modelo conceptual donde se podrán visualizar los primeros resultados obtenidos.

4.3.1.1. Identificar preguntas

EL primer paso es recabar las necesidades de información, esto mediante diferentes técnicas. Por ejemplo: entrevistas, cuestionarios, observaciones, etc.

El análisis de requerimientos de los diferentes usuarios, es importante ya que será la guía de desarrollo y reflejo en el almacén de datos en cuanto a sus funciones y cualidades.

El objetivo principal de esta fase es obtener e identificar las necesidades de información, que es primordial para llevar a cabo las metas y estrategias de la organización, ya que está información ayudará a desarrollar los pasos posteriores.

Se deben formular preguntas complejas sobre el negocio que se consideren relevantes, lo que permitirá analizar la información desde diferentes perspectivas.

Partiendo de la empresa comercial, el proceso de negocio elegido es Ventas, ya que se quiere entender las ventas realizadas.

A continuación, se procede a identificar qué es lo que interesa conocer acerca de este proceso y en base a ello poder tomar decisiones:

- Identificar cuáles son los hechos que representan el proceso de negocio Ventas.
- Identificar cuáles serán las dimensiones desde las cuáles se consultarán dichos hechos.

Para este proceso negocio Ventas nos interesa consultar los siguientes datos, obtenidas del ejemplo (ver capítulo 2):

- Conocer cuántas unidades de cada producto fueron vendidas a sus clientes en un periodo determinado.
- Conocer cuál fue el monto total de ventas de productos a cada cliente en un periodo determinado.

4.3.1.2. Identificar hechos y dimensiones

Se debe descubrir los hechos y dimensiones de análisis a partir de las preguntas de negocio que se establecieron anteriormente.

Los hechos por lo general son valores numéricos que representan lo que se desea analizar concretamente. Por ejemplo: saldos, promedios, cantidades, sumatorias, fórmulas, etc. Y las dimensiones son objetos mediante los cuales se quiere examinar los hechos, con el fin de responder a las preguntas planteadas. Por ejemplo: sucursales, países, productos, rubros, etc. El tiempo se considera comúnmente como una dimensión.

Ejemplo:

En la figura 4.2 se muestra el análisis de las preguntas obtenidas en el paso anterior, señalando cuáles son sus hechos y dimensiones correspondientes.

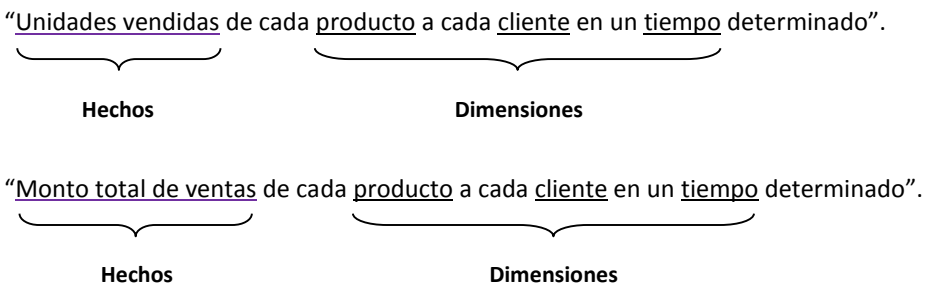

Figura 4.2. Empresa comercial, hechos y dimensiones. Propuesta por Dario, B. [7].

4.3.1.3. Modelo Conceptual

A partir de los hechos y dimensiones que se obtuvieron anteriormente, se empieza a construir el modelo conceptual, donde se observarán cuáles serán los alcances del proyecto y a su vez pueda ser entendible ante los usuarios.

La figura 4.3 muestra el modelo conceptual, en la parte izquierda se colocan las dimensiones encerradas en un rectángulo que son unidas al ovalo central, en la parte central del ovalo se tiene el nombre de la relación que es el área de estudio elegida. De dicha relación y entrelazadas con flechas se desprenden los hechos que se encuentran encerrados en un rectángulo en la parte izquierda.

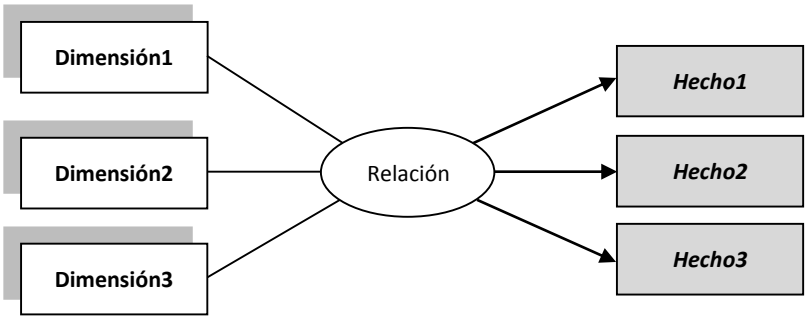

Figura 4.3. Modelo Conceptual. Propuesta por Dario, B. [7].

En la figura 4.4 se muestra el modelo conceptual del ejemplo, que se obtuvo a partir de los datos antes recaudados.

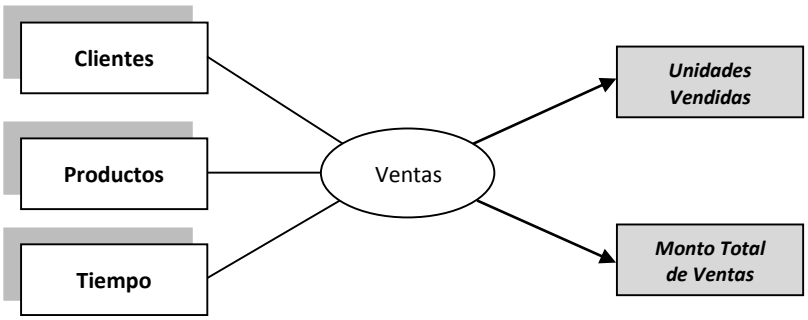

Figura 4.4. Empresa comercial, Modelo Conceptual. Propuesta por Dario, B. [7].

4.3.2. Análisis de los OLTP

En este paso, se analizan las fuentes OLTP para determinar cómo serán calculados los hechos y establecer correspondencias entre el modelo conceptual que se creó anteriormente y las fuentes de datos, posteriormente se determinarán qué campos se incluirán en cada dimensión y finalmente ampliar el modelo conceptual con la información obtenida en este paso.

4.3.2.1. Conformar hechos

En este paso, se debe explicitar cómo se calcularán los hechos, definiendo los siguientes conceptos para cada uno de ellos [7]:

- Hecho (s) que lo componen, con su respectiva fórmula de cálculo. Por ejemplo: Hecho1 + Hecho2.
- Función de resumen (sumarización) que se utilizará para su agregación. Por ejemplo: SUM, AVG, COUNT, etc.

Ejemplo:

A continuación se calculan los hechos de la siguiente manera:

- "Unidades Vendidas": representa la sumatoria de las unidades que se han vendido de un producto en particular.

    - Hechos: Unidades Vendidas.
    - Función de sumarización: SUM.

- "Monto Total de Ventas": representa la sumatoria del monto total que se ha vendido de cada producto y se obtiene al multiplicar las unidades vendidas, por su respectivo precio de venta.

    - Hechos: (Unidades Vendidas) * (Precio de Venta).
    - Función de sumarización: SUM.

4.3.2.2. Establecer correspondencias

En este paso, se examinan los OLTP disponibles que contengan la información requerida y característica, para poder identificar las correspondencias entre el modelo conceptual y las fuentes de datos. Es decir que los elementos del modelo conceptual correspondan en los OLTP.

En la figura 4.5 se muestra la correspondencia que existe entre el modelo de conceptual y el modelo relacional planteado y desarrollado en el capítulo 2.

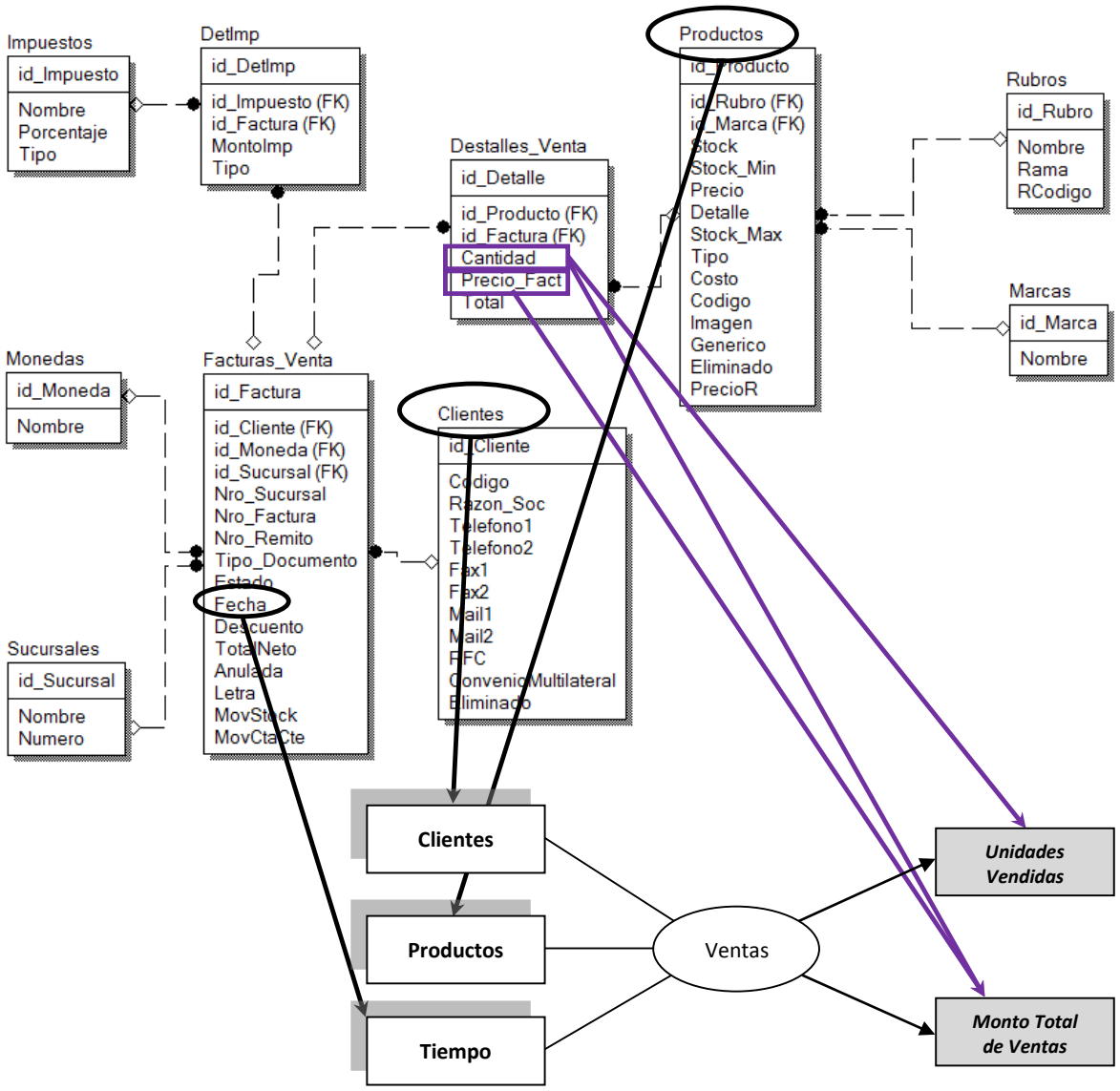

Figura 4.5. Empresa comercial, correspondencia entre el modelo conceptual en los OLTP. Propuesta por Dario, B. [7].

Las relaciones identificadas entre el modelo conceptual y el modelo relacional son las siguientes:

- La tabla “Productos” se relaciona con la dimensión “Productos”.
- La tabla “Clientes” se relaciona con la dimensión “Clientes”.
- El campo “Fecha” de la tabla “Facturas_Venta” se relaciona con la dimensión “Tiempo” (debido a que es la fecha principal en el proceso de ventas).
- El campo “Cantidad” de la tabla “Detalles_Venta” se relaciona con el hecho “Unidades Vendidas”.
- El campo “Cantidad” de la tabla “Detalles_Venta” multiplicado por el campo “Precio_Fact” de la misma tabla, se relaciona con el hecho “Monto Total de Ventas”.

4.3.2.3. Nivel de granularidad

Ya que se han establecido las relaciones con los OLTP, se deben seleccionar los campos correspondientes que contendrá cada dimensión, ya que a través de estos se examinarán y filtrarán los hechos.

Es importante conocer detalladamente qué significa cada uno de los campos y/o valores de los datos de entradas en los OLTP, por lo tanto es conveniente investigar su sentido a través de reuniones con los encargados del sistema, análisis de los datos propiamente dicho, etc.

Cuando ya se han expuesto los datos existentes frente a los usuarios explicando en sí sus significados, valores posibles y características, se debe decir cuáles son los que se consideran relevantes para consultar los hechos y cuáles no.

En la dimensión tiempo hay que definir claramente cómo se agruparán o resumirán (sumarización) los datos. Por ejemplo: día, quincena, mes, trimestre, semestre, año, etc.

Ejemplo:

De acuerdo a las correspondencias establecidas, deben analizarse los campos residentes en cada tabla a la que hace referencia, ya sea a través de diferentes métodos. Primer método: intuir los significados de cada campo. Segundo método: consultar con el encargado del sistema sobre algunos aspectos de los cuales no se comprenda su sentido o significado.

Como pudo apreciarse el Modelo Entidad Relación en el capítulo 2, los nombres de los campos son bastante explícitos y se deducen con facilidad.

Con respecto a la dimensión “Tiempo”, que es la que determinará la granularidad del almacén de datos, los datos más típicos que pueden emplearse son los siguientes:

- Año.
- Semestre.
- Cuatrimestre.
- Trimestre.
- Número de mes.
- Nombre del mes.
- Quincena.
- Semana.
- Número de día.
- Nombre del día.

Una vez que se obtiene toda la información necesaria, se consulta con los usuarios cuales son los datos que consideran de interés para analizar los hechos (ya expuestos). Los resultados obtenidos para este ejemplo son los siguientes:

- Dimensión “Clientes”:
    - “Razon_Soc” de la tabla “Clientes”. Ya que hace referencia al nombre del cliente.

- Dimensión “Productos”:

    - “Detalle” de la tabla “Productos”. Ya que hace referencia al nombre del producto.
    - “Nombre” de la tabla “Marcas”. Ya que hace referencia a la marca a la que pertenece el producto. Este campo es obtenido a través de la unión con la tabla “Productos”.

- Dimensión “Tiempo”:

    - “Año”.
    - “Trimestre”.
    - “Mes”.
    - “Dia”.

4.3.2.4. Modelo conceptual ampliado

En este paso, se amplía el modelo conceptual a partir de los resultados obtenidos anteriormente, como se muestra en la figura 4.6.

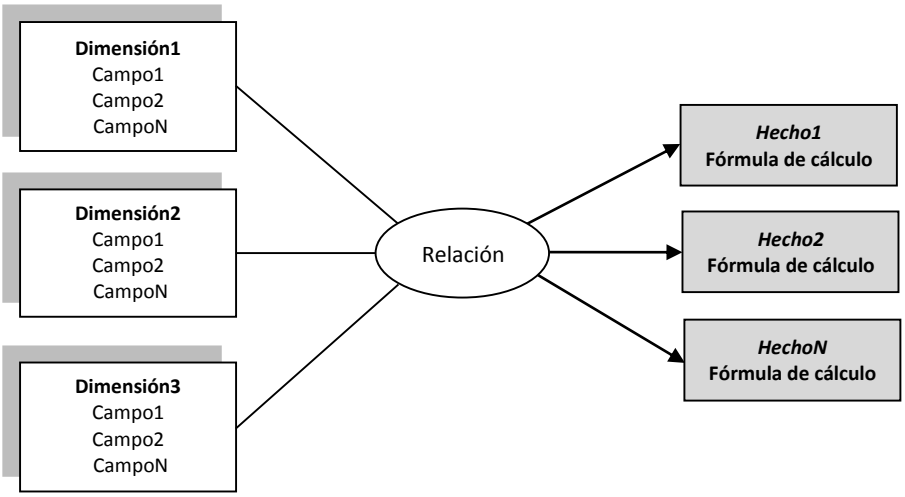

Figura 4.6. Modelo Conceptual ampliado. Propuesta por Dario, B. [7].

Para este ejemplo en la figura 4.7 se muestra el modelo conceptual ampliado; ampliado significa que a cada dimensión se le asigna su nombre que se identificaron anteriormente (Clientes, Productos, Tiempo) y con sus atributos correspondientes, la relación con el nombre de Ventas y los hechos (Unidades Vendidas, Monto Total de Ventas) y sus fórmulas de cálculos correspondientes.

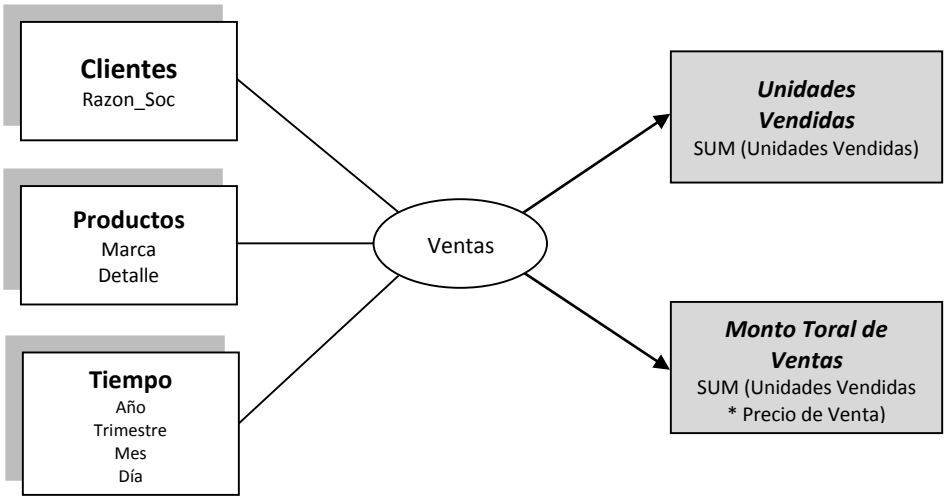

Figura 4.7. Empresa comercial, Modelo Conceptual ampliado. Propuesta por Dario, B. [7].

4.4.3. Modelo lógico del DW

Teniendo como base de partida al modelo conceptual que ya ha sido creado en las fases anteriores, se empezará a elaborar el modelo lógico de la estructura del almacén de datos.

Hay que definir el tipo de modelo que se va a utilizar y posteriormente llevar a cabo el diseño de las tablas de dimensiones y las tablas de hechos. Por último analizar las uniones correspondientes entre las tablas.

4.4.3.1. Tipo de modelo lógico del DW

En este paso, se debe seleccionar cuál va hacer el modelo que se utilice para contener la estructura del almacén de datos. Se debe seleccionar adecuadamente el tipo de modelo: estrella, copo de nieve o de constelación, ya que está decisión afectará a la elaboración del modelo lógico.

El modelo a utilizar para este ejemplo es el modelo estrella.

4.4.3.2. Tablas de dimensiones

En este paso, se diseñan las tablas de dimensiones que formarán parte del almacén de datos.

Para los tres tipos de modelos que se mencionaron anteriormente, cada dimensión definida en el modelo conceptual formará una tabla de dimensión.

En la figura 4.8 se muestra el diseño de tablas de dimensiones.

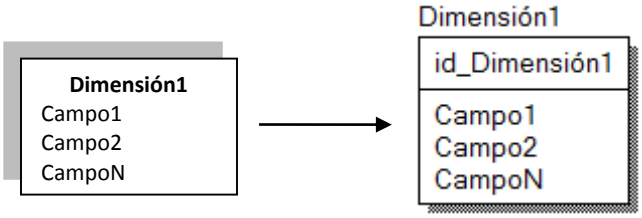

Figura 4.8. Diseño de tablas de dimensiones. Propuesta por Dario, B. [7].

Cuando existen jerarquías dentro de una tabla de dimensión en los modelos copo de nieve se debe normalizar como se muestra en la figura 4.9 donde se toma como referencia la tabla de dimensión "Geografia" y sus respectivas relaciones padre-hijo entre sus campos:

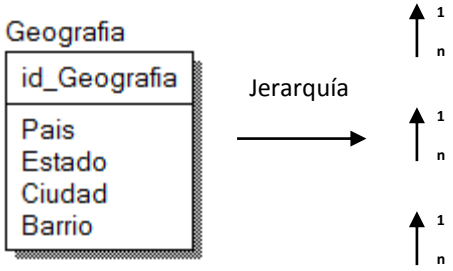

Figura 4.9. Jerarquía de "Geografia". Propuesta por Dario, B. [7].

Entonces al normalizar queda como se muestra en la figura 4.10:

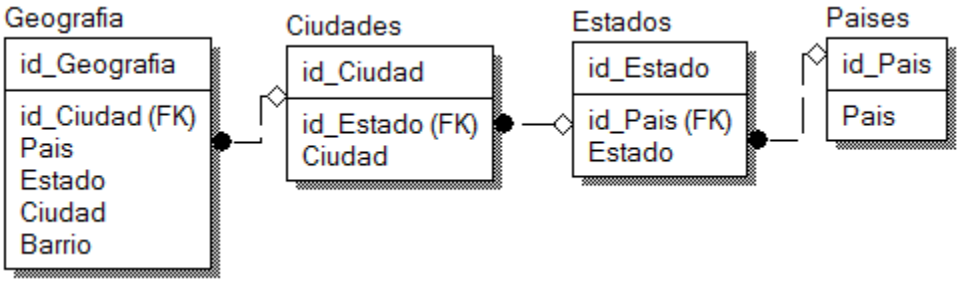

Figura 4.10. Normalización de la tabla "Geografia". Propuesta por Dario, B. [7].

Para este ejemplo se diseñan las tablas de dimensiones que a continuación se mostrarán:

- Dimensión "Clientes":
  - La tabla de dimensión tendrá el nombre "Clientes".
  - Se le agregará una llave principal con el nombre de "id_Cliente".
  - El nombre del campo será "Razon_Soc".

En la figura 4.11 se muestra el resultado gráficamente:

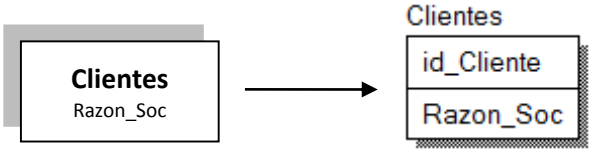

Figura 4.11. Empresa comercial, tabla de dimensión “Clientes”. Propuesta por Dario, B. [7].

- Dimensión “Productos”:

  - La tabla de dimensión tendrá el nombre “Productos”.
  - Se le agregará una llave principal con el nombre de “id_Producto”.
  - Los nombres de los campos serán “Marca” y “Detalle”.

  En la figura 4.12 se muestra el resultado gráficamente:

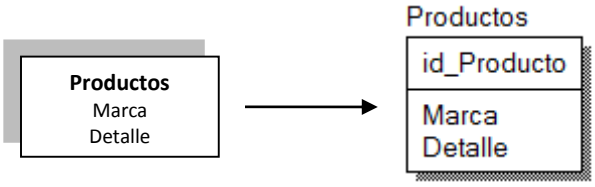

Figura 4.12. Empresa comercial, tabla de dimensión “Productos”. Propuesta por Dario, B. [7].

- Dimensión “Tiempo”:

  - La tabla de dimensión tendrá el nombre “Fecha”.
  - Se le agregará una llave principal con el nombre “id_Fecha”.
  - Los nombres de los campos serán “Año”, “Trimestre”, “Mes” y “Día”.

  En la figura 4.13 se muestra el resultado gráficamente:

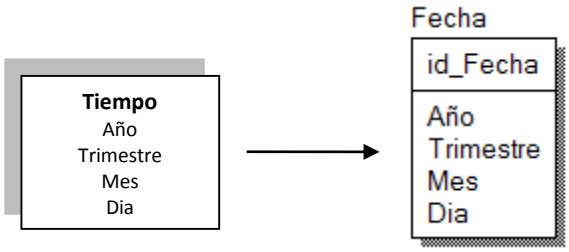

Figura 4.13. Empresa comercial, tabla de dimensión “Fecha”. Propuesta por Dario, B. [7].

4.4.3.3. Tablas de hechos

En este paso, se definen las tablas de hechos que contienen los hechos a través de los cuales se construirán los hechos de estudio.

- Para el modelo estrella y copo de nieve se deben realizar los siguientes puntos:

  - Asignar un nombre a la tabla de hechos que represente la información analizada, área de investigación, negocio, etc.

  - Definir la llave primaria, que está a su vez se compone de la combinación de las llaves primarias de cada tabla de dimensión relacionada.

  - Crear tantos campos hechos como se hayan definido en el modelo conceptual y asignar los mismos nombres que estos. Podrán ser renombrados en el momento que se requiera.

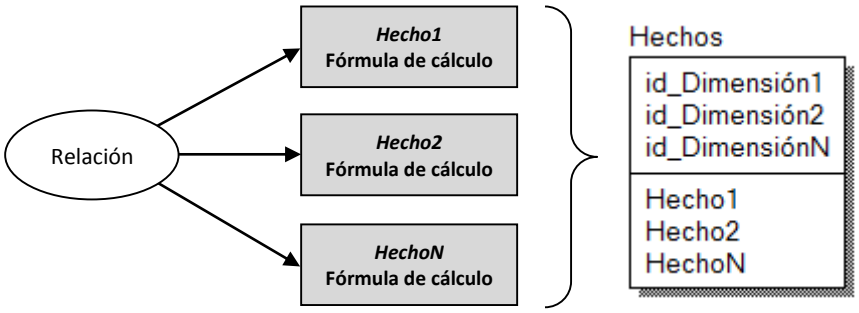

Figura 4.14. Tabla de hechos. Propuesta por Dario, B. [7].

- Para el modelo de constelación se deben realizar los siguientes puntos:

  - Las tablas de hechos se deben modelar teniendo en cuenta el análisis de las preguntas que se les fueron cuestionadas a los usuarios en pasos anteriores con sus hechos y dimensiones correspondientes.

  - Cada tabla de hechos debe tener un nombre que lo identifique, que contenga sus hechos correspondientes y la llave primaria debe estar compuesta por la combinación de las llaves de las tablas de dimensiones relacionadas.

  - Tener en cuenta los siguientes puntos al diseñar las tablas de hechos:

    - Caso 1: Si en dos o más preguntas de negocio tienen los mismos hechos pero con diferentes dimensiones de análisis, se deben crear tantas tablas de hechos como preguntas cumplan está condición. Por ejemplo:

      "Analizar el Hecho1 por **Dimensión1** y por **Dimensión2**".
      "Analizar el Hecho1 por **Dimensión2** y por **Dimensión3**".

      Entonces tenemos la figura 4.15:

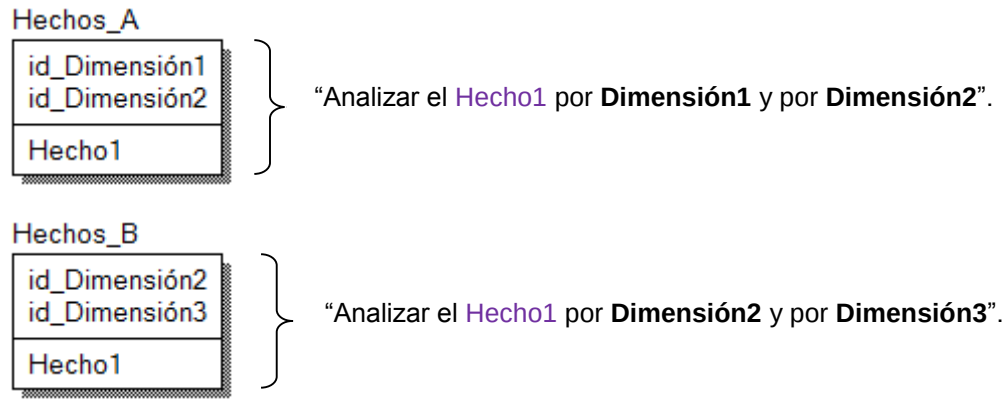

Figura 4.15. Caso 1, diseño de las tablas de hechos. Propuesta por Dario, B. [7].

- Caso 2: Si en dos o más preguntas de negocio figuran diferentes hechos con diferentes dimensiones de análisis, se deben crear tantas tablas de hechos como preguntas cumplan está condición. Por ejemplo:

  "Analizar el Hecho1 por **Dimensión1** y por **Dimensión2**".
  "Analizar el Hecho2 por **Dimensión1** y por **Dimensión3**".

Entonces tenemos la figura 4.16:

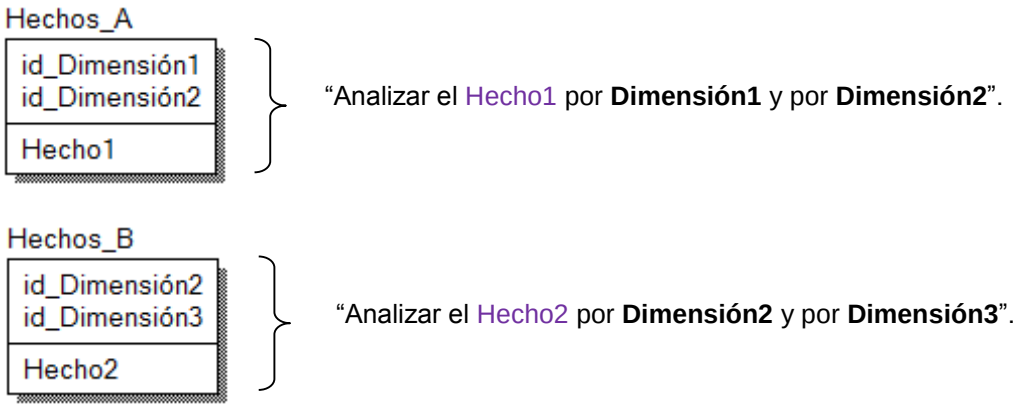

Figura 4.16. Caso 2, diseño de tablas de hechos. Propuesta por Dario, B. [7].

- Caso 3: Si el conjunto de preguntas de negocio cumplen con las condiciones de los dos puntos anteriores, se deben unir aquellas preguntas que posean diferentes hechos pero iguales dimensiones de análisis, para luego reanudar el estudio de las preguntas. Por ejemplo:

  "Analizar el Hecho1 por **Dimensión1** y por **Dimensión2**".
  "Analizar el Hecho2 por **Dimensión1** y por **Dimensión2**".

Se unen, quedando como se muestra a continuación:

"Analizar el Hecho1 y el Hecho2 por **Dimensión1** y por **Dimensión2**".

Ejemplo:

A continuación, se crea la tabla de hechos:

- La tabla de hechos tiene el nombre "Ventas".
- Su llave principal es la combinación de las llaves principales de las tablas de dimensiones antes definidas: "id_Cliente", "id_Producto" e "id_Fecha".
- Se crean dos hechos y son renombrados, "Unidades Vendidas" por "Cantidad" y "Monto Total de Ventas" por "MontoTotal".

Entonces tenemos la figura 4.17 para poder ver su representación:

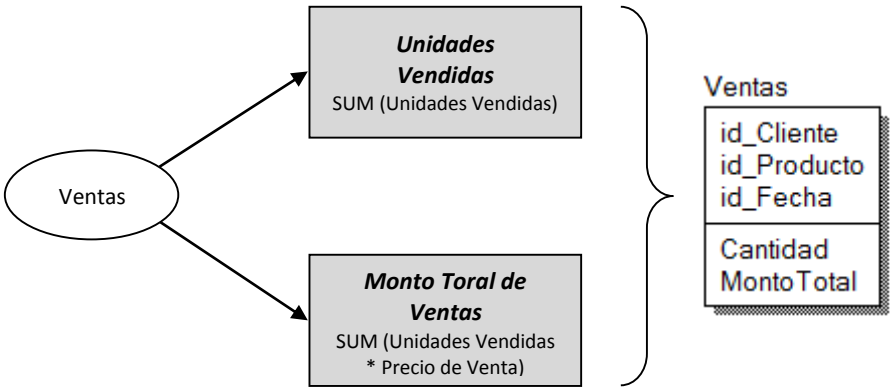

Figura 4.17. Empresa comercial, diseño de la tabla de hechos. Propuesta por Dario, B. [7].

4.4.3.4. Uniones

Para los tres modelos antes mencionados (estrella, copo de nieve y de constelación), se hacen las uniones correspondientes entre las tablas de dimensiones y sus tablas de hechos.

Para este ejemplo se hacen las uniones correspondientes, quedando el modelo estrella como modelo final del diseño del almacén de datos, como se muestra en la figura 4.18.

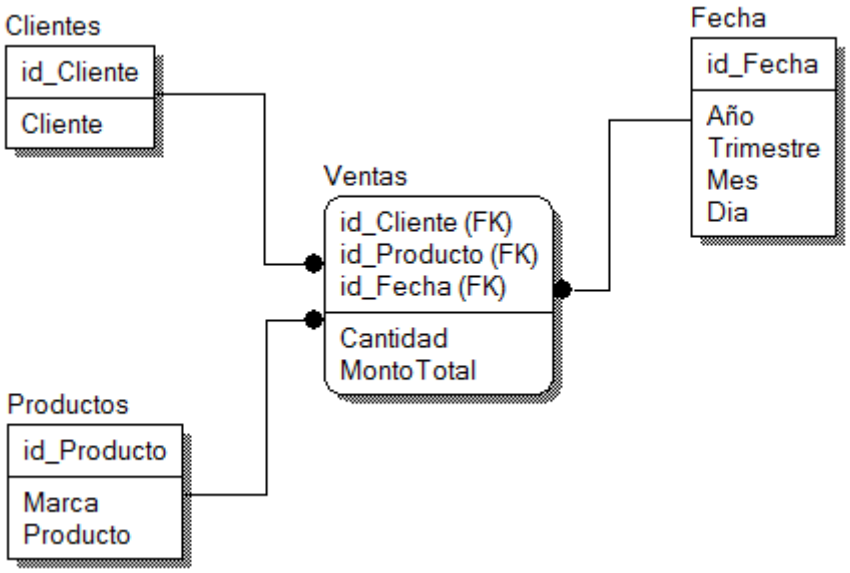

Figura 4.18. Empresa comercial, uniones. Propuesta por Dario, B. [7].

4.4.4. Integración de datos

Cuando se ha terminado de construir el modelo de lógico, se lleva a cabo el proceso ETL, posteriormente se definen las reglas y políticas para sus respectivas actualizaciones.

4.4.4.1. Carga Inicial

La carga inicial se inicia poblando el modelo de datos que se construyó anteriormente, donde se tiene que llevar una serie de tareas que conlleva el proceso ETL.

Las tareas son complejas y requiere de mucho tiempo, pero existen diferentes herramientas que facilitan realizar este tipo de trabajo.

Hay que tener en cuenta que cuando se diseña un modelo de constelación muchas de las tablas serán compartidas con diferentes tablas de hechos, por lo tanto puede darse el caso de que ciertas restricciones aplicadas a una tabla de dimensión en particular para analizar una tabla de hechos, se puede contraponer con otras restricciones o condiciones de análisis de otras tablas de hechos.

Primeramente se cargan las tablas de dimensiones y posteriormente las tablas de hechos. Cuando se esté en un modelo copo de nieve cada vez que existan jerarquías de dimensiones, se comenzarán de lo más general al nivel más detallado.

4.4.4.2. Actualización

Ya que se ha cargado el almacén de datos por completo, se deben establecer las políticas y estrategias de actualización de los datos.

Llevar a cabo las siguientes acciones:

- Tareas de limpieza de datos, calidad de los datos y el proceso ETL.
- Especificar de forma general y detallada las acciones que deberá realizar cada software.

“Intenta no volverte un hombre de éxito, sino volverte un hombre de valor.”

- Albert Einstein (1879-1955); Físico alemán.

Capítulo 5

Estudio comparativo de técnicas para el diseño y construcción de almacenes de datos

Inmon, Kimball y Bernabeu, han creado almacenes de datos para la gestión de información para la toma de decisiones. Cada uno de ellos con diferentes filosofías, técnicas de diseño y estrategias de implementación.

Está investigación tiene como objetivo comparar los principales enfoques acerca del diseño de un almacén de datos centrándose en aspectos básicos de cada enfoque. En este capítulo comenzaremos con aspectos importantes contemplados por cada autor como objetivo de un almacén de datos, metodología (diferencias, ventajas y desventajas), arquitectura (alcances, diferencias, similitudes, ventajas y desventajas), modelado de datos, filosofía (diferencias) y consideraciones específicas para elegir alguna metodología.

5.1. Objetivo de un almacén de datos para cada autor

Está claro que para los tres autores antes estudiados, el objetivo de un almacén de datos es que la información almacenada sea correcta y fiable, además de asegurar que las personas encargadas de tomar decisiones puedan acceder fácilmente a él. Sin embargo existen objetivos específicos de cada autor como se muestra en la tabla 5.1.

Autor	Objetivo
Inmon	Entregar una solución técnica basada en métodos y tecnologías de bases de datos.
Kimball	Entregar una solución fácil para que los usuarios finales puedan consultar directamente los datos y obtener tiempos de respuesta razonables.
Bernabeu	Entregar una primera implementación que satisfaga parte de las necesidades, para demostrar las ventajas del almacén de datos y motivar a los usuarios.

Tabla 5.1. Objetivos.

5.2. Metodología de diseño para cada autor

5.2.1. Metodología de William H. Inmon

Metodología o enfoque de Inmon también conocido como “De arriba abajo”. En esta metodología los datos son extraídos de los sistemas operacionales por el proceso ETL y cargados en el Data Staging Area, donde son validados y consolidados al almacén de datos, los datos están normalizados en una estructura relacional. Una vez realizado este proceso los data marts obtienen la información del almacén de datos y los usuarios finales puedan acceder al almacén por diferentes aplicaciones o herramientas. En la figura 5.1 se muestra el enfoque de arriba abajo de Inmon.

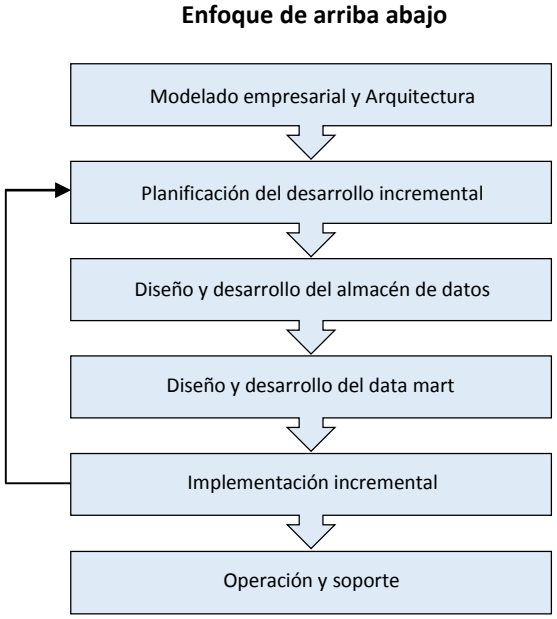

Figura 5.1. Metodología de Inmon. Propuesta por The Data Warehousing Institute. (s.f.). TDWI Data Warehousing Concepts and Principles [45].

5.2.2. Metodología de Ralph Kimball

En la metodología o enfoque de Kimball también conocido como "De abajo arriba". El almacén de datos es la unión de los diferentes data marts, que están en una estructura dimensional de una forma en común a través de la "Arquitectura bus". Esta característica la hace más flexible y sencilla de implementar, pues se puede construir un data mart como primer elemento del sistema de análisis y luego ir añadiendo otros que comparten las dimensiones ya definidas o incluyen otras nuevas. En este sistema, el proceso ETL extrae los datos de los sistemas operacionales y los procesan igualmente en el Data Staging Area, realizando posteriormente el llenado de cada uno de los data marts de una forma individual, aunque siempre respetando la estandarización de las dimensiones y finalmente los usuarios finales puedan acceder al almacén por diferentes aplicaciones y herramientas. En la figura 5.2 se muestra el enfoque de abajo arriba de Kimball.

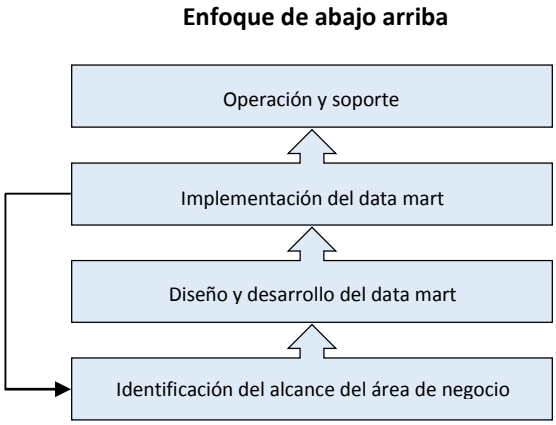

Figura 5.2. Metodología de Kimball. Propuesta por The Data Warehousing Institute. (s.f.). TDWI Data Warehousing Concepts and Principles [45].

5.2.3. HEFESTO: Metodología para la construcción de un Almacén de Datos

Esta metodología facilita la construcción de un almacén de datos y la comprensión de cada paso a realizar, sin caer en lo tedioso que provoca el no entender los pasos del mismo. Esto para motivar a los usuarios y demostrar las ventajas del almacén y que satisfaga las necesidades requeridas y finalmente los usuarios finales puedan acceder al almacén de datos por diferentes aplicaciones y herramientas.

Un almacén de datos comienza siendo un data mart, esto para minimizar riesgos. Pero una vez que se han implementado exitosamente, su alcance irá aumentando y es por eso que de acuerdo a las operaciones que se requieran desarrollar, los data marts pueden adoptar cualquiera de los siguientes enfoques: "De arriba abajo" o "De abajo arriba".

La metodología HEFESTO se resume a través de la figura 5.3:

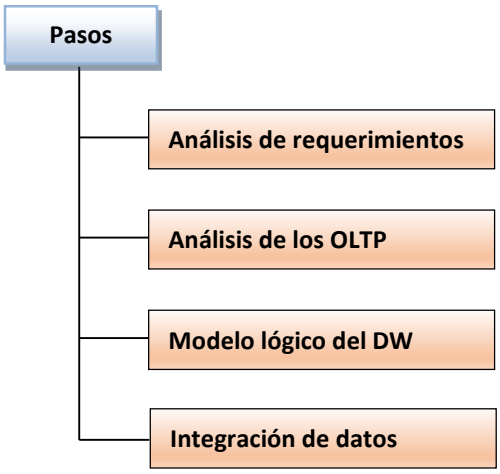

Figura 5.3. Metodología HEFESTO. Propuesta por The Data Warehousing Institute. (s.f.). TDWI Data Warehousing Concepts and Principles [45].

5.3. Diferencias y similitudes entre metodologías

En la tabla 5.2 se muestran cuáles son las diferencias y similitudes entre metodologías.

Diferencias	Inmon	Kimball	Bernabeu
Enfoque general	Enfoque "De arriba abajo": Un almacén de datos es la fuente de información para los data marts.	Enfoque "De abajo arriba": Un almacén de datos es la concentración de todos los data marts en una arquitectura bus.	"De arriba abajo" o "De abajo arriba".
Complejidad de la metodología	Bastante complejo: Tiempo de desarrollo de 4 a 6 meses.	Moderado: Un data mart se puede construir en 90 días y los posteriores de 60 a 90 días.	Moderado: Se busca que las fases de desarrollo sean cortas y que no lleven demasiado tiempo.
Comparación con metodologías de desarrollos establecidas	Derivado de la metodología en espiral.	Proceso de cuatro pasos para el modelado dimensional.	Puede adaptarse muy bien a cualquier ciclo de vida de desarrollo de software.
Organización de los requerimientos de toma de decisiones	Estratégico.	Táctico.	Estratégico o Táctico.
Personal y habilidades	Equipo grande de especialistas.	Equipos pequeños de generalistas.	Equipos de generalistas.
Duración del proyecto	Muy alta.	Alta.	Se busca que la duración del proyecto sea corta.
Tiempo de entrega	Los requerimientos de la organización permiten un mayor tiempo de puesta en marcha del almacén de datos.	Necesidad de que el primer almacén de datos sea una aplicación urgente.	Se busca rápida entrega.
Costo de implementar	Altos costos de puesta en marcha, con menores costos posteriores de desarrollo de proyectos.	Bajos costos iniciales, cada proyecto posterior de desarrollo cuesta aproximadamente lo mismo.	Bajo costo.
Similitudes			

	Las tres metodologías satisfacen las necesidades de los usuarios finales de depender de la fiabilidad de los datos contenidos en el almacén de datos. Así como de acceder a la información a través de consultas fáciles y comprensibles.
	El principal componente de las metodologías es el Área de presentación de datos o Data Staging Area; se lleva a cabo todo lo que conlleva el proceso ETL.
	Los tres autores coinciden que los data marts independientes no satisfacen las necesidades de información precisa, oportuna y de fácil acceso para los usuarios.

Tabla 5.2. Diferencias y similitudes entre metodologías. Basado en Breslin, M. [12].

5.4. Ventajas y desventajas de cada metodología

En la tabla 5.3 se muestran las ventajas y desventajas de cada metodología.

	Inmon	Kimball	Bernabeu
Ventajas	• El almacén de datos está en 3NF, por lo tanto es más fácil de construir modelos de minería de datos. • Modelo estructurado y fácil de mantener. • Mejor definición del proyecto, ya que abarca a toda la empresa. • Procesos de integración. • Metodología más dirigida a definir los componentes del back-end.	• No requieren altos costos iniciales. • El valor del negocio puede ser recuperado lo antes posible (creando los primeros data marts). • Infraestructura más adaptable a los requerimientos de un DSS[10]. • Tienen un plazo de ejecución más rápido. • Mejor rendimiento de tiempo de respuestas a las consultas. • Enfoque centrado en las necesidades de información. • Garantiza una mayor participación de los usuarios.	• Se puede elegir el enfoque que mayor se adapte a las necesidades de la empresa u organización (enfoque de arriba abajo o de abajo arriba). • Participación de cualquier tipo de usuario. • Se busca una rápida entrega del proyecto final.
Desventajas	• El costo de implementación del almacén de datos es alto. • El tiempo desde el inicio y hasta el final del proyecto es muy tardado. • El enfoque de arriba abajo representa la construcción de un proyecto muy grande con un alcance muy amplio. • Impide la participación del usuario final en el proyecto.	• Su modelo de datos es inflexible. • Difícil de mantener debido a la redundancia de los datos. • Cambios en los sistemas operacionales implican cambios en los procedimientos dedicados a los diferentes modelos de diseños.	• La metodología HEFESTO está en constante cambio. • No existe mucha documentación sobre esta metodología. • Si no se define claramente el enfoque a utilizar (de arriba abajo o de abajo arriba) el proyecto puede ser inestable.

Tabla 5.3. Diferencias y similitudes entre metodologías. Basado en Breslin, M. [12].

[10] Por sus siglas en Inglés Decision Support System.

5.5. Arquitectura contemplada por cada autor

5.5.1. Arquitectura de Fábrica de Información Corporativa

Como se mencionó en el capítulo 2 la Fábrica de Información Corporativa se basa en la creación de un repositorio de datos corporativos como fuente de información consolidada, persistente, histórica y de calidad. Además describe cómo los componentes trabajan juntos, cómo fluyen los datos de un elemento a otro y da una visión del producto final. Los componentes de esta arquitectura están divididos en dos grupos de componentes (ver capítulo 2) con sus respectivos procesos como se muestra en la figura 5.4.

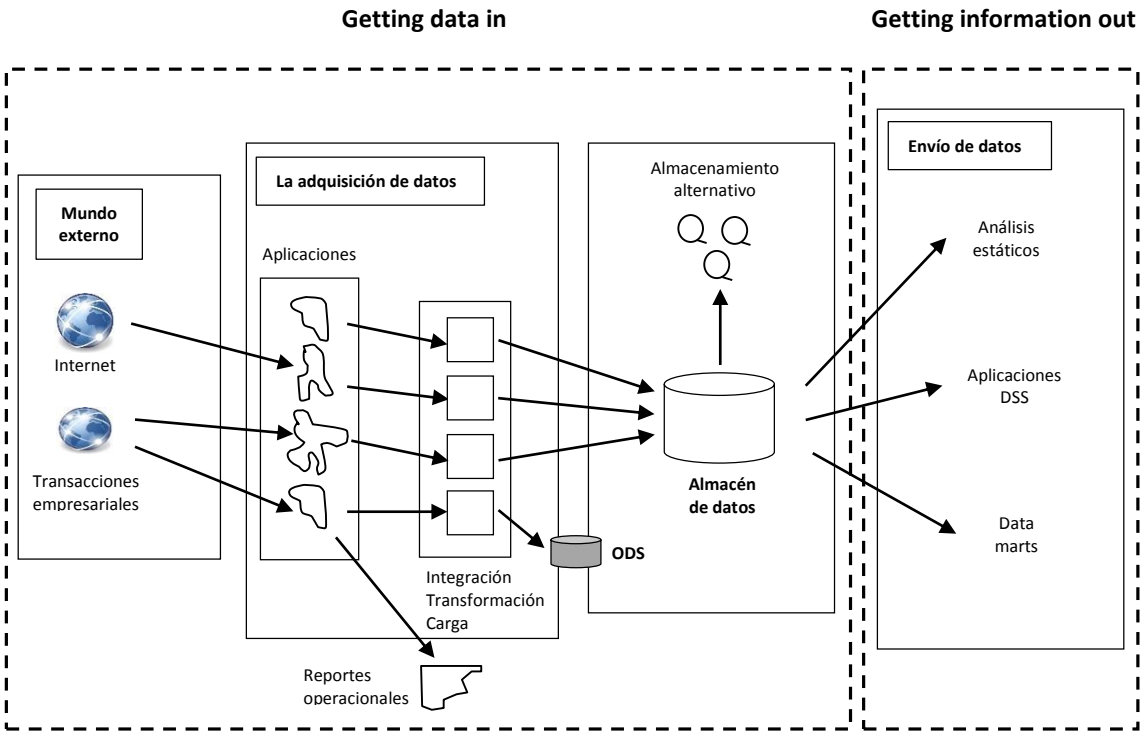

Figura 5.4. Arquitectura básica de la Fábrica de Información Corporativa (CIF). Propuesta por H Inmon, W., Imhoff, C., & Sousa, R. [17].

5.5.2. Arquitectura MultiDimensional

La arquitectura multidimensional propuesta por Kimball se basa en la idea de que todos los análisis de inteligencia de negocios tienen su base en un modelo de datos MultiDimensional (MD), de aquí en adelante lo llamaremos como arquitectura MD. Se divide en dos grupos de conjuntos qué son el Back Room y el Front Room (ver capítulo 3) como se muestra en la figura 5.5.

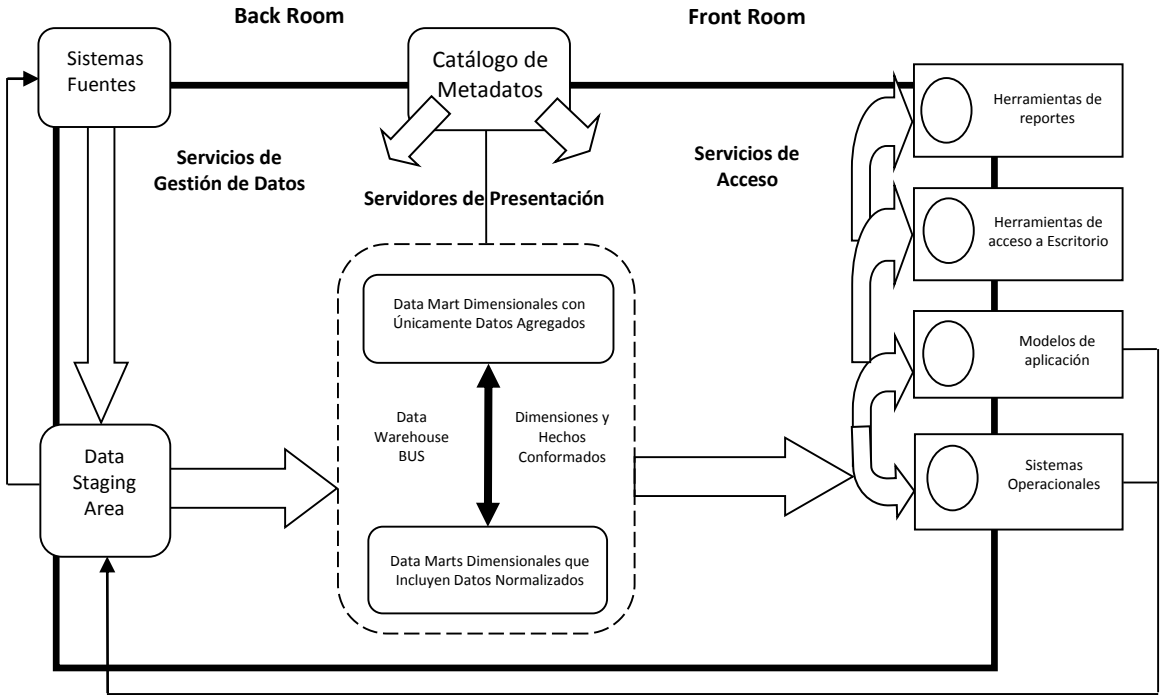

Figura 5.5. Modelo de Arquitectura Técnica de alto nivel. Propuesta por Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. [4].

5.5.3. Arquitectura para HEFESTO

Bernabeu (HEFESTO) menciona los siguientes componentes que conforman la arquitectura del data warehousing (ver figura 5.6):

- OLTP y diversas fuentes de datos.
- Load manager (administrador de carga).
- Data Warehouse Manager (administrador del almacén de datos).
- Query Manager (administrador de consultas).
- Herramientas de consulta.
- Usuarios.

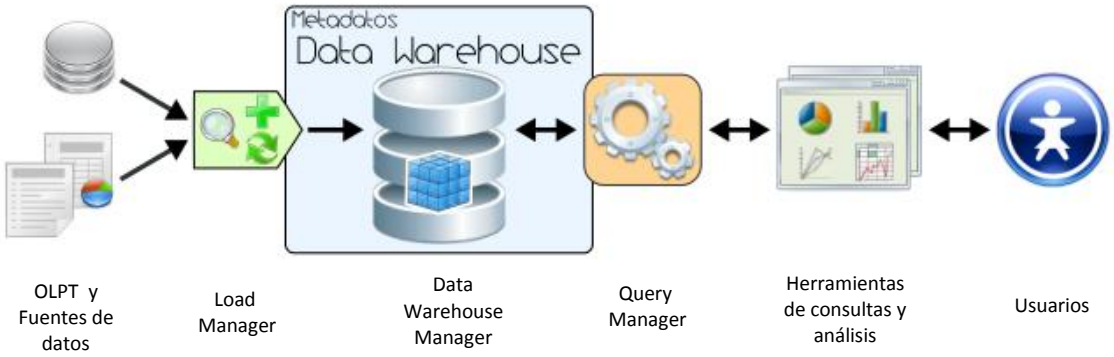

Figura 5.6. Modelo de Arquitectura Técnica de alto nivel. Propuesta por Dario, B. [7].

La forma de operar de la figura 5.6 (figura tomada de 7) básicamente es la siguiente: los datos son extraídos desde diversas aplicaciones, bases de datos, archivos de diferentes fuentes de origen, etc. Posteriormente los datos son integrados, transformados y limpiados, para luego ser cargados al almacén de datos. Principalmente la información del almacén de datos está en una estructura dimensional (ver capítulo 3) conectados a través de una arquitectura bus, ya que estos preparan esta información para responder a las consultas. Pero también puede utilizarse otro tipo de estructura de datos que es el caso del modelo relacional (ver capítulo 2). Posteriormente se explicará cual es la complejidad de cada estructura de datos.

5.6. Alcance de la arquitectura

Cada una de las arquitecturas se favorece de una u otra manera. La prioridad de la arquitectura de Inmon es sobre el ámbito de toda la empresa (recomendable cuando se tiene una empresa grande), mientras que la arquitectura de Kimball tiene una prioridad al negocio de la empresa (recomendable cuando se tiene una empresa pequeña) y la arquitectura que se opte utilizando la metodología HEFESTO de una u otra manera se favorecerá. Por ejemplo, si se tiene una empresa grande lo más conveniente es optar por utilizar la arquitectura de Inmon, ya que abarca toda la empresa y puede satisfacer las necesidades de información a todos los departamentos que se encuentren en ella. Pero si la empresa es pequeña lo conveniente es utilizar la arquitectura de Kimball, ya que la implementación de un data mart con alcance restringido a un área de negocio en específico minimiza los riesgos y produce una entrega en tiempos razonables.

Por lo tanto el alcance que puede tener la arquitectura CIF con respeto a la arquitectura MD es mucho más grande y la arquitectura que se opte para construir los data marts en la metodología HEFESTO pueden tener mayor o menor alcance en la empresa (teniendo en consideración las ventajas y desventajas de cada una de las arquitecturas antes expuestas para su elección).

5.7. Diferencias y similitudes entre arquitecturas

5.7.1. Diferencias

En general la arquitectura de un almacén de datos para Inmon, no existe un almacén de datos independiente y físicamente separado. Esta arquitectura es conveniente cuando se quiere tener un repositorio de datos centralizado para toda la empresa. Mientras tanto para Kimball los data marts del modelo estrella o copo de nieve pueden estar o no en la misma base de datos. A este modelo en una sola base de datos se le conoce como “Arquitectura bus”, esta arquitectura es buena cuando se quiere crear data marts departamentales (no necesariamente para toda la empresa). Bernabeu se basa en la arquitectura de data warehousing con sus respectivos componentes y de acuerdo a las operaciones que se deseen desarrollar, como se explicó anteriormente los data marts pueden adoptar cualquiera de los siguientes enfoques: de arriba abajo (arquitectura CIF) o de abajo arriba (arquitectura MD). Pero la prioridad para Bernabeu se centra en utilizar una arquitectura MD que una arquitectura CIF, ya que es más entendible para las personas encargadas de construir el almacén de datos.

5.7.1.1. Flujo de datos

A pesar de que el flujo de datos de la inteligencia de negocios empieza y termina en el mismo lugar para la arquitectura CIF y MD, teniendo en cuenta el alcance de cada arquitectura, ven el flujo de datos de diferente manera para la inteligencia de negocios. En general la arquitectura CIF es de arriba abajo es decir la atención se centra en los datos de la empresa que se integran en un todo (arquitectura CIF) para el uso de cualquier data mart. En cambio la arquitectura MD es de abajo arriba por lo tanto la atención se centra en la obtención de datos en unidades específicas de negocio (data marts) de forma rápida para los usuarios y para Bernabeu la prioridad de centra en la arquitectura MD para los data marts. Por ejemplo, si se opta por la arquitectura CIF para los data marts primeramente se define el almacén de datos y luego se desarrollan, construyen y cargan los data marts a partir del almacén como ya se explicó anteriormente. Pero si se opta por la arquitectura MD se definen previamente los data marts y luego se integran al almacén de datos centralizado.

5.7.1.2. Volatilidad

Para la arquitectura CIF el modelo de datos relacional (ver capítulo 2) es independiente de cualquier proceso de negocio. Es decir que el modelo de datos depende de las reglas de negocio más no de las consultas, por esta razón el modelo para esta arquitectura es bueno ya que lo hace tolerante ante cualquier cambio en el entorno de la empresa, pero malo en el sentido de que es más complejo en el modelado del almacén de datos y difícil de entender para cualquier persona que no tenga experiencia en TI.

Mientras que el modelo de datos dimensional (ver capítulo 3) de la arquitectura MD, depende de las posibles preguntas de negocio con el fin de eliminar o reducir la necesidad de reconstruir el modelo. Por ejemplo, si se produce un cambio en el proceso de negocio, el modelo multidimensional debe ser reconstruido. Algunos de esos cambios permitidos son: hechos nuevos siempre y cuando estén en el mismo nivel de granularidad con el resto de la tabla de hechos y nuevos atributos en las dimensiones. Pero no se aconseja una reconstrucción ya que la tabla de hechos puede contener miles o millones de filas. Este modelo es bueno en el sentido de que es mucho más entendible para los desarrolladores que el modelo de la arquitectura CIF. Pero malo si surge un cambio en el proceso de negocio como se mencionó anteriormente.

Para Bernabeu la prioridad es el modelo de datos multidimensional (ver capítulo 4). El cuál se debe seleccionar el tipo de modelo que se utilizará para contener la estructura del almacén de datos que se adapte mejor a los requerimientos y necesidades de los usuarios. Es bueno y malo en los mismos aspectos que se mencionaron anteriormente para Kimball.

5.7.1.3. Flexibilidad

La arquitectura CIF para el almacén de datos se puede apoyar de tecnologías que no son de carácter multidimensional. Ya que además las herramientas de análisis estadísticos requieren de archivos planos y conjuntos de datos que no dependan de diseños multidimensionales. El modelo relacional es bueno porque es flexible ante casi cualquier herramienta de análisis y malo porque personas con poca experiencia en TI no pueden manejar este tipo de herramientas.

La arquitectura MD se basa en que todos sus componentes deben ser multidimensionales en el diseño excepto el Area Data Staging. El modelo multidimensional es muy eficiente en las necesidades de algunos usuarios, pero la flexibilidad de su modelo no es buena, ya que no se puede acoplar a cualquier herramienta que no sea multidimensional, pero si es buena en el sentido que las herramientas utilizadas para el diseño de estos modelos de datos es mucho más entendible para casi cualquier tipo de usuario que no tenga experiencia en TI.

La arquitectura para Bernabeu se basa primordialmente en la arquitectura MD, ya que a pesar de que la arquitectura no es muy flexible, da seis pasos (ver 7) para que el proceso de consulta y análisis sea entendible para el usuario sin preocuparse de cuales estructuras de datos son utilizadas, ni el saber emplear el lenguaje SQL y solo enfocarse en el análisis. En este aspecto beneficia a Bernabeu cuando los usuarios no deben importarles la parte del Back Room y malo en el mismo aspecto que se mencionó anteriormente para la arquitectura MD.

5.7.1.4. Complejidad

La complejidad entre arquitecturas existe, pero de alguna u otra manera, afecta más a una arquitectura que a otra en diferentes aspectos.

La complejidad para la arquitectura CIF tiende a causar menos problemas que la arquitectura MD, ya que empieza con un modelo de datos complejo para toda la empresa y a partir de allí la creación de los data marts suelen ser más sencillos, esto por la razón de que empieza con un modelo de datos complejo y termina con un modelo de datos menos complejo. Además de que minimiza la incoherencia de los datos, ya que no se basa en preguntas, funciones o procesos de negocio. La complejidad de desarrollo de esta arquitectura al final es buena si se tiene un equipo de trabajo profesionales en TI, ya que la base de datos difícilmente puede causar problemas ante cualquier cambio inesperado (más adelante se menciona) pero es mala cuando no se tiene el personal con experiencia en el desarrollo de este tipo de arquitectura.

Por otro lado el modelo de datos para la arquitectura MD es buena para personas que no tengan mucha experiencia en este tipo de arquitectura, ya que el modelo de datos es menos complejo (modelo dimensional estrella o copo de nieve (ver capítulo 1)) y fáciles de entender para los integrantes de la empresa. Sin embargo es malo en su modelo de datos ya que se basa en preguntas de negocio causando problemas ante cualquier cambio inesperado (más adelante se menciona) y además la independencia de los data marts al momento de unirlos pueden causar inestabilidad.

Bernabeu se centra principalmente en la arquitectura MD. La ventaja que tiene respecto a las razones que considera Kimball es que también se puede optar por la arquitectura CIF y pueden involucrar personas que se encargan de tomar decisiones, planificar las actividades del negocio y cualquier usuario que no tenga relación directa con la empresa y las desventajas son las mismas a las de Kimball como se mencionó anteriormente. Pero hay que tener en cuenta que si se opta por la arquitectura CIF lo recomendable es considerar personas con conocimientos en TI.

5.7.1.5. Funcionalidad

La arquitectura CIF se basa en un modelo ER que soporta a las reglas de negocio de la empresa. El almacén de datos apoya a todas las formas de análisis de datos estratégicas, no solo multidimensionales. Por ejemplo: la minería de datos, análisis estadísticos y ad hoc. Es buena en el sentido de que podemos explotar el almacén de datos con una diversidad de herramientas de análisis como se muestran algunas en la tabla 5.4 y mala si no se utilizan las herramientas adecuadas para su explotación.

Herramientas	Descripción
Clementine/SPSS	Herramienta de data mining que permite desarrollar modelos predictivos y desplegarlos para mejorar la toma de decisiones. Está diseñada teniendo en cuenta a los usuarios empresariales.
InfoMaker	InfoMaker es una interfaz fácil de usar que permite generar informes útiles y que facilitan la toma de decisiones. Los usuarios no necesitan conocer el lenguaje de base de datos para saber cómo acceder a los datos.
Oracle Exadata y Oracle Database	Ofrecen una plataforma rápida, confiable y rentable para el almacén de datos e inteligencia de negocios que escala fácilmente para satisfacer las necesidades de informes complejos y análisis de las organizaciones.
SAS Analytics	Proporciona un entorno integrado para el modelado predictivo y descriptivo, la minería de datos, la analítica de textos, la predicción, la optimización, la simulación y el diseño experimental, entre muchos otros.

Tabla 5.4. Herramientas. Tomados de [46, 47, 48].

La arquitectura MD proporciona un entorno de procesamiento multidimensional, es buena ya que asegura un buen rendimiento en “slice and dice”, “drill-up” y “drill-down”, alrededor de las consultas. Ya que todas las dimensiones son equivalentes entre sí, lo que significa que todas las preguntas dentro de los límites del modelo estrella se procesa más o menos de la misma manera. Sin embargo el modelo multidimensional es malo ya que no se acomoda fácilmente a métodos de análisis como la minería de datos y análisis estadísticos o de cualquier tipo de herramienta que no sea multidimensional. En la tabla 5.5 se muestran algunas herramientas que pueden utilizarse para la explotación del almacén de datos.

Herramientas	Descripción
AtlasSBI	Plataforma de Business Intelligence “todo en uno”, que incluye: Informes estáticos y dinámicos (OLAP), dashboards, cuadro de mando y posicionamiento de datos en mapas.
Designer (Universos)	Es una herramienta potente y el plus que la Suite de Business Objects tiene en comparación con otras herramientas de inteligencia de negocios, ya que crea una capa intermedia (capa semántica), que despliega los modelos transaccionales y multidimensionales a través de un lenguaje neutro, orientado al usuario, ideal para la comprensión y utilización de él.
Explorer (polestar)	Es una herramienta que se diferencia del resto ya que permite el análisis sencillo y rápido, debido a que se realizan consultas ya calculadas, estas consultas se denominan espacios y consisten en cubos OLAP que son indexados previamente antes de su consulta.
Pentaho Analysis	Con esta solución se puede navegar por la información con el uso de tablas dinámicas, cubos multidimensionales, dashboards, entre otros, permitiendo analizar de manera rápida el comportamiento de la empresa en sus diferentes rubros.

Tabla 5.5. Herramientas multidimensionales. Tomados de [49].

Bernabeu se basa en un modelo multidimensional, las herramientas de consulta y análisis son sistemas que permiten a los usuarios realizar la exploración de datos del almacén de datos. Básicamente constituyen el nexo entre el almacén y los usuarios. Existen diferentes tipos de herramientas de consulta y análisis, y de acuerdo a la necesidad, tipos de usuarios y requerimientos de información, se deben seleccionar las más adecuadas al caso. Entre ellas se destacan las siguientes: Reportes y consultas, OLAP, Dashboards[11], Data Mining, EIS[12], etc.

5.7.2. Similitudes

Las similitudes que existen entre los tres autores, son tiempo similar (datos con fecha y hora), proceso ETL y resultado de la consulta. A continuación se explica cada similitud:

5.7.2.1. Tiempo similar

El atributo tiempo es la característica definitoria más importante de los datos de un almacén de datos. Por ejemplo, el atributo tiempo permite el análisis de apoyo a las decisiones para comparar las ventas de un producto X de este año con el año pasado o para determinar si más de un producto se vende el fin de semana en días festivos. Cada autor llama de diferente nombre el atributo tiempo, como se muestra en la tabla 5.6.

Atributo	Inmon	Kimball	Bernabeu
Tiempo	Elemento Tiempo.	Dimensión Fecha o Dimensión Tiempo.	Dimensión Tiempo.

Tabla 5.6. Atributo tiempo.

Independientemente de que metodología se utilice, los usuarios finales pueden consultar los datos de día, mes, trimestre, año, día de fiesta, día de la semana, fin de semana, etc.

5.7.2.2. Proceso ETL

Para Inmon, Kimball y Bernabeu en el entorno de almacén de datos se empieza con el proceso ETL (ver capítulo 1). Los datos se extraen de las fuentes de datos, es transformado para llevar a cabo la estandarización de almacenamiento y posteriormente cargado. Los datos se cargan al almacén de datos ya sea en una sola base de datos (arquitectura CIF) o en una serie de bases de datos pequeñas conocidas como data marts (arquitectura MD). En la tabla 5.7 se muestran algunas herramientas utilizadas para llevar a cabo el proceso ETL.

Herramientas	Descripción
Oracle Warehouse Builder	Es una herramienta ETL producida por Oracle, que ofrece un entorno gráfico para construir, gestionar y mantener la integración de datos en los procesos de inteligencia de negocios de sistemas.
Businessobjects data integrator	Es una herramienta ETL, el software incluye características de calidad de los datos. Se utiliza comúnmente para la construcción de data marts y almacenes de datos.

[11] Dashboards es una interfaz donde el usuario puede administrar el equipo y/o software (tablero de instrumentos).

[12] Por sus siglas en Inglés Executive Information System es una herramienta software, basada en DSS, que provee a los gerentes de un acceso sencillo a la información interna y externa de su compañía.

IBM - InfoSphere DataStage	Es una herramienta de IBM que permite crear y mantener fácil y rápidamente data marts y data warehouse. Soporta la extracción, integración y transformación de altos volúmenes de datos desde estructuras simples hasta muy complejas.
Pentaho - Kettle	Programa ETL que incluye un conjunto de herramientas para realizar la extracción y transformación de datos. Uno de sus objetivos es que el proyecto ETL sea fácil de generar, mantener y desplegar.

Tabla 5.7. Herramientas ETL. Tomados de [50, 51, 52].

5.7.2.3. Resultado de la consulta

El resultado de la consulta debe ser el mismo independientemente del enfoque que se opte por desarrollar. Hay que tener cuidado que las herramientas que se utilicen para llevar a cabo la consulta no deben casarse con el almacén de datos, esto para no depender de alguna herramienta o tecnología en particular.

En la tabla 5.8 se muestra un resumen de las diferencias y similitudes entre arquitecturas.

Diferencias	**Inmon**	**Kimball**	**Bernabeu**
Arquitectura	CIF: Diseño de un almacén de datos para toda la empresa.	MD: Las etapas de desarrollo de un data mart se basan en procesos específicos de negocio.	Arquitectura Data Warehousing (bajo el enfoque de Kimball).
Diseño físico de la arquitectura	Bastante complejo.	Moderado.	Moderado.
Alcance de la arquitectura	Grande.	Medio.	Medio.
Flujo de datos	De arriba abajo	De abajo arriba.	De abajo arriba.
Volatilidad	Independiente de cualquier proceso de negocio: se basa en las reglas de negocio.	Se basa en las preguntas de negocio.	Se basa en las preguntas de negocio para responder cualquier consulta.
Flexibilidad	Es muy flexible.	Flexibilidad no muy buena.	Flexibilidad no es muy buena.
Complejidad	Compleja.	Moderada.	Moderada.
Similitudes			
Tiempo similar	El uso de los datos con fecha y hora.		
Proceso ETL	En el entorno de almacén de datos se empieza con el proceso ETL.		
Resultado de la consulta	El resultado es el mismo independiente del enfoque que se opte por desarrollar.		

Tabla 5.8. Resumen de las diferencias y similitudes entre arquitecturas. Basado en Breslin, M. [12].

5.8. Modelado de datos

El modelo de datos conceptual de Inmon está orientado a temas (ver capítulo 2). Donde existe compatibilidad con las herramientas tradicionales tales como los ERDs[13] y DISs[14]. Para el diseño del modelado de datos de Inmon es importante comenzar con un modelo en 3FN ya que es la mejor estructura de datos para el almacén mediante la aplicación de ocho pasos para su transformación final, como se explicó en el capítulo 2. Mientras la normalización (a veces se aplica la desnormalización) de los datos del modelo lógico (modelo relacional: 3NF) de Inmon no es óptima en el rendimiento de los datos, pero los datos detallados es fácilmente disponible para los data marts

[13] Por sus siglas en Inglés Entity-Relationship: Modelo de datos de alto nivel.
[14] Por sus siglas en Inglés Data Item Set es un conjunto de elementos de datos, cada uno de los cuales se relaciona directamente con la llave de la agrupación de datos, en la que los elementos de datos se encuentran.

y que además reduce el espacio de almacenamiento. Este tipo de modelo es flexible ante cualquier cambio inesperado en la empresa. Por ejemplo, si se integra un nuevo departamento en la empresa que se quiera incluir en la base de datos o cualquier otro tipo de consulta que posteriormente pueda surgir en el negocio. Es recomendable desarrollar para equipos de trabajos estables y que tengan experiencias en diseño de bases de datos y de TI. Este modelo no es recomendable para equipos de trabajos que no cuenten con personas con formación técnica en TI.

Por otro lado el modelo de datos conceptual de Kimball está orientado a procesos (ver capítulo 3), que significa que está determinado por las necesidades del usuario final, donde existe compatibilidad con las herramientas multidimensionales. El diseño de modelado dimensional de Kimball consiste en una serie de cuatro pasos para su diseño, como se explicó en el capítulo 3. Este modelo lógico (modelo dimensional estrella o copo de nieve) es bueno desarrollarse cuando la empresa no cuenta con muchas personas con formaciones técnicas en TI, ya que es fácil de entender para los desarrolladores, pero no es recomendable en el sentido de que no es muy flexible. Por ejemplo, si en la empresa surgen nuevas preguntas en el proceso de negocio, lo más probable es reconstruir el modelo lo que ocasionaría mayor tiempo en el desarrollo e inestabilidad en el modelo de datos.

Para Bernabeu tiene la prioridad para su diseño el modelado dimensional como se explicó en el capítulo 4. El cuál se debe seleccionar el tipo de modelo que se utilizará para contener la estructura del almacén de datos que se adapte mejor a los requerimientos y necesidades de los usuarios. Es muy importante definir objetivamente si se empleará un modelo estrella o copo de nieve, ya que esta decisión afectará considerablemente la elaboración del modelo lógico. Pero también se puede optar por el modelo relacional, tomando en cuenta los beneficios y contras que pueda tener el modelo de datos como se mencionó en los dos últimos párrafos.

Modelado de datos	Inmon	Kimball	Bernabeu
Orientación de los datos	Orientado a temas o basados en datos.	Orientado a procesos.	Orientado a procesos.
Estructura de los datos	Datos no métricos: Modelo relacional (tercera forma normal: 3NF).	Métricas de negocio, medidas de rendimiento y scorecards[15]: Modelo dimensional (estrella o copo de nieve).	Modelo estrella o copo de nieve: Se estructura en cubos multidimensionales.
Nivel de granularidad	El más alto nivel de granularidad posible y debe incluir todos los datos históricos posibles dentro de una empresa.	Alto.	Alto.
Requerimientos de integración de los datos	Integración corporativa.	Áreas de negocio individual.	Análisis multivariables.
Persistencia de los datos	Alto.	Estable.	Estable.
Herramientas	Tradicionales (ERDs, DISs).	Modelado dimensional.	Modelado dimensional.

Tabla 5.9. Modelado de datos. Basado en Breslin, M. [12].

5.9. Filosofía

Inmon, Kimball y Bernabeu, han comenzado con diferentes filosofías para la recolección de información en toda la empresa, gestión de información y análisis para la toma de decisiones, en busca de un objetivo en común. A continuación se presentan las filosofías de cada autor:

[15] Es un método para medir las actividades de una compañía en términos de su visión y estrategia. Proporciona a los gerentes una mirada global del desempeño del negocio.

5.9.1. Filosofía de Inmon: Evolutiva, no revolucionaria

Inmon ve un almacén de datos como una parte integral de la Fábrica de Información Corporativa [12]. Esto quiere decir que un almacén de datos y las bases de datos operacionales son parte de un todo. Está percepción ayuda a explicar que el enfoque evolutivo de Inmon puede ser fácilmente justificada en términos de diseño y desarrollo.

El modelo de datos en si evoluciona iterativamente y puede ser revisado a fin de garantizar que se cumplan las expectativas de los analistas de Sistemas de Soporte a las Decisiones (DSS). Por lo tanto significa que el modelo de datos está cambiando constantemente para reflejar las necesidades de la organización. En contraste con el enfoque "Big Bang", que es de naturaleza revolucionaria. Visto de este contexto el modelo de Inmon es mucho más evolutivo que revolucionaria.

5.9.2. Filosofía de Kimball

Los requerimientos de negocio impulsan tanto el proceso y construcción del almacén de datos. En el primer capítulo de su libro *"The Data Warehouse Toolkit, segunda edición"* [12], define los objetivos de un almacén de datos:

- Hacer que la información de fácil acceso.
- Presentar la información de la organización consistente.
- Ser adaptable y resistente al cambio.
- Proteger la información.
- Servir como base para la mejor toma de decisiones.

Entonces para Kimball, la aceptación de un almacén de datos se mide en base al tiempo de uso, que está relacionado con su sencillez de usarlo, esto lo hace esencial para la filosofía de Kimball y con sus cuatro pasos para el diseño del modelo dimensional sea fácil de entender para el usuario final.

5.9.3. Filosofía de la metodología HEFESTO

La construcción e implementación de un almacén de datos puede adaptarse muy bien a cualquier ciclo de vida de desarrollo de software, con la salvedad de que para algunas fases en particular, las acciones que se han de realizar serán muy diferentes. Lo que se debe tener muy en cuenta, es no entrar en la utilización de metodologías que requieran fases extensas de reunión de requerimientos y análisis, fases de desarrollo que conlleve demasiado tiempo y fases de despliegue muy largas. Lo que se busca, es entregar una primera implementación que satisfaga una parte de las necesidades, para demostrar las ventajas del almacén de datos y motivar a los usuarios.

5.9.4. Diferencias filosóficas

Inmon ve a los profesionales de TI como los principales desarrolladores del almacén de datos, por esta razón el rendimientos del almacén se maximiza. Mientras que Kimball ve a los usuarios finales y los profesionales de TI como personas con casi las mismas funciones a cumplir. Está claro que para Kimball la participación de los usuarios finales en el desarrollo del almacén de datos, ayuda a tener más probabilidad de aceptación del almacén y para Bernabeu las personas que se encargan de planificar las actividades del negocio, tomar decisiones y profesionales de TI deben involucrarse en el diseño y construcción del almacén de datos.

Tanto Inmon, Kimball y Bernabeu están de acuerdo que en la construcción de un almacén de datos los usuarios no tienen que participar en todas las fases de construcción.

En la tabla 5.10 se muestra la parte principal de la filosofía de cada metodología.

Filosofía	Inmon	Kimball	Bernabeu
Principales usuarios	Los profesionales de TI.	Los usuarios finales.	Las personas que se encargan de planificar las actividades del negocio y tomar decisiones.

Tabla 5.10. Filosofía. Basado en Breslin, M. [12].

5.10. Entonces ¿Cuál es la mejor metodología a elegir?

Para poder responder a esta pregunta tenemos que hacer un análisis general de aspectos específicos que se han mencionado con anterioridad.

En una empresa si se opta por la metodología de Inmon lo más probable es que tenga éxito, ya que este enfoque tiene un gran equipo de especialistas en TI para el diseño de un almacén de datos, planeación del proyecto de toda la empresa. Por tal motivo Inmon propone construir una infraestructura sólida del modelo de datos para toda la empresa. Esta metodología es ideal para empresas grandes, que cuenten con un buen presupuesto inicial, se den el tiempo necesario para su desarrollo. Sin embargo el inconveniente principal que tiene esta metodología es que para ver los resultados, puede llevar varios meses (de 4 a 9 meses [12]).

Por otra parte para una organización con características diferentes (empresa pequeña, bajo presupuesto inicial, equipos pequeños de trabajos generalistas) a la de Inmon es mejor el enfoque de Kimball, ya que se puede construir un data mart de un área específica de negocio en 90 días y los siguientes data marts que se construyan serán de 60 a 90 días cada uno [12]. El enfoque de Kimball indica que la organización es capaz de desplegar equipos de trabajo generalistas de desarrollo de proyectos más pequeños y espera almacenar métricas de negocio. Esta metodología es ideal para empresas pequeñas, que cuenten con bajos costos iniciales y quieran ver resultados rápidos. Pero los principales inconvenientes que tiene esta metodología son el mantenimiento de las bases de datos y la unión de los data marts.

La idea principal de la metodología de Bernabeu (HEFESTO) es construir un almacén de datos centrándose en lo más práctico, de forma sencilla, ordenada y que se comprenda cada paso a realizar, para no caer en el aburrimiento y sin saber qué es lo que se está haciendo. Las fases de construcción de su metodología (ver capítulo 4) para la creación de un almacén de datos deben ser cortas (independientemente si la empresa es grande o pequeña), que no lleven demasiado tiempo y adaptándose a cualquier ciclo de vida de desarrollo de software, esto con el objetivo de generar resultados rápidos, para motivar a los usuarios las ventajas que puede tener un almacén de datos en la toma de decisiones, independiente del enfoque se desarrolle “De arriba abajo” o “De abajo arriba”. Sin embargo el tiempo de construcción del almacén dependerá mucho del alcance del negocio y su costo de construcción e implementación, de las herramientas y modelos de datos que se utilicen, etc.

La elección del enfoque de diseño y construcción de un almacén de datos a simple vista parece sencilla como se explicó en los párrafos anteriores, pero no es fácil como parece. Los proyectos tienen éxito no por el diseño innovador o una tecnología nueva, sino que se debe a aspectos de

liderazgo, comunicación, planeación y las relaciones interpersonales. Por eso hay que tener en consideración las habilidades sociales de las empresas incluso más importantes que las habilidades técnicas.

Ahora sí entonces ¿Cuál es la mejor metodología a elegir? En lo personal no creo que una metodología sea mejor que otra, sino que cada metodología tiene un punto de vista diferente a la hora de diseñar y construir un almacén de datos. Como ya se ha visto anteriormente, el enfoque a adoptar para el diseño y construcción de un almacén de datos dependerá de los objetivos de la organización, naturaleza de negocio, tiempo y el costo que supone. Inclusive puede llegar a suceder que se haga una combinación de las metodologías. Una cosa hay que resaltar que tenemos que partir siempre de las necesidades del negocio más no de las herramientas o tecnologías que tengamos, para no depender de ellas a la hora de diseñar, construir y explotar el almacén de datos.

Conclusiones

El presente trabajo tuvo como objetivo principal de comparar técnicas de diseño y construcción de almacenes de datos e integrarlas en un solo documento, además de desarrollar un ejemplo práctico para cada una de las metodologías y ver cuáles son las diferencias y similitudes entre ellas en su desarrollo, que me ha permitido obtener un aprendizaje y derivar una serie de conclusiones que presento a continuación y el por qué elegí las tres metodologías expuestas en capítulos anteriores.

Elegí la metodología de Inmon porque me pareció una metodología con bases sólidas de amplio alcance en una organización y detallada en cada una de sus fases los pasos que componen la metodología. Además Inmon enfatiza la limpieza en el diseño y normas que garanticen la exactitud, integración y coherencia de los datos.

Por otro lado elegí la metodología de Kimball ya que me llamo la atención de que es la metodología más aceptada y efectiva para desarrollar almacenes de datos de todas las que existen. Es una metodología que se puede implementar y ver resultados en poco tiempo, para demostrar el valor de la solución al negocio. Por otra parte Kimball representa la relación de los datos con el usuario final, flexibilidad y rápida explotación de la información.

Y finalmente elegí la metodología de Bernabeu (HEFESTO), ya que a pesar de que es una metodología reciente y que está en constantes cambios realizados por su creador. Me pareció una metodología con pasos cortos y sencillos, centrándose en un ejemplo práctico para su mejor entendimiento. Es decir que se comprenda cada paso en el diseño y construcción de un almacén de datos, para no caer en lo tedioso y sin saber lo que se está construyendo.

Al realizar este trabajo he comprendido que nos encontramos en una era en que la información es indispensable para la toma de decisiones y éxitos de las organizaciones, y de analizar grandes volúmenes de datos con la tecnología de desarrollo adecuada, independientemente de qué metodología se utilice para diseñar y construir un almacén de datos.

Originalmente la tecnología de almacén de datos se desarrolló con el objetivo de brindar una ventaja competitiva a las organizaciones y empresas. Pero actualmente brinda administración de una base de datos, haciendo más fácil la obtención de la información necesaria y oportuna, de esta forma tomar decisiones acertadas. Los almacenes de datos hoy en día son muy utilizados por las organizaciones de diversos tipos que van desde empresariales hasta científicas. Sin embargo el éxito de un almacén de datos depende en gran parte de su diseño adecuado, que permita cubrir las necesidades de la organización. Para cumplir esto el diseñador debe poner especial atención a los requerimientos de los usuarios y es muy importante que domine los conceptos y técnicas recomendadas en este trabajo de tesis.

Dado que los recursos son siempre limitados para el diseño y construcción de un almacén de datos en una organización, es necesario conocer y explotar de la manera más adecuada la tecnología que tengamos, pero sin que el almacén de datos dependa de alguna tecnología o herramienta en particular. Al tener grandes volúmenes de datos en un almacén, para que esta se convierta en información útil para el usuario final, es necesario minimizar el tiempo de acceso a esa información durante la realización de consultas, ya que los recursos de rendimientos pueden saturarse y no satisfacer las necesidades de los usuarios. El mantenimiento de los datos almacenados en la base de datos puede volverse una tarea compleja si los volúmenes de datos son grandes y no se utilizan métodos de particionamiento adecuados, que en este caso no es el objetivo de este trabajo, pero si

teniendo en cuenta cual modelo de datos, que en capítulos anteriores se han expuestos se pueden ajustar a las necesidades de los usuarios, ya que esto también ayuda a optimizar el rendimiento de las consultas.

A lo largo de los capítulos anteriores se dieron a conocer conceptos generales, ventajas y desventajas, características, componentes y estructuras, técnicas, metodologías, entre otros aspectos y consideraciones para su diseño, construcción de un almacén de datos: Inmon, Kimball, Bernabeu. Entonces ¿Cuál es la mejor metodología a elegir de las antes mencionadas? En mi opinión no existe como tal una metodología que garantice el éxito del diseño y construcción de un almacén de datos. Si no que depende de las necesidades que queramos que satisfaga, como se mencionaran algunas a continuación.

El enfoque de desarrollo de Inmon “De arriba abajo” es adecuando para empresas estables, que puedan darse el tiempo necesario para el diseño y costos involucrados. Por otro lado su arquitectura CIF gestiona los datos de requerimientos de negocio para servir a toda la empresa y a partir de esta arquitectura, las bases de datos departamentales (data marts) son desarrolladas para servir a un departamento o un área en específica de negocio para el apoyo de la toma de decisiones. Sin embargo este tipo de enfoque es ambicioso, consume mucho tiempo y el costo de construcción de un almacén de datos es elevado. Inmon dice que construir el modelo de datos, consiste en la transformación del modelo de negocio (modelo en tercera forma normal) con ochos pasos definidos (ver capítulo 2).

Pero si la optimización de un departamento de alguna empresa es lo suficientemente buena y la atención se centra en la rapidez de obtención de resultados, es recomendable utilizar el enfoque de Kimball “De abajo arriba”. El tiempo de construcción es más rápido y su costo es bajo en comparación del enfoque de Inmon, ya que se empieza por la creación de los data marts de los principales procesos de negocio y el bus de datos parte de la arquitectura de Kimball garantiza la interoperabilidad entre los distintos data marts. El bus de datos requiere que todos los data marts se modelen dentro de los estándares de datos llamados dimensiones compatibles. Sin embargo los problemas asociados con los data marts no se hacen evidentes hasta que se han construidos varios data marts, ya que puede haber incompatibilidad de estructura de datos. Kimball recomienda seguir cuatro pasos (ver capítulo 3) para cada data mart en el modelo de datos dimensional.

Para Bernabeu la construcción e implementación de un almacén de datos se puede acoplar a cualquier ciclo de vida de desarrollo de software independiente del enfoque que se utilice “De arriba abajo” o “De abajo arriba”, teniendo en cuenta que algunas fases en particular las acciones que se realicen serán diferentes. Bernabeu se enfoca en el diseño de un almacén de datos lo más práctico y sencillo posible teniendo como prioridad el enfoque de Kimball. Pero como se mencionó en capítulos anteriores dependiendo de las necesidades de la empresa y tomando en consideración las ventajas y desventajas se puede optar por el enfoque de Inmon.

Un almacén de datos puede dar una serie de beneficios para la organización: accesible, correcta, uniforme y actualizada. Si el almacén de datos está bien implementado los beneficios obtenidos serán muchos de lo contrario su implementación generará inconvenientes como se ha mencionado en capítulos anteriores.

Para terminar el concepto de almacén de datos va más allá que un repositorio de datos. El sistema debe ofrecer una solución completa para gestionar y controlar el flujo de información desde las bases de datos corporativas y externas para dar soporte a la toma de decisiones a los usuarios finales. Además de permitir a los usuarios conozcan la información que existe en el almacén de datos, como poder acceder a ella y manipular con las herramientas adecuadas. Ya que un almacén de datos

puede contener la mejor información pero pierde su valor sino se tienen las herramientas adecuadas para explotarlo.

Bibliografía

[1] H Inmon, W. (2005). *Building the Data Warehouse* (4 ed.). Indianapolis, Indiana: Wiley Publishing, Inc.

[2] Kimball, R., & Caserta, J. (2004). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* Wiley Publishing, Inc.

[3] Marchese, A., Picco, A., Pluss, J. J., Díaz, D., Gaibazzi Ma, F., Repetto, L., y otros. (2002). Séptimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística: nuevas tecnologías para la administración de empresas. *Séptimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística*, (págs. 1-17). Argentina.

[4] Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. (1998). *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses.* Indianapolis, Indiana: Wiley Publishing, Inc.

[5] Ponniah, P. (2010). *Data warehousing fundamentals for it professionals* (2 ed.). Hoboken, New Jersey: Wiley a John Wiley & Sons, Inc.

[6] Daneshpour, N., & Abdollahzadeh Barfourosh, A. (2011). Data Engineering Approach to Efficient Data Warehouse: life cycle development. *IEEE Xplorer Digital Library*, 109-120.

[7] Dario, B. (2010). *HEFESTO: DATA WAREHOUSING: Investigación y Sistematización. Metodología para la Construcción de un Data Warehouse* (Versión 2.0 ed.). Córdoba, Argentina: Licencia de Documentación Libre de GNU.

[8] Simitsis, A., Vassiliadis, P., & Sellis, T. (2005). Optimizing ETL Processes in Data Warehouses. *IEEE Xplorer Digital Library*, 564-575.

[9] Muñoz, L., Mazón, J. N., & Trujillo, J. (2010). Systematic Review and comparison of modeling ETL processes in Data Warehouse. *IEEE Xplorer Digital Library*, 1-6.

[10] Muñoz, L., Mazón, J. N., & Trujillo, J. (2011). ETL Process Modeling Conceptual for Data Warehouses: A Systematic Mapping Study. *IEEE Latin America Transactions, 9*(3), 360-365.

[11] Ribeiro, L., Goldschmidt, R., & Cavalcanti, M. (2011). Complementing Data in the ETL Process. *Springer-Verlag Berlin Heidelberg*, 112-123.

[12] Breslin, M. (2004). Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models. *BUSINESS INTELLIGENCE JOURNAL, WINTER*, 6-20.

[13] Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (2 ed.). Wiley Computer Publishing.

[14] H Inmon, W. (2002). *Building the Data Warehouse* (3 ed.). John Wiley & Sons, Inc.

[15] Dario, B. (2012). *PACKT PUBLISIHING*. Recuperado el 03 de junio de 2012, de http://www.packtpub.com/authors/profiles/bernabeu-r-dario

[16] Imhoff, C., Galemmo, N., & Geiger, J. G. (2003). *Mastering Data Warehouse Design: Relational and Dimensional Techniques.* Wiley Publishing, Inc.

[17] H Inmon, W., Imhoff, C., & Sousa, R. (2001). *Corporate Information Factory* (2 ed.). John Wiley & Sons, Inc.

[18] Frankel, D. S., Harmon, P., Mukerji, J., Odell, J., Owen, M., Rivitt, P., y otros. (2003). The Zachman Framework and the OMG's Model Driven Architecture. *Business Process Trends*, 1-14.

[19] Widom, J. (1995). Research Problems in Data Warehousing. *Proc. of 4th Int'l Conference on Information and Knowledge Management (CIKM)*, (págs. 1-6).

[20] H Inmon, W. (2012). *Corporate Information Factory*. Recuperado el 20 de agosto de 2012, de http://www.inmoncif.com/home/

[21] Rainardi, V. (2008). *Building a Data Warehouse: With Examples in SQL Server.* Apress.

[22] Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3 ed.). Indianapolis, Indiana: John Wiley & Sons, Inc.

[23] Date, C. J. (2001). *Introducción a los sistemas de bases de datos* (7 ed.). (S. M. Ruiz Faudón, & S. Kourchenko Barrena, Trads.) Prentice Hall.

[24] Mendez, A., Mártire, A., Britos, P., & Garcia Martínez, R. (2003). Fundamentos de Data Warehouse. *Reportes Técnicos en Ingeniería del Software, 5*(1), 19-26.

[25] Wanumen S, L. F., Mosquera Palacios, D., & Rivas Trujillo, E. (2012). Comparación entre enfoques de prueba en la verificación de una arquitectura de integración. *Tecnura, 16*, 218-224.

[26] Sai Satyanarayana Reddy, S., S S Reddy, L., & Lavanya, A. (2009). Advanced Techniques for Scientific Data Warehouses. *IEEE Computer Society*, 576-580.

[27] Arora, R., Pahwa, P., & Bansal, S. (2009). Alliance Rules for Data Warehouse Cleansing. *IEEE Conference Publications*, 743-747.

[28] Rudra, A., & Nimmagadda, S. L. (2005). Roles of Multidimensionality and Granularity in Warehousing Australian Resources Data. *IEEE Xplore Digital Library*, 1-6.

[29] Silberschatz, A., Korth, H. F., & Korth, S. (2010). *Database System Concepts* (6 ed.). McGraw-Hill.

[30] Elmasri, R., & Navathe, S. B. (2010). *Fundamentals of Database Systems* (6 ed.). Addison Wesley.

[31] Jian, L., & Bihua, X. (2010). ETL Tool Research and Implementation Based on Drilling Data Warehouse. *IEEE Xplore Library Digital*, 2567-2569.

[32] Espino Barrios, L. F. (2009). Sistemas de Bases de Datos Federadas. 1-9.

[33] Mohammad, R., Kianmehr, K., & Ridley, M. J. (2008). Data Warehouse Architecture and Design. *IEEE Xplore Library Digital*, 58-63.

[34] Levene, M., & Loizou, G. (2003). Why is the snowflake schema a good data warehouse design? *ACM LC Digital Library*, 225-240.

[35] Jian-bo, W., & Chong-jun, F. (2012). Research on Airport Data Warehouse Architecture. *International Journal of Business, Humanities and Technology, 2*(4), 107-111.

[36] Saroop, S., & Kumar, M. (2011). Comparison of Data Warehouse Design Approaches from User Requirement to Conceptual Model: A Survey. *IEEE Xplore Library Digital*, 308-312.

[37] Golfarelli, M., & Rizzi, S. (2009). A Survey on Temporal Data Warehousing. *IGI Global, 5*(1), 1-17.

[38] Munoz Palacio, J. M. (2010). *Information systems development methodologies for Data-driven Decision Support Systems.*

[39] Silvers, F. (2008). *Building and Maintaining a Data Warehouse.* CRC Press, Taylor & Francis Group.

[40] WebMining Consultores. (11 de octubre de 2011). *WebMining Powering Web Intelligence*. Recuperado el 21 de enero de 2013, de http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/

[41] Trujillo, J. C., Mazón, J. N., & Pardillo, J. (2009). *Diseño y explotación de almacenes de datos: Conceptos Básicos de Modelado Multidimensional.* ECU EDITORIAL CLUB UNIVERSITARIO.

[42] The Kimball Group. (2013). *KIMBALL GROUP*. Recuperado el 09 de abril de 2013, de http://www.kimballgroup.com/

[43] SIS KLE. (2013). *KE*. Recuperado el 22 de junio de 2013, de http://kle.sisorg.com.mx/

[44] George, S. (2013). *SearchBusinessIntelligence.in*. Recuperado el 11 de septiembre de 2013, de http://searchbusinessintelligence.techtarget.in/tip/Inmon-vs-Kimball-Which-approach-is-suitable-for-your-data-warehouse

[45] The Data Warehousing Institute. (s.f.). TDWI Data Warehousing Concepts and Principles.

[46] Sybase. (2013). *Sybase*. Recuperado el 02 de octubre de 2013, de http://www.sybase.com.mx/products/datawarehousing

[47] Oracle. (2013). *Oracle Data Warehousing*. Recuperado el 16 de noviembre de 2013, de http://www.oracle.com/lad/products/database/datawarehousing/overview/index.html

[48] SAS. (2013). *SAS*. Recuperado el 02 de diciembre de 2013, de http://www.sas.com/offices/europe/spain/news/apredictivo.html

[49] Universidad de Concepción Chile. (2013). *DTI Dirección de Tecnologías de Información*. Recuperado el 06 de diciembre de 2013, de http://www.udec.cl/dti/node/110

[50] SISMEXICO. (2013). *SISMEXICO*. Recuperado el 06 de diciembre de 2013, de http://sismexico.sisorg.com.mx/ibmtools.html

[51] Oracle. (2013). *Oracle*. Recuperado el 06 de diciembre de 2013, de http://www.oracle.com/technetwork/developer-tools/warehouse/overview/introduction/index.html

[52] Pentaho. (2013). *Pentaho Data Integration*. Recuperado el 09 de diciembre de 2013, de http://www.pentaho.com/product/data-integration

[53] ABAST. (2013). *ABAST*. Recuperado el 19 de diciembre de 2013, de http://www.abast.es/business_objects_data_integrator.shtml