UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CONTADURIA Y ADMINISTRACION

DATA WAREHOUSE

TESIS PROFESIONAL

ALMA REBECA IBÁÑEZ CARRANZA

MEXICO D.F. 2008

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CONTADURIA Y ADMINISTRACION

DATA WAREHOUSE

TESIS PROFESIONAL PARA OBTENER EL TITULO DE:
LICENCIADA EN INFORMATICA

PRESENTA:
ALMA REBECA IBÁÑEZ CARRANZA

ASESOR:
L.I. CARLOS FRANSICO MENDEZ CRUZ

MEXICO D.F. 2008

Agradecimientos

A mis padres Francisco Ibáñez Mariño y Mercedes Carranza, y a mi hermana Martha, en agradecimiento a su amor, comprensión, confianza y su gran apoyo incondicional que me ha permitido alcanzar varios logros entre ellos cerrar un ciclo más de mi vida exitosamente, mi formación profesional.

A mi asesor Carlos Francisco Méndez Cruz en agradecimiento por todo su tiempo que me otorgo en realizar este proyecto con toda la paciencia, comprensión, motivación y crítica constructiva.

A mis amigos en agradecimiento por su confianza, consejo y por ser una gran compañía en todo momento.

A la familia Montedónico García en agradecimiento por cariño brindado estos años.

A Israel Quiroz y Ramón Hernández que con su ejemplo me han ayudado a tener la motivación e inspiración en el transcurso de este ciclo.

Deseo expresarles mi respeto, admiración y agradecimiento sincero a todas las personas antes mencionadas y compartir este logro que constituye un aliciente para continuar con mi superación.

INDICE

Índice de ilustraciones

ANTECEDENTES

Actualmente, las organizaciones cuentan con una gran cantidad de información, por lo que la facilidad de acceso, la seguridad y la rapidez en la obtención de ella para su análisis en tiempo real se ha complicado, no teniendo la óptima productividad. La gran cantidad de información se debe en gran medida al incremento de las bases de datos de todas las áreas funcionales del negocio, ya sea por medio de datos históricos, aumento de servicios, clientes, etc.

El incremento de los datos ha causado conflictos en cómo manejarlos, sin embargo es necesario que las personas que requieran de la información la tomen de la forma más sencilla posible, sin necesidad de realizar procesos muy complicados.

Una característica de las diferentes áreas del negocio es que su información la manejan de diferente forma por medio de diferentes manejadores de bases de datos, por lo que en un nivel superior, como lo es una gerencia, al necesitar la información clara, concreta y en el mejor tiempo, se tienen conflictos para obtenerla.

Un Data Warehouse (DW) provee la posibilidad de extraer la información que se encuentra en varios manejadores de bases de datos sin tener que replicar la información. Éste unifica la forma de realizar las consultas de tal forma que todos los usuarios finales tengan acceso a la información de la forma más simple sin tener que saber programar, permitiendo la toma de decisiones en tiempo y forma.

A continuación les presento brevemente el contenido de esta tesis.
El capítulo primero habla sobre los conceptos básicos de lo que es un Data Warehouse, sus componentes, la importancia de un DW, las definiciones de varios autores de un Data Mart (DM) y las diferencias entre ambos.
El capítulo dos muestra las dos metodologías mas importantes para desarrollar un DW, éstas son la de Kimball y la de Inmon, así también se comparan para ver en qué difieren y qué similitudes tienen.
El capítulo tres, habla sobre las herramientas que se pueden usar para el diseño y construcción de un DW, ya sean de software libre o comercial.
El capitulo cuatro nos comenta la forma de construir un DM y el ciclo de vida de un proyecto de construcción de un DW.
El último capitulo es la parte más importante ya que es una pequeña aplicación de lo que se venía trabajando en los demás capítulos, se presenta el diseño de un caso dentro de la Facultad de Contaduría y Administración de la UNAM.

Importancia de un Data Warehouse (DW)

Un Data Warehouse existe para facilitar el soporte de decisiones dentro de la organización. Un sistema de soporte de decisiones ayuda a usuarios con el análisis *ad hoc* y a realizar una estrategia para tomar decisiones. Generalmente, el sistema de soporte de decisiones requiere de datos históricos, resumidos y a un nivel de transacción detallado. Los usuarios necesitan poder consultar cantidades masivas de datos fácilmente. (Albert citado en Breslin 2004:7)

Para un profesional en informática es importante saber este tema porque en el campo laboral actualmente dentro de las empresas lo más importante es la información que manejan, ya que con ésta se toman decisiones de alto impacto, por lo que él informático debe saber como explotarla, y saber qué es un DW, en qué ambiente se maneja, cómo diseñarlo, con qué herramientas se puede explotar y qué análisis se pueden realizar, para que el profesional pueda enseñar alternativas para solucionar conflictos con grandes cantidades de información almacenada.

Para una empresa los beneficios que trae implementar un Data Warehouse son los siguientes (Cf. Kimball 2002:3):

- La información de la compañía es fácilmente accesible.
- La información de la compañía es consistente.
- La información es adaptable y resistente al cambio.
- El Data Warehouse es el "Asegurador" que protege su información
- El Data Warehouse sirve como fundamento para la toma de decisiones.

Un Data Warehouse dentro de una organización es aplicable a todas las áreas del negocio. Actualmente se ha implementado en mercado bancario, al área de salud, área de seguros, para el comercio y en el sector de telecomunicaciones entre otras.

Objetivos

Una vez presentado el antecedente demos continuación a los objetivos que se pretenden alcanzar en este proyecto de tesis.

Objetivo General

. Crear un concentrado de información en español que trate el tema de Data Warehouse de una forma sencilla para que con base en este texto alguien pueda implementar uno, Además de ser una guía para capacitación en el tema de Data Warehouse.

Objetivos Específicos

- Determinar la diferencia entre un Data Warehouse y un Data Mart.
- Crear una guía de referencia para posibles necesidades de diseño y construcción de Data Warehouse.
- Plantear un caso dentro de la Facultad de Contaduría y Administración en la Coordinación de Informática, cuyas necesidades requieran una propuesta de Data Mart o Data Warehouse.
- Comparar las metodologías de los padres del Data Warehouse.
- Justificar la importancia de una DWH dentro de una organización ya sea privada o pública.
- Encontrar si hay software libre y comercial para la construcción de un Data Warehouse.

Hipótesis

Esta sección pretende plantear cuestiones ligadas a los objetivos previamente escritos para así llegar a los resultados esperados por medio de una metodología de trabajo.

He encontrado que el tema es principalmente enfocado a soluciones empresariales por lo que cuestiono sí, ¿es posible realizar el diseño de un DW dentro del área administrativa de una entidad académica? Propongo que sí es posible.
Por otro lado, es común encontrar en la bibliografía de Data Warehouse la pregunta ¿existen diferencias entre un Data Mart y un Data Warehouse?, por lo que la postulo como cuestión en este proyecto y también planteo que sí hay.
En una investigación previa, al iniciarme en el tema del DW, encontré que los principales autores son dos: Kimball e Inmon cada cual con su metodología por lo que me pregunto si ¿existen similitudes y diferencias entre ambos autores en cuanto al DW? Es de esperarse que sí las haya.
Hoy en día la tecnología esta rebasando muchos límites, la creación de software a aumentado, pero ¿existe un software libre o comercial que pueda ayudar al diseño e implementación de un DW? Mi hipótesis es que no existe software libre para DW.

Metodología de trabajo

Para la metodología a realizar en este proyecto de tesis se realizarán las siguientes actividades:

- Obtención de fuentes de información sobre el tema a tratar en libros, tesis, revistas y artículos de Internet.
- Obtención de información acerca del tema a tratar en distintas fuentes como libros, tesis, revistas y artículos de internet.
- Investigar las referencias de las tesis encontradas que puedan servir.
- Elaboración de resúmenes que ayuden a comprender el tema.
- Definir en dónde se va a realizar el diseño de una Data Warehouse
- Recolectar los requerimientos donde se diseñará la Data Warehouse
- Analizar toda la información obtenida
- Elaborar un plan de trabajo para el diseño y desarrollo de la DWH
- Entrevistas con los expertos y con los usuarios.
- Selección de la información obtenida
- Redacción de la información seleccionada

CAPITULO I INTRODUCCION AL DATA WAREHOUSE.

Lo que se va a tratar en este capítulo es una breve historia de la evolución de los sistemas para manejar la información, desde los manejados de archivos hasta la arquitectura de Data Warehouse. Además, abordaremos el concepto de Data Warehouse de diversos autores, el concepto de un Data Mart y algunos componentes de una Data Warehouse.

1.1 Breve Historia

A continuación revisaremos la evolución de los sistemas para almacenar información de forma cronológica y basándonos en Inmon. (2002:2-16)

En los años 60, se ejecutaban los programas con archivos maestros en el lenguaje COBOL. Con el incremento de estos archivos, se generó redundancia en los datos, ocasionando los siguientes problemas: la necesidad de sincronizar los datos sobre fechas; la complejidad de mantener los programas; la complejidad para desarrollar nuevos programas; y la necesidad de extender la cantidad de hardware para soportar todos los archivos maestros.

En los años 70, se comenzó un nuevo método para almacenar los datos llamado Dispositivos de Almacenamiento de Acceso Directo, (*DASD, Direct Access Storage Device*). El almacenamiento en disco tenía una diferencia básica con el almacenamiento en cintas magnéticas, ésta es que al consultar los datos se podían acceder directamente en ellos. Con los DASD, surgieron los sistemas manejadores de bases de datos (DBMS, Database Management Systems) que hacían fácil la programación para almacenar y acceder al DASD. En conjunto, los DASD y los DBMS solucionaron el problema de los archivos maestros.

A mediados de los años 70, el procesamiento de transacciones en línea (OLTP *On line Transaction Process)* hicieron más rápido el acceso a los datos, abriendo un nuevo panorama para negocios y procesos.

En los años 80 nuevas tecnologías surgieron, tales como la computadora personal *(Personal Computer)* y lenguajes de cuarta generación que hicieron qué el usuario final asumiera el papel correspondiente al especialista en procesamiento de datos. Anteriormente los datos y la tecnología eran exclusivamente usados para manejar decisiones operacionales detalladas, no cualquier base de datos servía para procesos operaciones transaccionales y procesos analíticos al mismo tiempo. Por esta época se implementaron los sistemas de información para la administración (MIS *Managements Information Systems*) que en conjunto con las nuevas tecnologías se pensaba que podían hacer más con los datos que simplemente hacer transacciones de procesos en línea como análisis estadísticos al mismo tiempo.

Poco después de que aparecieron los sistemas masivos OLTP a mediados de los 80, comenzaron a aparecer los programas de extracción. Los programas de extracción consistían en la obtención de algunos datos por medio de criterios de búsqueda dentro de una base de datos, al encontrar los datos calificados, se transportaban a otro archivo o a otra base de datos, se volvieron populares por poder mover los datos fuera de los caminos de las grandes representaciones de los procesos en línea y porque cuando se movían fuera del dominio de procesos-transacciones, el control del dato cambiaba y el usuario final se convertía en propietario.

Con el tiempo, las extracciones de datos se fueron incrementando, podían tener hasta 45000 extracciones al día, a esto se le llamó "*the naturally evolving architeture*" por lo que derivó en varios problemas: credibilidad de los datos, poca productividad y la imposibilidad de transformar los datos en información.

A finales de los años 80, varias empresas manejaban los llamados Sistemas de información ejecutiva (*Executive Information Systems*), pero comenzaron a crecer en complejidad a causa del valor de la información que proveían a los ejecutivos de las empresas. Para hacer más simple el proceso, otras empresas comenzaron a permitir, en los departamentos de tecnología, personal técnico orientado al negocio para acceder directamente a los datos usando un método de computación llamado lenguaje de consulta estándar (SQL *Structure Query Language*)

A comienzo de los años 90, existían dos formas de extracción de la información usando SQL; el primer camino era usando el QMF (*Query Management Facility*) de IBM y el PC/SQL-link de Micro Decisionware (el cual después fue adquirido por Sybase) El proceso QMF se creó en SQL, ejecutaba consultas, guardaba datos, exportaba el dato a un archivo, lo transfería del archivo a la PC y finalmente lo importaba del archivo a una hoja de cálculo.
A causa del volumen de información este proceso se tenía que reparar constantemente. Este proceso fue simplificado por el PC/SQL-link, el cual ejecutaba la consulta y realizaba la descarga automáticamente. A pesar de que ambos métodos resolvían varios problemas, requerían de pruebas y que la gente se capacitara causando una gran pérdida de tiempo. Por lo cual, se fueron desarrollando los procesos hasta llegar al Data Warehouse, que tenía sus bases de datos independientes del sistema funcional, esto significa que los datos son copiados desde el sistema funcional OLTP, limpiados, transformados y duplicados en la base de datos del Data Warehouse (Westerman 2001: 4-8).

1.2 Concepto de Data Warehouse

Existen varios autores que definen el Data Warehouse, a continuación menciono:

A single, complete and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use in a business context (DEVL97)

A data store for a large amount of corporate data (BIGU96)
The focal point for dissemination of information to end user for decision support and management reporting needs (SOFT95)
The data… and the process managers… that make information available, enabling people to make informed decisions (ANAH97).” (Mallach 2000:468)

En resumen estas definiciones de una Data Warehouse tienen los siguientes elementos:

- Consistencia de datos
- Almacenamiento de grandes cantidades de información
- El enfoque que se les da hacia el negocio
- Disponibilidad de los datos para el usuario final
- Permisible para el soporte de decisiones

Las siguientes definiciones corresponden al autor Kimball, que es uno de los precursores del Data Warehouse.

*“The conglomeration of an organization’s Data Warehouse staging and presentation areas, where operational data is specifically structured for query and analysis performance and ease-of-use.”(*Kimball 2002: 397*)*
“The conglomeration of an organization’s Data Warehouse staging and presentation areas. Others in the industry refer to the EDW as an centralized, atomic, and normalized layer of the Data Warehouse, without making it clear if such a system is available for end-user querying and drill-down. We discourage this interpretation of the EDW, preferring to think of the EDW as the largest possible union of staging and presentation services, taken as a whole.” (Kimball 2002:400)

El DW provee acceso a los datos de la organización, en él los datos son consistentes y pueden ser separados y combinados mediante cada indicador en el negocio. Es también un conjunto de herramientas para consultas y análisis. Es el lugar donde se publican las características de los datos usados que son conductores de la reingeniería del negocio.

A mi parecer la siguiente definición es la más completa:

“A Data Warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management’s decisions. The Data Warehouse contains granular coorporate data”.(Inmon 2002:31)

Ahora expliquemos los elementos de esta definición (Cf.. Inmon 2002:32-45):

Orientados a temas.
Anteriormente los datos eran orientados a aplicaciones de la compañía, por ejemplo en una compañía de seguros las aplicaciones pueden ser autos, vida, gastos médicos y daños, en cambio en este mismo ejemplo las mayores áreas del negocio pueden ser cliente, reclamos, políticas, premios, es decir, lo importante es que esté orientado a los aspectos de mayor interés para la organización, un DW almacena la información por temas y no por aplicación.

Integridad de datos.

Los datos son alimentados por múltiples recursos dentro del Data Warehouse. En el momento en que los datos son alimentados éstos también son convertidos, reformados, resumidos y así sucesivamente. El resultado de estos datos es que tienen una imagen corporativa común por lo que deben de mantener una consistencia en la estructura de tablas, archivos, nombres de los campos, tipos de datos y características de los atributos físicos.

No volátil
Cuando los datos son cargados y consultados se crea una foto la cual no es actualizada si no que se guarda generando datos históricos, dependiendo de los intereses de la organización pueden ser semanales, mensuales, etc.

Variante en el tiempo
Implica que cada unidad de datos en la Data Warehouse pertenece a un determinado momento del tiempo, para una organización tanto es importante los datos actuales como los datos históricos para la realización de análisis.

Granularidad
Se refiere al grado de detalle o resumen de las unidades de datos dentro del DW. Estos niveles de detalle van de lo general a lo particular en donde el nivel más bajo es el nivel más detallado.

1.3 Justificación de una DW

La necesidad de analizar los datos no es nueva, en el pasado era satisfecha por la labor de análisis manual, más recientemente por reportes programados desde archivos y bases de datos en *mainframes* (Mallach 2000:471)

Como en un inicio sólo comente los beneficios de un DW ahora en esta sección voy a explicarlos (Cf. Kimball 2002:3)

- En una DW la información debe ser fácil de acceder.
  El contenido de una DW debe ser comprensible, los datos deben ser intuitivos y casi obvios para los usuarios del negocio. Comprensibles implica legibilidad,, adicionalmente, debe ser fácil el manejo y el acceso para los usuarios y las consultas deben ser en un corto tiempo

- La información presentada debe ser consistente en un DW.
  Los datos deben tener credibilidad; deben estar cuidadosamente agrupados por medio de varios recursos alrededor de la organización; ser claros; seguros; y realizados sólo cuando lo requieran los usuarios. La información debe ser la misma para todos los usuarios, si dos medidas son interpretadas con el mismo nombre de igual forma deben de tener el mismo significado y en caso que no signifiquen lo mismo deben de estar etiquetadas de diferente modo. Consistencia en la información significa una alta calidad de la información.

- El DW debe ser adaptable y resistente al cambio.
  No debe evitar los cambios. El usuario necesita condiciones del negocio, datos y tecnologías que constantemente estén cambiando. El DW estará diseñado para ajustarse a los inevitables cambios; los cambios dentro del DW deben ser de tal forma que no invaliden los datos o las aplicaciones existentes. Las aplicaciones no deben ser cambiadas ni alteradas por ningún motivo cuando se agregan nuevos datos.

- El DW es un mecanismo que protege nuestra información.
  La información dentro de una organización es uno de los activos más importantes. El DW debe mantener confidencial la información de la organización.

- El DW sirve como fundamento para la toma de decisiones.
  El DW es el que tiene correctos los datos para soporte en toma de decisiones.

- El personal de la compañía que acepta el DW puede estimarse exitosa.
  Las personas en la compañía que aceptan el DW trabajan con él y pueden explotar toda la información que necesiten, contemplando todos los beneficios anteriores: fácil acceso, seguridad y consistencia para todos los datos.

Por otro lado tenemos que un DW tiende a crecer por los siguientes principales motivos: (Ken, Buss y Ryan 1998: 5-78)

El DW colecciona datos históricos. Los sistemas anteriores a él sólo miraban a los datos actuales, como tal, estos primeros sistemas operaban en una cantidad limitada de datos. Sin embargo, después de 5 a 10 años, los datos son coleccionados en el Data Warehouse ya que simplemente no hay manera de evitar la acumulación de grandes cantidades de datos

El DW implica la colección de datos para satisfacer requerimientos desconocidos. El diseñador de bases de datos debe acomodar ambos requerimientos los conocidos y los no conocidos. Para ello, el diseñador incorpora, datos ajenos y no evidentes dentro del Data Warehouse, esto necesita una cantidad considerable de almacenamiento para acomodar los datos requeridos potenciales y desconocidos ya sean usados o no.

El DW incluye datos tanto a nivel detallado como a nivel resumen. La necesidad de acomodar los detalles y los resúmenes de los datos conduce a la acumulación de largas cantidades de datos.

El DW contiene datos externos. Gran cantidad de datos externos son recolectados para soportar variedad de actividades de minería de datos. Por ejemplo, las herramientas de minería de datos que usan datos externos para predecir cual es probablemente el mejor cliente.

1.4 Definición de Data Mart

A continuación comentaremos algunas definiciones de Data Mart para la comprensión del tema, además de explicar los tipos de Data Mart

Data Mart is a logical subset of the complete Data Warehouse. (Kimball, Reeves, Ross, Thornthwaite 1998:18)

Data Mart a departmentalized structure of data feeding from the data warehouse where data is denormalized based on the department's need for information. (INMON 2002:389)

Data Mart *A logical and physical subset of the Data Warehouse's presentation area. Originally, Data Marts were defined as highly aggregated subsets of data, often chosen to answer a specific business question. This definition was unworkable because it led to stovepipe Data Marts that were inflexible and could not be combined with each other. This first definition has been replaced, and the Data Mart is now defined as a flexible set of data, ideally based on the most atomic (granular) data possible to extract from an operational source, and presented in a symmetric (dimensional) model that is most resilient when faced with unexpected user queries. Data Marts can be tied together using drill-across techniques when their dimensions are conformed. We say these Data Marts are connected to the Data Warehouse bus. In its most simplistic form, a Data Mart represents data from a single business process.(Kimball 2002:396)*

A Data Mart is a smaller version of a Data Warehouse, typically containing data related to a single functional area of the firm or having limited scope in some other way. (Mallach 2000: 469)

En resumen, para Inmon, un Data Mart (DM) es una estructura departamental que es alimentada desde el DW, donde los datos están basados en las necesidades de la información de los departamentos o áreas del negocio. Para Kimball es un conjunto flexible de datos idealmente basados en los posibles datos atómicos extraídos desde una fuente operacional y presentada en un modelo sistemático que es resistente a las consultas inesperadas de los usuarios, anteriormente en conjunto con Reeves, Ross y Thornthawaite decían que un DM es un subconjunto lógico de un DW. Para Mallach es una versión pequeña de un DW.

Generalmente los DM se clasifican en dos tipos, Unos autores dividen entre dependientes e independientes y otros lo manejan como multidimencionales y de propósito general.

Los tipos de Data Mart, dependientes e independientes, se muestran en la siguiente figura. La categorización está basada principalmente en las fuentes de los datos que alimentan el Data Mart. Los DM dependientes extraen los datos desde un Data Warehouse central que ya se ha creado previamente. Los DM independientes en contraste son sistemas independientes construidos por medio de la extracción de datos directamente desde las fuentes de datos operacionales o externas.

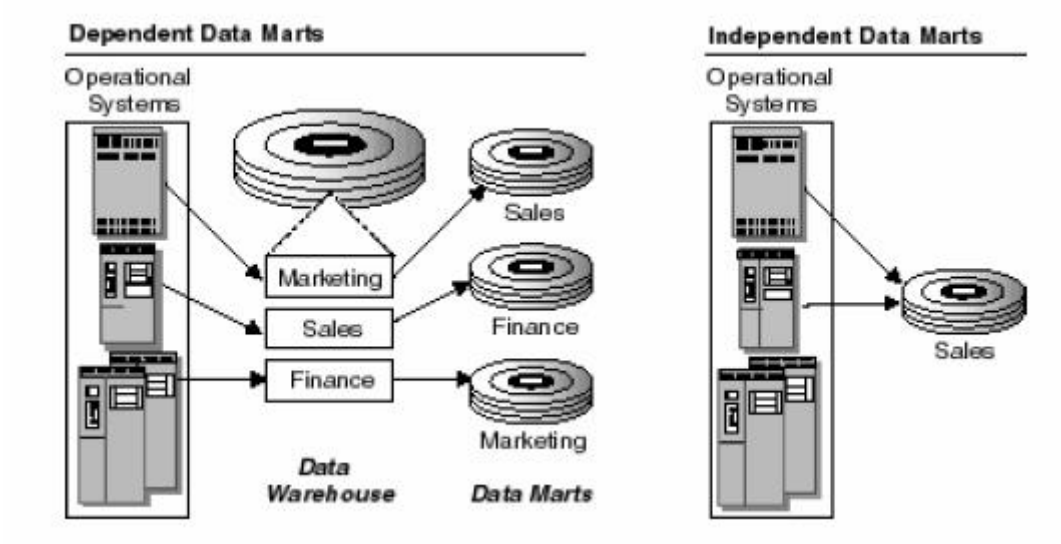


Ilustración 1 Clasificación de Data Mart (Oracle Data Mart Suite 1999:1-2)

Con los Data Mart dependientes, el proceso es simplificado, porque los datos resumidos y formateados ya han sido cargados en el Data Warehouse centralizado. El proceso de ETL (el cual se mencionará con mayor detalle en los siguientes capítulos). Para Data Marts dependientes es principalmente el proceso de identificar los subconjuntos correctos de datos relevantes para el subconjunto escogido del Data Mart y mover una copia quizá en forma resumida. Un Data Mart dependiente es usualmente construido para lograr un mejor rendimiento y disponibilidad, un mejor control, y una reducción de los costos resultantes de las telecomunicaciones de acceso local de los datos correspondientes a un determinado departamento.
Con los Data Mart independientes se tiene que llevar a cabo todos los aspectos del proceso de ETL como se debiera hacer para un Data Warehouse centralizada. El número de recursos son comúnmente menores que un Data Warehouse. Un Data Mart independiente es construido por la necesidad de disponer de una solución en un corto tiempo. (Oracle® Data Mart Suite 1999:1-3)

Por otro lado tenemos otra clasificación de los Data Marts estas son de bases de datos multidimensionales y el de propósito general. El primero, es usado como soporte de habilidades gerenciales y analíticas para mirar a los datos en diferentes formas. Algunas características de este tipo son: matrices escasamente pobladas, datos numéricos, estructura rígida de datos una vez que entra en la base de datos multidimencional, consistente y veloz tiempo de respuesta.
El otro tipo de Data Mart es llamado de propósito general éste contiene tanto datos numéricos como de texto. Sirve a una mayor audiencia que el anterior, los Data Mart de bases de datos multidimensionales son soportados por bases de datos especializadas en sistemas manejadores, los de propósito general son soportados con tecnología relacional. Algunas características son: soporta números, texto y otras formas de datos; soporta análisis de propósito general, soporta a estructuras libres de datos, tiene la habilidad de tener numerosos índices y soporta esquemas de estrella.(Ken, Buss y Ryan 1998:82)

El usuario del ambiente de los Data Marts puede ser llamado analista departamental, éste es un individuo que toma decisiones sobre tendencias departamentales de forma estratégica en periodos largos o cortos, no es un técnico pero es una persona del negocio ante todo.(Ken, Buss y Ryan:1998,77)

1.5 Componentes de un DW

Para entender los componentes a continuación se muestra la estructura de un DW

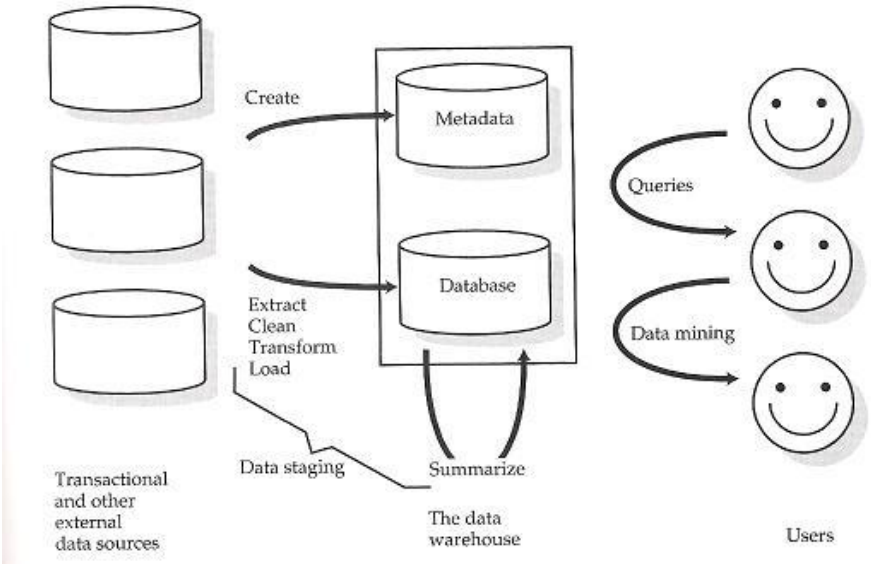


Ilustración 2: Componentes de un DW (Mallach 2000:473)

Los componentes que se incluyen en esta estructura son los siguientes (Mallach 2000:473-474)

1. Basas de datos Transaccionales y otros recursos externos. Son aquellas que alimentan la Data Warehouse, éstas toman una copia de los datos transaccionales, pero no los almacenan directamente.
2. El proceso de extracción. Extrae los datos de diversas bases para traerlos a la Data Warehouse. Este proceso debe siempre transformar los datos en la estructura de una base de datos y formatos internos de la Data Warehouse.
3. El proceso de limpieza de datos. Para estar seguros que tiene suficiente calidad para los propósitos de toma de decisiones se usa la limpieza.
4. Proceso de carga de datos limpios en la base de datos de la Data Warehouse. Se refiere a la colectividad como almacén de datos.
5. Resumen de datos. Cualquier dato precalculados que genera un total, un resultado.
6. Metadato Es útil cuando se tiene información centralizada en un repositorio. Dice al usuario qué hay en la Data Warehouse, de dónde viene, quién es el responsable y más detalles.
7. La base de datos del Data Warehouse. Contiene los detalles de la base de datos y la suma de datos de la Data Warehouse

8. Herramientas de consulta. Normalmente incluye una interfaz de usuario para procesar cuestiones de la base de datos, en un proceso llamado OLAP (*On line analytical processing).* Estos incluyen herramientas automatizadas para descubrir datos, esto es *data mining*
9. Por ultimo tenemos al usuario. Para el cual fue hecha la Data Warehouse, es él quien la usa.

Esta misma estructura aplica para un Data Mart, aunque cada componente individual debe ser más modesto en cuanto al alcance. Un Data Mart típicamente tiene pocos recursos porque detalla sólo un área, requiere de menos hardware, requiere de menos capacitación y soporte. Un Data Mart es una proposición fácil a comparación de un completo Data Warehouse.

1.6 Procesos básicos de un DW

Los procesos básicos de un DW son los siguientes: Extracción, transformación, carga e indexado y revisión de control de calidad.
Extracción: significa leer y entender las fuentes de datos y copiar las partes que requiere el área de almacenamiento de datos.
Transformación: una vez extraídos los datos, sigue una serie de pasos que son: la limpieza de los datos (corrigiendo los campos vacíos, resolviendo conflictos de dominio y análisis dentro de los formatos estandarizados), la depuración de los campos seleccionados que no usara el DW, la combinación de los recursos por medio de los valores clave, la creación de llaves alternas para evitar una herencia de llaves y la edificación de consultas comunes.
Indexado y Carga: después de la transformación de los datos, usualmente se colocan las tablas de hechos y de dimensiones en recipientes de Data Marts, cada Data Mart es indexado.
Revisión de control de calidad: al finalizar el paso anterior se debe correr un reporte para revisar que el proceso se haya hecho con las reglas de calidad establecidas.
Después de haber obtenido una apreciación positiva del reporte de control de calidad deben dar aviso a la comunidad de usuarios que los datos están listos para ser consultados. Cada usuario tal vez requiera de una actualización, los Data Mart frecuentemente requieren de ser actualizados, para ello se debe tener un adecuado control por lo que los datos incorrectos deberán ser corregidos.
Las consultas son las tablas en términos que abarca completamente las actividades y requerimientos de datos del Data Mart incluyendo consultas ad hoc realizadas por los usuarios finales, reportes escritos, aplicaciones complejas de soporte a las decisiones, y respuestas de modelos. También, retroalimentación de datos y reproceso, auditoria, seguridad y respaldos son procesos que se realizan en un DW. (Kimball, Reeves,Ross,Thornthwaite 2000:23,24)

1.7 El DW en diversas industrias

Comercio minorista

Algunos de los negocios que se incluye en esta clasificación son farmacias, almacenes, negocios de especialidades, catálogos (venta por correo), comercio de ramos generales, servicios de comida, tiendas virtuales en Internet (*e-tailing*) y bienes de consumo.

Actualmente el comercio minorista esta en continua innovación, varios sectores tienen la atención en el *e-commerce*. Por lo que el Data Warehouse se ha vuelto un arma fundamental.

Los hombres de negocio del sector minorista combinan datos a nivel de transacción de sus puntos de venta con otros tipos de datos y realizando análisis que pueden servir para una variedad de actividades comerciales. Los usuarios de un Data Warehouse en este tipo de industria son los vendedores compradores gerentes de tienda, gerentes de producto, personal de depósito y distribución, analista de marketing y ejecutivos.

Dentro del comercio minorista se realizan análisis detallado de compras, ya que es muy complejo poner en los estantes el volumen optimo de productos; también se llevan a cabo análisis de afinidades del producto por ejemplo el remedio para un resfrío puede ser con jugos o aspirinas, los datos que obtenemos de esos análisis en este caso son los porcentajes de veces que se compra el producto y el promedio de dinero del producto entre otros factores.

La obtención de estos factores de los diversos análisis son usados para definir estrategias de venta por ejemplo poner productos básicos como pan y leche alejados entre sí ya que el cliente tendría que recorrer toda la tienda y a lo largo de su trayecto elegir otros productos. Además los minoristas pueden negociar con mayor efectividad con sus proveedores.

El Data Warehouse también simplifica la practica conocida como "cotización diferencial" al permitir a las compañías cotizar ciertos productos de acuerdo a diferentes situaciones demográficas o segmentos de clientes. Los minoristas que tienen acceso a datos de sus clientes pueden fijar precios con mayor precisión.

El historial de ventas de un producto es un indicador de su movimiento futuro, por lo que si se examina el historial se puede determinar los segmentos del mercado para predecir cuanto tiene que pedir más en el stock de un producto. El Data Warehouse por lo tanto es en la industria de bienes de consumo es el que puede proveer de datos importantes acerca de pedidos y reabastecimiento. Además, algunos minoristas optan por el reabastecimiento automático o en el *e-commerce* de negocio a negocio, es decir interconectar sus sistemas de pedidos con los de sus proveedores a fin de permitir la entrega "justo en tiempo".

Servicios financieros.

Esta industria abarca una amplia gama de compañías, por ejemplo bancos, compañías de ahorro y préstamo, compañías de tarjetas de crédito, agentes de bolsa y bancos de inversión.

En los últimos años, la mayoría de los bancos adoptaron la noción de "banca minorista", la banca mayorista por su parte se concentra en clientes comerciales y es claro que estos son los clientes de mayor impacto en los bancos ya que generan mayores ingresos.

Las firmas de servicios financieros combinan detalles de transacciones, información de productos, y datos acerca de los clientes y la historia de sus cuentas para realizar los siguientes tipos de análisis.

Análisis de rentabilidad.
La mayoría de los bancos realizan rigurosos programas de costos de actividades necesarios para resolver los costos individuales de los productos y así asegurar su rentabilidad o comprando costosos productos de software para determinar la rentabilidad de clientes y grupos familiares.
La rentabilidad del cliente es la forma más codiciada de información sobre rentabilidad por una muy buena razón: una compañía no puede ser el verdadero valor de un cliente, el banco está formado a adivinar el modo de tratarlo.
Los bancos pueden relacionar los resultados de los análisis de rentabilidad de clientes con los canales parar determinar sucursales de alto valor a los que conviene apuntar. Los cambios en la rentabilidad de un cliente a través del tiempo también pueden revelar patrones de comportamiento de los clientes previamente ignorados.

Gestión de riesgo y prevención de fraude
Algunos bancos administran riesgos otros tratan de los evitarlos. Un Data Warehouse le brinda a una compañía bancaria un enfoque científico de la gestión del riesgo.
La gestión del riesgo puede individualizar mercados o segmentos de clientes específicos que pueden ser de mayor riesgo que otros, y a un nivel más detallado puede determinar los factores de riesgo de individuos específicos. Existen varias herramientas de software que especializadas que ayudan a la realización de gestión de riesgo.
Los tipos más avanzados de gestión de riesgo y detección de fraude permiten la reevaluación continua del comportamiento del cliente. Al evaluar cómo cambian a lo largo del tiempo el fraude y los patrones de mora, y además determinar qué tipo de clientes y cuentas son merecedoras de ofertas de crédito preactivas, los bancos pueden tanto prevenir pérdidas de ingresos como afinar sus actividades de marketing.

Análisis de propensión y marketing dirigido por eventos
Es posible predecir la propensión de un cliente a cometer fraude. Este tipo de análisis predictivo se usa también para generar ingresos adicionales significativos.
El análisis de propensión a la compra ayuda a los bancos a reconocer si es probable que determinado cliente compre un producto o servicio dado e, incluso, cuándo podría ocurrir dicha compra. Al saber que productos atraen a qué clientes los bancos pueden reducir sustancialmente sus costos de marketing y además prevenir que los clientes se pasen a otros bancos en busca de productos similares.

Modelado de respuesta y de duración
Este tipo de análisis avanzado puede decirle a un banco la probabilidad de que un cliente responda a una promoción dada y compre un producto o servicio publicitado. Puede predecir, en caso del modelado de duración, el tiempo en que un cliente nuevo responda a una promoción o un cliente ya existente cuanto tiempo se quedará. El modelo de duración en verdad es útil para compañías de tarjetas de crédito que ofrecen tarjetas de afinidad (tarjetas de crédito que permiten al usuario enviar una pequeña suma de dinero a una organización caritativa cada vez que se usa la tarjeta) a clientes o segmentos específicos, además de la frecuencia de uso y el tiempo de conservación.

Análisis y planificación de distribución

Al usar un análisis cuidadoso de los diversos canales de distribución permite a los bancos tomar de decisiones sobre la organización interna de las sucursales, incrementos, reducciones de personal, agregado de nuevas tecnologías o el cierre o unificación de las sucursales de bajo tráfico.
Combinando la actividad del canal con los datos de rendimiento del canal tomados del Data Warehouse, las firmas de servicios financieros pueden inducir a sus clientes a cambiar a canales de distribución de bajo costo mediante la reducción de aranceles o la provisión de servicios adicionales.

Telecomunicaciones

Los Data Warehouse pueden transformar las telecomunicaciones de la misma manera que el comercio minorista: cambiando de raíz los procesos comerciales y los puestos de trabajo y convirtiendo el análisis de datos en una "habilidad básica" de la industria.
El grado con el que las compañías de telecomunicaciones están realizando análisis avanzados mediante sus Data Warehouse es proporcional al segmento de mercado al que sirven. A diferencia de la mayor parte de las otras industrias, los segmentos individuales de mercado en la industria de telecomunicaciones definen el tipo de nivel de análisis de soporte de decisión que se realiza.

La industria de telecomunicaciones ha pasado de una mentalidad de servicios a una mentalidad minorista, lo que implica que los tipos de análisis que las compañías emprenden han evolucionado hasta perecerse a los que se ven en el comercio minorista y en la banca.

La regla practica del análisis de Churn (la fuga de clientes), en las telecomunicaciones es que cuanto mejor establecido este el portador, más probable es que ya exista una solución al *churn.* El Churn se refiere a aquellos que abandonan un proveedor de servicio, usualmente por otro. Por lo tanto el término *churn* puede implicar tanto la predicción como el análisis *post facto.* El Churn de productos significa la cancelación de un producto o servicio, a veces a favor de otro.

El fraude siempre ha sido un problema para las compañías de teléfonos, y las tecnologías de fraude aparecen a la misma velocidad que las nuevas tecnologías de comunicación. Para solucionar este problema tenemos dos tipos de análisis de fraude: proactivo, el que la empresa es capaz de detectar antes de que ocurra el fraude, y reactivo, en el que la compañía toma medidas para lidiar con los fraudes ya cometidos. Las herramientas de *data minig* pueden predecir el fraude detectando patrones en la información consolidada del cliente y los registros detallados de llamadas.

Gobierno

Desde historias detalladas de los contribuyentes hasta "incidentes" de tráfico o patrones climáticos globales, el gobierno federal viene confiando desde hace algún tiempo en el *Data Warehouse* para tener el país en marcha.

Los gobiernos estatales y locales también emplean sus Data Warehouses en diversas formas. Algunas razones que los guiaron a seguir en líneas innovadoras de conocimiento fueron:

- Algunos funcionarios electos, muchos de los cuales propusieron una plataforma de rebaja de impuestos, ahora necesitan medidas de reducción de gastos.
- El gobierno federal ha impuesto nuevas atribuciones a los gobiernos locales y estatales.
- Hay recortes de presupuestos
- Los delitos han ido en aumento y se requiere de la información como factor determinante en la resolución de estos delitos.
- Los contribuyentes demandan un retorno de su inversión, insisten en un mejor nivel de servicios con mayor velocidad de respuesta.

El *Data Warehouse* en el gobierno no es una tarea administrativa seria y aburrida que uno podría imaginar. En estados repartidos a lo largo de Estados Unidos, los utilizan en departamentos de vehículos automotores, ejecución de la ley y castigo, recaudación de impuestos, administración de aeropuertos, control de servicios públicos de salud, presupuesto e inversiones en capital, administración de contratos, con proveedores externos, departamentos de obras públicas y caminos, servicios, agencias forestales y de incendios, servicios de asistencia social y administración de infraestructura de ciudades y condados.

Salud.

La industria de salud abarca instituciones públicas, privadas y sin fines de lucro, incluyen las siguientes empresas: hospitales, farmacias y compañías de distribución de medicamentos, organizaciones de medicina, consultorios y clínicas privadas y compañías de seguros.

Las compañías de atención de la salud están usando los datos en *Data Warehouses* para una variedad de propósitos: almacenamiento y análisis de registros de pacientes, administración de casos, gestión de riesgo, creación de perfiles de proveedores y control de proveedores, seguimiento de quejas, seguimiento y procesamiento de contratos, funciones actuariales y de aseguración, análisis de tasas, control de tendencias en la calidad del tratamiento, análisis de calidad/gastos, análisis de ganancias y pérdidas, análisis de reclamos y relaciones con los pacientes.

Los hospitales y compañías de seguros analizan los reclamos para determinar los precios para ciertos clientes o proveedores, nuevas tasas de renovación de contratos, cálculos de reembolsos esperados y compararlos con los pagos reales para descubrir casos en los que se haya pagado menos de lo debido.

La gestión del riesgo permite a las organizaciones de atención de la salud continuar el proceso de administración de reclamos con el seguimiento de informes de incidentes, con el objetivo de evitar fraudes por parte de los pacientes.

Los avances en la tecnología de soporte de decisiones específicas de atención de la salud permiten el uso de los datos del paciente para seguir cada caso a lo largo del

continuo del tratamiento. La respuesta de los pacientes con episodios agudísimos (desde la cirugía y la convalecía hasta la atención a domicilio) ofrece a los administradores del hospital y equipos de tratamiento veloces datos acerca de las tendencias que muestra la respuesta de los pacientes al tratamiento.

Seguros

En la mayoría de las compañías de seguros el grueso de la información lo construyen los datos sobre reclamos, y el análisis de reclamos sigue siendo la motivación comercial con mayor prioridad.

Otros análisis que se realizan en esta industria a partir del Data Warehouse son: gestión de riesgo, análisis costo y precio de productos, marketing directo, desgaste, venta cruzada, empaquetamiento de productos, gestión de campaña, perfiles de proveedores, detección de fraude, entre otros.

El desafío de las compañías de seguros básicamente son: controlar sus costos, mantener su cuota de mercado y mejorar el servicio al cliente y la imagen.

Los informes que usan a partir del Data Warehouse algunas compañías de seguros son:

- Informes ejecutivos de reclamos, que incluyen resúmenes de la cantidad y tipo de los reclamos y los pagos durante un periodo de tiempo determinado
- Comparaciones de reclamos contra los promedios de la industria
- Seguimiento de reclamos para ciertos prestadores de servicios, comparándole historial de pagos y el volumen con la de otros prestadores o canales de venta.
- Medición de satisfacción de clientes, obtenidas usando los datos de reclamos como una ayuda para la realización de encuestas a los clientes.

1.8 Diferencias entre un DW y un DM

Para que quede un poco más clara las definiciones a continuación mostrare algunas diferencias entre los conceptos de DW y DM.

Un Data Warehouse, en contraste, trata con múltiples temas y es típicamente implementado y controlado por una unidad central organizacional como es el grupo *the Corporate Information Technology* (IT). Generalmente, es llamado central o *enterprise* Data Warehouse. Típicamente, un Data Warehouse reúne datos desde múltiples fuentes de sistemas. (Oracle® Data Mart Suite 1999:1-2)

Nada de estas definiciones básicas limitan el tamaño de un Data Mart o la complejidad de toma de decisiones de los datos que contiene, aunque un Data Mart es típicamente más pequeño y menos complejo que un Data Warehouse; por lo tanto son típicamente más fáciles de construir y mantener.

La siguiente tabla muestra las diferencias entre un DW y un DM

	Data Warehouse	Data Mart
Scope	Corporate	Line-of-Business (LoB)
Subjects	Multiple	Single subject
Data Sources	Many	Few
Size (typical)	100 GB - TB+	< 100 GB
Implementation Time	Months to years	Months

Ilustración 3: Cuadro comparativo DW y DM (Oracle Data Mart Suite 1999:1-2)

Un Data Mart no sustituye a un Data Warehouse, anteriormente se pensaba que sí porque para la construcción de un Data Warehouse se requería de un gran esfuerzo y altos costos.

La estructura de un Data Mart para un departamento es diferente a la de otro, esto se debe por las necesidades de cada departamento en particular, los datos que alimentan el Data Mart de cada departamentos de una organización son extraídos, dependiendo el tipo de Data Mart, del Data Warehouse o de las bases de datos operacionales.

La estructura de un Data Mart no es reusable, ni flexibles, no es usable como fundamento de reconciliación y tampoco están listas para un nuevo conjunto de requerimientos desconocidos.

En síntesis un DM nos sirve cuando tenemos poca información o la tenemos separada por departamentos llámese por poner un ejemplo el departamento de finanzas el departamento de mercadotecnia, etc. Un DW es más viable cuando tenemos una gran cantidad de datos que manejar y los procesos de extracción y análisis de la información son más complejos, podemos decir que un conjunto de Data Marts engloban a un DW. (Inmon 2002:142)

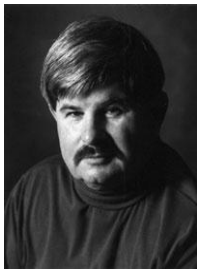
CAPITULO II INMON & KIMBALL

En este capítulo vamos a ver unos breves datos biográficos de estos personajes quienes son parte importante del concepto de Data Warehouse, comentaremos sobre sus esquemas, modelos y filosofía, concluyendo con una breve comparativo.

2.1 Breves datos Biográficos

Considero que los dos personajes dentro del tema de Data Warehouse más importantes son William Inmon y Ralph Kimball. En los siguientes párrafos comentaré sobre sus biografías.

2.1.1 William Inmon (WEB:00)



William H. Inmon es llamado el padre del Data Warehouse por ser reconocido como experto y creador del Data Warehouse. Es el autor de "*The Corporate Information Factory*".

Tiene 35 años, aproximadamente, de experiencia en tecnologías de manejadores de Base de Datos y en diseño de Data Warehouse, es conocido globalmente por sus seminarios en esta materia. Ha sido un ponente clave en varias organizaciones de computación y ha dado varias conferencias en las industrias, seminarios y en distintos eventos.

Ha escrito alrededor de 650 artículos sobre varios tópicos referentes a la construcción, uso y administración de un Data Warehouse y sobre el contenido de su libro "*The Corporate Information Factory*". Sus trabajos han sido publicados en diversas revistas de computación incluyendo *Data Managment Review y The Bussines Intelligence Network,* donde él continua siendo columnista. Ha escrito 46 libros, varios de ellos traducidos en 9 idiomas.

Como empresario, fundó Prism Solutions en 1991. En 1995, fundó Pine Cone Systems, después llamado Ambeo. En 1999, creo su Web Site para educar a profesionales y para tomar decisiones sobre Data Warehouse y *The Corporate Information Factory*. En 2003, cofundó *Inmon Data Systems, Inc.* y creó *The Government Information Factory.*(WEB:01)

Ha trabajado para *American Management Systems, Inc. y Coopers & Lybrand.* Estudió ciencias matemáticas en la Universidad de Yale y su maestría en Ciencias de la Computación por la universidad de Nuevo México.

2.1.2 Ralph Kimball (WEB:02)

Es conocido mundialmente como innovador, escritor, educador, expositor y consultor en el campo del Data Warehouse. Sus libros en técnicas de diseño dimensional son considerados entre los más vendidos en el tema de Data Warehouse. Al día de hoy ha escrito más de 100 artículos y columnas para Intelligent Enterprise y ha sido ganador del Readers' Choice Award 5 años seguidos.

Después de recibir el Ph. D. en 1972 por Stanford en ingeniería electrónica (con especialización en sistemas hombre-máquina), Kimball se hizo socio de Xerox Palo Alto Research Center (PARC). En PARC coinventó the Xerox Star Workstation, el primer producto comercial que usa ratón, iconos y ventanas.

Tiempo después se convirtió en vicepresidente de aplicaciones de Metaphor Computer Systems, pionero del software para toma de decisiones y proveedor de servicios. Desarrolló el llamado *The Capsule Facility* en 1982. *The Capsule* era una técnica de programación gráfica que conectó iconos juntos en un flujo lógico, permitiendo un estilo muy visual de la programación para los no programadores.

Kimball fundó *Red Brick Systems* en 1986, sirviendo como CEO hasta 1992. Dicha compañía, ahora poseída por IBM, era conocida por su base de datos con optimización para el almacenamiento de los datos.

2.2 Modelos de Data Warehouse

2.2.1 Esquema de un Data Warehouse de William Inmon

Inmon divide el ambiente de las bases de datos organizacionales en cuatro niveles, los tres últimos componen la arquitectura de ambiente para un Data Warehouse. En el siguiente cuadro se ilustran:

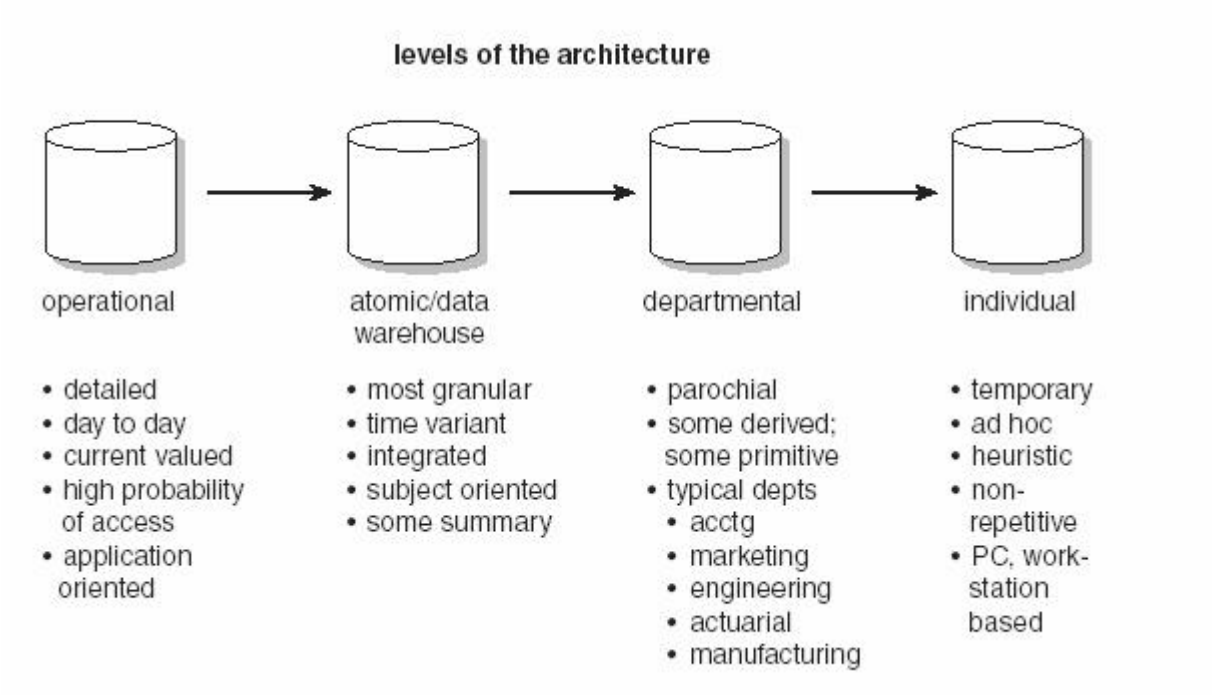


Ilustración 1: Niveles de la arquitectura de Inmon (Inmon 2002:16)

- El nivel operacional de datos contiene aplicaciones orientadas a datos primitivos y otros sistemas de procesos transaccionales. Este nivel soporta las operaciones diarias dentro una organización, en esta capa los datos son manipulados y se mueven a la siguiente capa.
- El nivel de Data Warehouse contiene los datos primitivos integrados, son datos históricos que no pueden ser actualizados, además ciertos datos derivados también se encuentran aquí.
- El departamental o nivel de Data Mart contiene datos derivados casi exclusivos. Este nivel es formado por requisitos del usuario final en forma específica que satisfacen las necesidades del departamento o área.
- El nivel individual es donde se realizan análisis heurísticos de los datos.

Para que quede más clara la arquitectura, Inmon pone un ejemplo:

En el nivel operacional hay una entidad que es un cliente, en este caso llamado J. Jones, donde su registro contiene valores que pueden ser actualizados en cualquier momento y que muestran su estatus actual, es decir, si cambia el nivel operacional, el registro tendrá que cambiar y reflejar el dato correcto de su información .

En el siguiente nivel encontramos datos históricos del cliente, los créditos que ha tenido por año, la suma anual, cada año en un registro, los cambios que ha realizado en su domicilio, etcétera.

En el nivel departamental o Data Mart encontramos la información interesante para los diferentes departamentos dentro la organización crediticia a la que pertenece, como lo son los montos por mes del crédito.

Y por último, el nivel individual donde vienen datos estadísticos de sus cuentas como son el año de inicio de su crédito, el monto promedio de crédito etc.

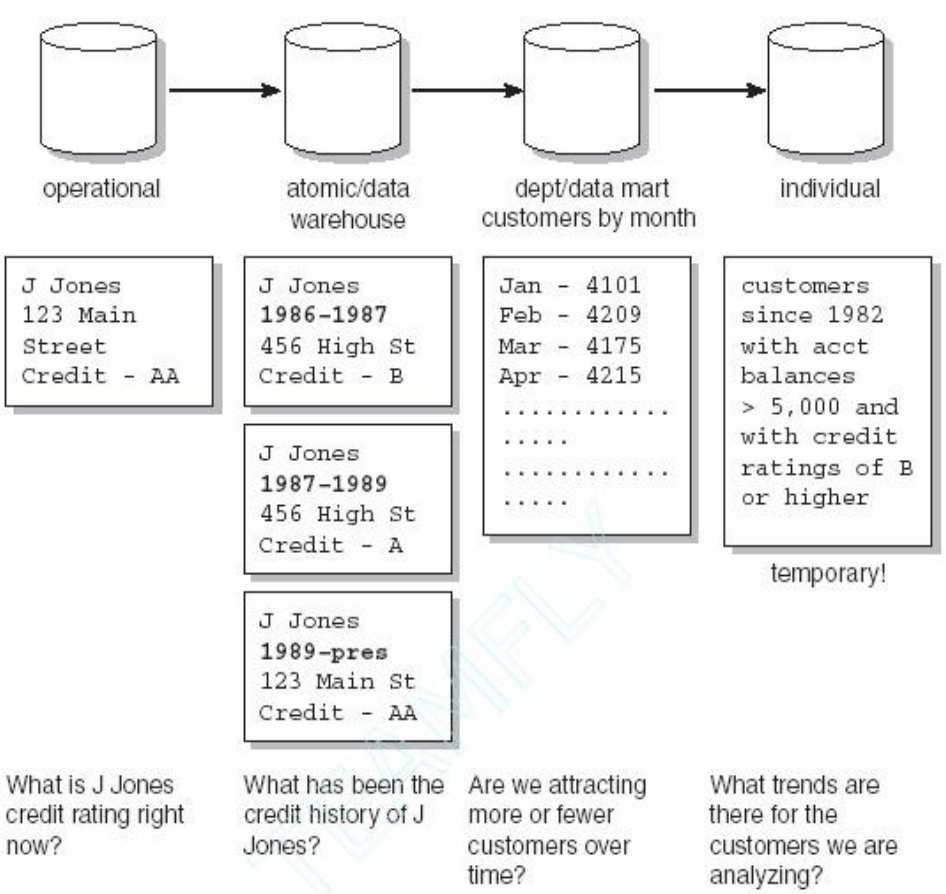


Ilustración 2: Ejemplo de la arquitectura de Inmon (Inmon 2002:18)

2.2.1.1 Modelos de Data Warehouse de INMON

Existen tres niveles para el modelado de datos según Inmon. Estos son: el nivel alto de modelado (llamado ERD *Entity Relationship Level*), el nivel medio (llamado DIS *Data Item Set*) y el bajo nivel de modelado (llamado también modelado físico).

El Nivel alto de modelado o ERD representa las entidades y sus relaciones, el nombre de las entidades son rodeadas por un ovalo. Relaciones entre entidades son representadas con flechas. La dirección y el número de cabezas de flechas indican la cardinalidad de la relación, y sólo las relaciones directas son indicadas. Por lo tanto, las dependencias transitivas son minimizadas.

El nivel medio del modelado incluye cuatro construcciones:

- El grupo primario de datos. Existe sólo una vez para cada área principal. Contiene atributos y llaves que existen sólo una vez por área.
- El grupo secundario de datos. Contiene atributos de los datos que pueden existir múltiples veces para cada área principal.
- El conector. Relaciona grupos de datos entre ellos.
- El tipo de dato. Es indicado por una línea punteada a la derecha del grupo de datos, el grupo de datos de la izquierda es un supertipo y el grupo de datos de la derecha es un subtipo.

En el siguiente diagrama se muestran las cuatro construcciones del nivel medio de modelado.

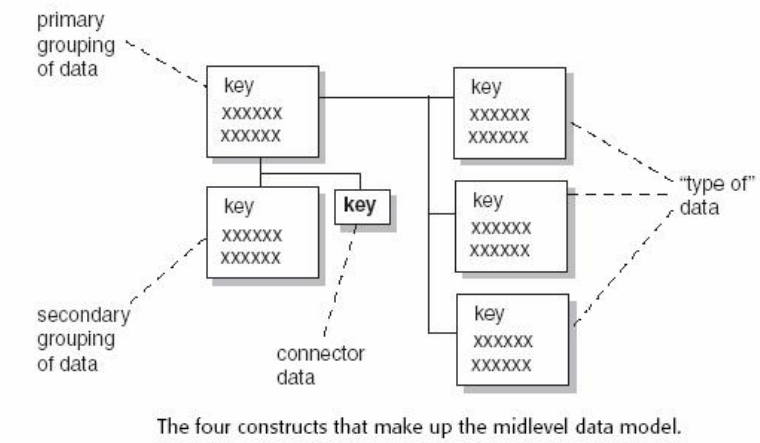


Ilustración 3: Construcciones del nivel medio de modelado (Inmon 2002:95)

El modelado físico es creado a partir del modelo medio de datos, incluye llaves y características físicas, aparece como una serie de tablas llamadas tablas relacionales.
Podría decirse que el modelo físico es un paso anterior al diseño de la base de datos, aunque se recomienda la granularidad y el particionamiento de datos antes de definir un diseño en la base de datos.

La granularidad consiste en el proceso de hacer que los datos estén en el máximo nivel de detalle en que se puedan encontrar los datos, a mayor detalle más bajo nivel y a menor detalle más alto nivel, por ejemplo, una simple transacción puede estar en el nivel más bajo de detalle mientras que un resumen de todas las transacciones mensuales puede estar en un nivel alto de detalle (véase Figura(x- a)).
La granularidad es el principal tema del diseño en un ambiente de Data Warehouse porque afecta profundamente el volumen de datos que reside en él y en el tipo de respuesta de las consultas. Los datos deben de tener un nivel alto de granularidad dentro de un Data Warehouse. (Inmon.2002:43)

El particionamiento se refiere a la desintegración de los datos en unidades físicas que pueden ser manejadas de forma independiente. Un particionamiento adecuado puede tener varios beneficios para el Data Warehouse en diferentes maneras: en la carga, acceso, limpieza, monitoreo y almacenamiento de los datos. El propósito de particionar el actual detalle de los datos es fragmentar en pequeñas unidades para que sea más manejable como se ilustra en la Ilustración (7-b) (Inmon.2002:56).

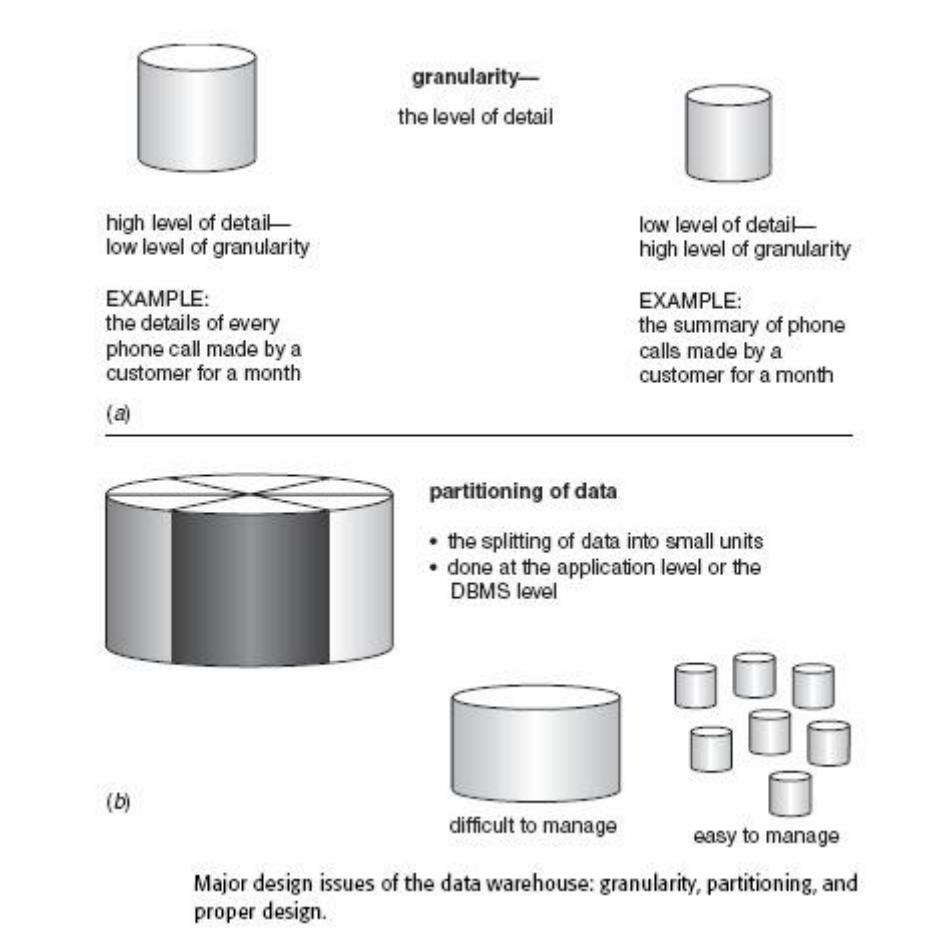


Ilustración 4: Granularidad y Particionamiento (Inmon 2002:44)

En el siguiente cuadro se representan los 3 niveles de modelado que como lo comentamos anteriormente son el nivel alto (llamado ERD *Entity Relationship Level*), el nivel medio (llamado DIS *Data Item Set*) y el nivel bajo (llamado modelado físico).

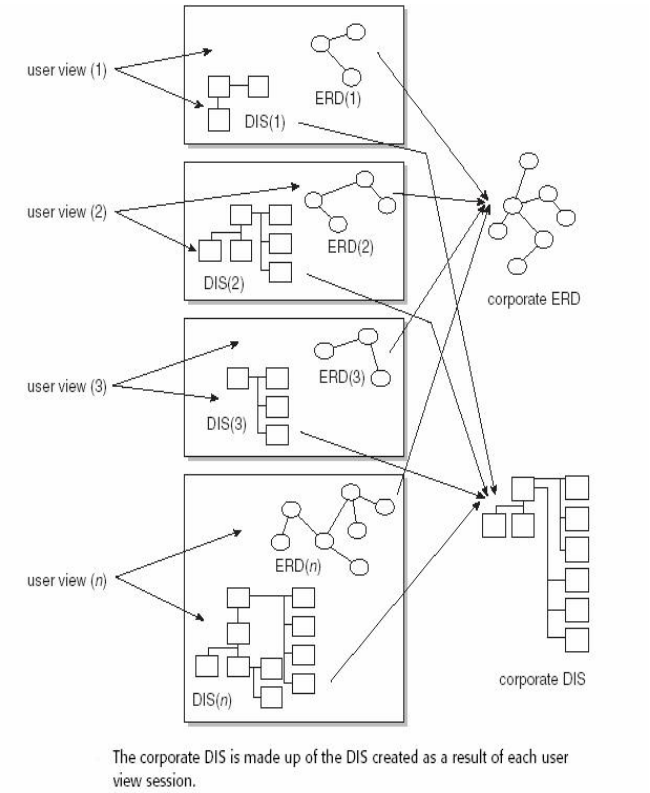


Ilustración 5: METH 2 DE INMON (Business Intelligence Journal 2003:15)

Es un requisito completar los tres niveles del modelo de datos para usar la adaptación especial de Inmon en la metodología en espiral la cual llamo Meth2 ya que para él, la Meth1 es desarrollar el sistema operacional y la Meth3 comprende la existencia de una DW.

Los pasos de dicha metodología son 10 El primer paso es el que llama de soporte a las decisiones esta consiste en el desarrollo de los tres niveles antes mencionados. En la segunda etapa se elabora un análisis del tamaño que tendrá el DW el cual llama análisis de granularidad. Al resolver el tema de granularidad se selecciona el primer ámbito, el cual se convertirá en la primer base de datos departamental (paso 5), se revisan los recursos del sistema (paso 7), se escriben las especificaciones (paso8), los códigos de los programas (paso 9) y se puebla la base de datos (paso 10). El diseño del DW atómico comienza simultáneamente que la revisión de los recursos del sistema (paso 6). Cuando hay bastante información, el equipo conduce una evaluación técnica (paso 3) esta es importante para realizar la preparación del ambiente técnico (paso 4).

A continuación se muestra un cuadro que ejemplifica la metodología 2 de Inmon para lograr su mayor comprensión.

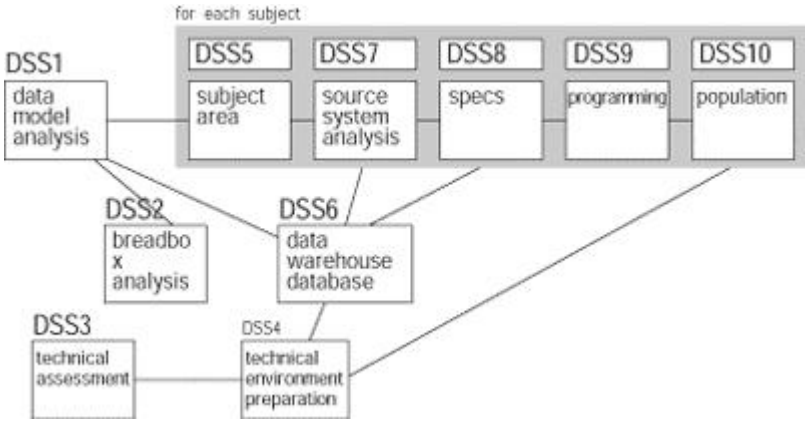


Ilustración 6: Ejemplificación de la Meth 2 (Business Intelligence Journal Volume. 2003:16)

2.2.1.2 Filosofía de INMON

"Data Warehouses are not built all at once. Instead, they are designed and populated a step at time and as such are evolutionary not revolutionary." (Inmon 2002: 41)

Inmon plantea que la construcción de una Data Warehouse es paso a paso y la describe por días. El primer día hay un conjunto esencial de sistemas que hacen operaciones, transacciones y procesos. En el día dos, pocas de las primeras tablas del área del negocio más grande son pobladas dentro del Data Warehouse; en este momento los usuarios descubren los procesos analíticos.
En el día 3, más datos son poblados en el Data Warehouse y con la población de más datos comunes se agregar más usuarios. Los usuarios pueden obtener los datos de una forma sencilla y realizar consultas históricas.

En el día 4, más datos son poblados, algunos de los datos que habían residido en el ambiente operacional se colocan correctamente en el Data Warehouse. Éste puede ya realizar procesos analíticos, además, aumentan los usuarios y las consultas que requieren ser procesadas.
En el día 5, comienzan a figurar la Data Mart o los sistemas OLAP (*Online Analytic Processsing*). Se supone que es más barato y fácil conseguir los procesos hechos extrayendo los datos de la Data Warehouse en cada un de los procesos departamentales.
En el día 6, los usuarios, de manera fácil y rápida, pueden obtener el detalle de los datos en el Data Warehouse. Finalmente, en el día n, la arquitectura está totalmente construida. La mayor parte de los procesos analíticos se llevan a cabo a nivel departamental. Claro que la evolución del primero hasta el día n implica un largo periodo.

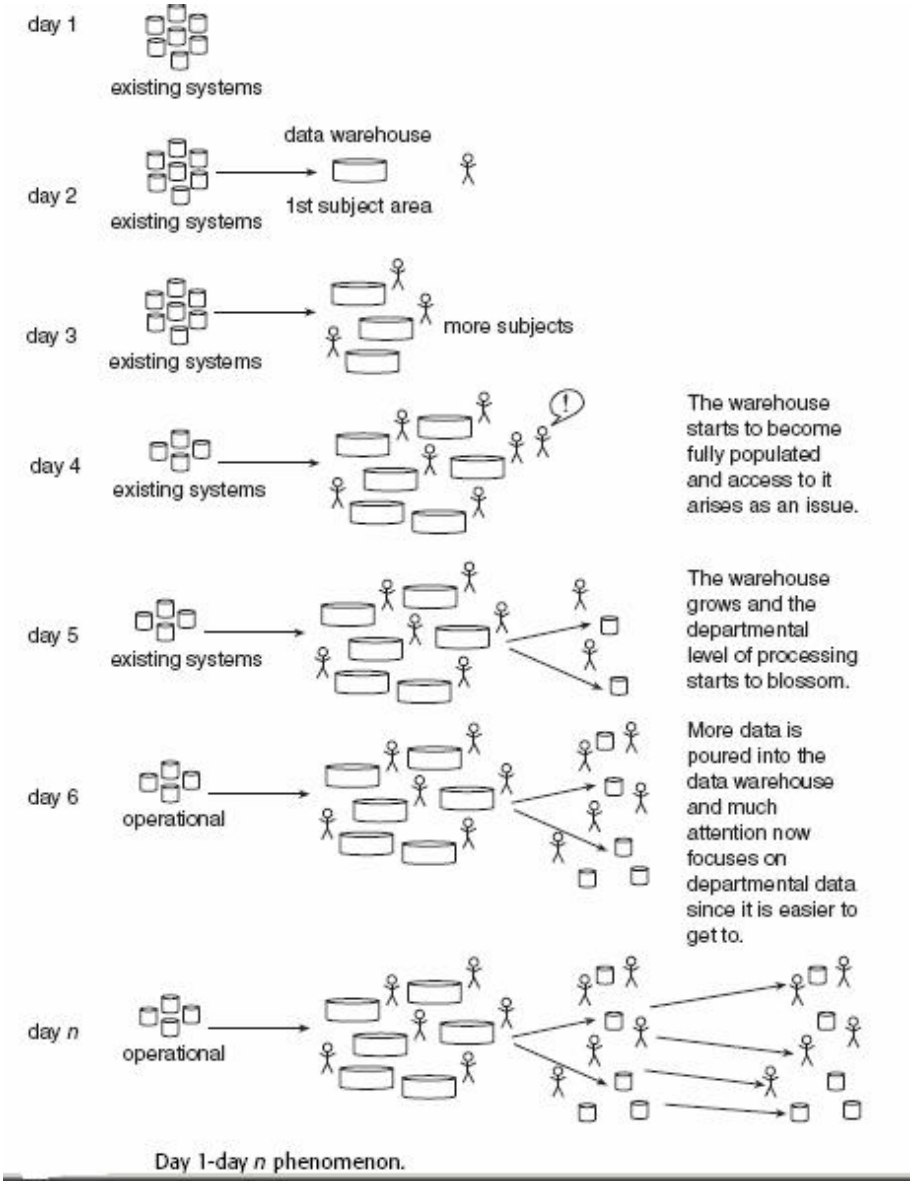


Ilustración 7: Filosofía de construcción de un DW. (Inmon 2002:42)

2.2.2 Modelo de Ralph Kimball

2.2.2.1 Kimball propone los siguientes elementos básicos para un ambiente de Data Warehouse.

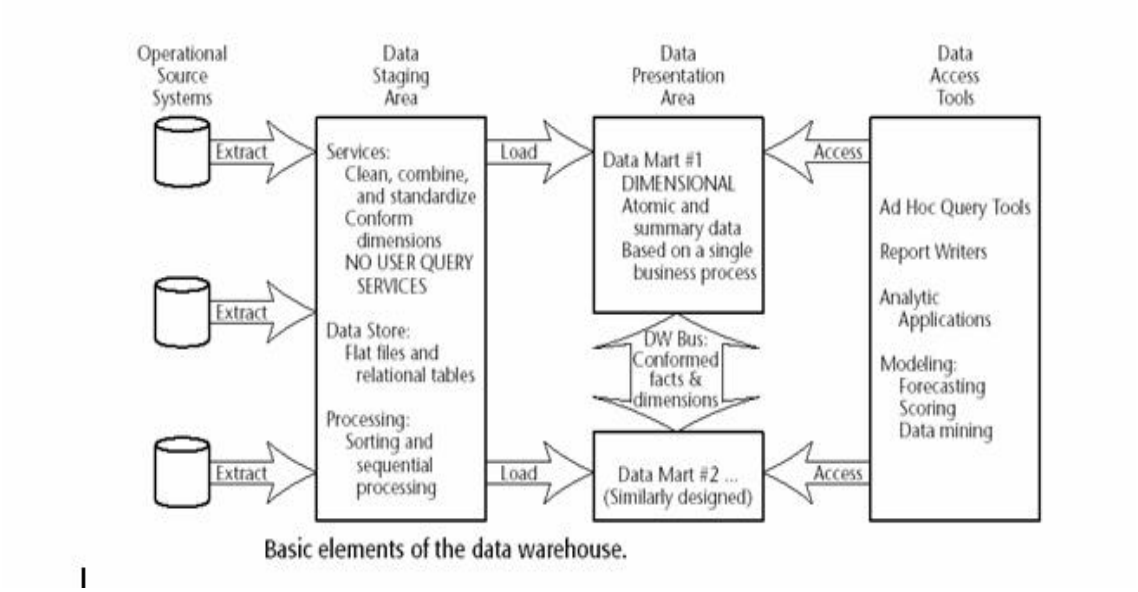


Ilustración 8: Elementos del modelo de Kimball (Kimball 2000:7)

Sistemas operacionales.

Éstos son los sistemas operacionales que almacenan los registros captados durante las transacciones del negocio. Este tipo de sistemas deben estar pensadas afuera del Data Warehouse porque no tenemos control sobre el contenido y forma de los datos en los sistemas operacionales heredados. Sus principales prioridades son procesar, representar y habilitar los datos. Las consultas, que son parte normal del flujo transaccional, deben estar restringidas a sus propias demandas en el sistema operacional. Estos sistemas mantienen datos históricos.

Área de almacenamiento de datos

Ésta es un área de almacenamiento y un grupo de procesos comúnmente referidos al proceso de extracción transformación y carga (*extract-transformation-load* ETL). El área de almacenamiento de datos es todo lo que está entre los sistemas operacionales y el área de presentación de los datos.

La extracción es el primer proceso para llevar los datos al ambiente de Data Warehouse. Extraer significa leer y entender la fuente de los datos y copiar los datos necesarios para el Data Warehouse dentro del área de almacenamiento para su futura manipulación.

Después de la extracción hay varios procesos que se llevan a cabo como la transformación, la limpieza de datos, la combinación de datos desde múltiples fuentes, procesamiento de datos duplicados y asignaciones de llaves.

Se requiere de una buena transformación para continuar con el proceso de cargado de datos dentro del área de presentación del Data Warehouse. Para realizar el

procedimiento de carga de datos se requiere de estructuras físicas normalizadas, en la industria éstas hacen referencia a lo que Kimball llama Enterprise Data Warehouse (EDW). La EDW es la unión de todos los diversos componentes del DW.

Presentación de datos

El área de presentación de datos es donde los datos son organizados, almacenados y habilitados para consultas directas por lo usuarios, para reportes y otras aplicaciones analíticas. El área de presentación es el DW, como se conoce en la comunidad de negocios. Es referido como una serie de Data Marts integrados; un Data Mart se presenta en un proceso singular del negocio. Los procesos del negocio cruzan la frontera de las funciones organizacionales. Los Data Marts deben ser construidos usando dimensiones y hechos.

En la industria, el modelado dimensional es una técnica más viable para obtener los datos del DW que manejar esquemas dimensiónales. El modelado dimensional es un nuevo nombre para la vieja técnica de hacer las bases de datos simples y entendibles.

Si el área de presentación de los datos está fundamentada en una base de datos relacional, entonces las tablas del modelo dimensional están referidas en un esquema de estrella. Pero si está establecida en una base de datos multidimensional o unos procesos de tecnología OLAP entonces son almacenados en cubos. El modelo dimensional es aplicable para ambos modelos: bases de datos relacionales y multidimensionales.

El área de herramientas de datos

Se usa el término herramienta libremente para referirse a varias capacidades que pueden ser provistas por los usuarios del negocio para influir en el área de presentación con el fin de tomar decisiones analizadas. Todas las herramientas para la consulta de los datos provienen del DW o el área de presentación.

Las herramientas de acceso pueden ser tan simples, complejas o sofisticadas como pueden ser las herramientas de consultas de datos; y pueden ser tan poderosas como eficientes para el negocio.

2.2.2.2 Modelo dimensional.

Para entender el modelo dimensional propuesto por Kimball es necesario definir algunos conceptos.

Tabla de hechos Fact table.
Es una tabla primaria en el modelo dimensional, donde las medidas de los resultados numéricos del negocio son almacenadas. Nos sirve para almacenar el resultado de datos medidos por un proceso de negocio en un Data Mart común. Usamos el termino Fact para representar medidas del negocio.
Las tablas de hechos (Fact tables) tienen una o varias llaves foráneas que conectan a las llaves primarias de las tablas de dimensiones (Dimension tables). Cuando todas las llaves en la tabla de hechos coinciden con las llaves primarias correctamente correspondientes a las tablas de dimensiones hablamos de que las tablas contienen integridad.

Tablas de dimensiones (Dimension table)
Las tablas de dimensiones son las tablas que describen o complementan las tablas de hechos. Contienen la descripción textual del negocio, tienen varias columnas o atributos, que describen los registros. Se describen lo mejor posible tendiendo de 50 a 100 atributos.

2.2.2.3 Pasos para el diseño dimensional

Para la construcción de un diseño dimensional Kimball propone cuatro pasos:

Paso 1.- Seleccionar el modelo del proceso del negocio.

El proceso es una actividad natural del negocio que típicamente es soportada por un sistema recolector de datos. Al escuchar a los usuarios es la forma más eficiente para seleccionar el proceso del negocio. Las medidas de resultados que ellos claman para analizar en el DW son resultado de procesos del negocio. Como ejemplo de procesos del negocio podemos incluir materias primas, compras, órdenes, envíos, facturas, inventarios y libro mayor.
Para enfocarnos en los procesos del negocio, podemos entregar información de la totalidad del negocio de una forma más económica que enfocándonos a los departamentos del negocio; si establecemos modelos dimensionales departamentalmente seguro duplicaremos datos con diferentes tablas y terminologías. Además, los datos múltiples fluirán por un modelo dimensional separado que nos hará vulnerables a datos.
Ejecutar una simple consulta también reduce el esfuerzo de la extracción-transformación-carga (ETL) desarrollado así como de la administración de datos y de la carga de almacenamiento de disco.

Paso 2.- Declaración del detalle del proceso del negocio.

Esto significa especificar exactamente lo que representa cada registro de una tabla de hechos individual. El fragmento transmite el nivel de detalle asociado a cada una de las tablas de hechos Éste provee la respuesta a la pregunta: ¿cómo puede describirse un registro individual en una tabla de hechos?
Ejemplo de declaraciones de fragmentos detallados (granos):
Una línea de artículos en una factura recibidos por un doctor.
Un boleto de abordaje individual en un vuelo,
Una fotografía diaria de niveles de inventario para cada producto en un almacén.
Una fotografía mensual de la contabilidad de un banco.

Paso 3.-Elegir las dimensiones que se aplicaran en los registros de las tabla de hechos.

Para definir las dimensiones se responde a la pregunta: ¿cómo la gente del negocio describe los datos que resultan del proceso del negocio? Queremos construir nuestras tablas de hechos con un grupo robusto de dimensiones representando todas las posibles descripciones que toman valores singulares en el contexto de cada medida.
Con la opción de cada dimensión, podemos listar todos los atributos que veremos por cada tabla de dimensión. Ejemplo de atributos comunes son la fecha, producto, cliente, tipo de transacción, y el status.

Paso 4.-Identificar el número de hechos reales con el que será poblado cada registro de la Fact table.

Los hechos reales son determinados respondiendo a la pregunta ¿qué estamos midiendo? Los usuarios del negocio siempre están muy interesados en analizar los procesos del negocio mediante medidas.

Kimball pone como ejemplo el caso de estudio de una cadena de tiendas. En donde incluye las áreas de puntos de venta, ventas al público, almacén, órdenes y envíos, donde en algunas organizaciones estas áreas interrelacionan entre sí.

El primer paso es definir cual es el proceso de negocio con más impacto que se debe construir. Para este ejemplo el proceso del negocio que se podría modelar es el sistema de puntos de venta, el cual nos permitirá analizar qué productos son vendidos en qué tiendas, qué días, y bajo qué condiciones promociónales.

Una vez que se eligió el proceso del negocio es necesario definir el nivel de detalle que debe tener el modelo dimensional. Preferiblemente se debe desarrollar el modelo dimensional con la mayor información atómica, puesto que los datos atómicos son los que no se pueden subdividir de nuevo. En este caso el dato más atómico es una línea individual en la transacción del punto de venta por ejemplo, se seleccionó uno que brindaba los datos de las ventas por producto y su promoción en un determinado almacén un determinado día. Lo que se quiere obtener son las diferencias en ventas del lunes al sábado del conjunto de varios tamaños individuales de marcas como el cereal.

Después de que se haya obtenido la división de la tabla de hechos, se deben escoger las dimensiones. En este caso se escogió el día, producto y almacenamiento. Una cuidadosa división de temas determina las principales dimensiones en la tabla de hechos; es posible agregar más dimensiones a su división básica.

El último paso del diseño es hacer cuidadosamente la determinación de cuales serán los detalles que aparecerán en la tabla de hechos. Otra vez la división elaborada en el paso dos nos ayudara a determinarla. En este caso será la línea individual del artículo en una transacción de punto de venta. Para que quede más claro se representa de la siguiente forma.

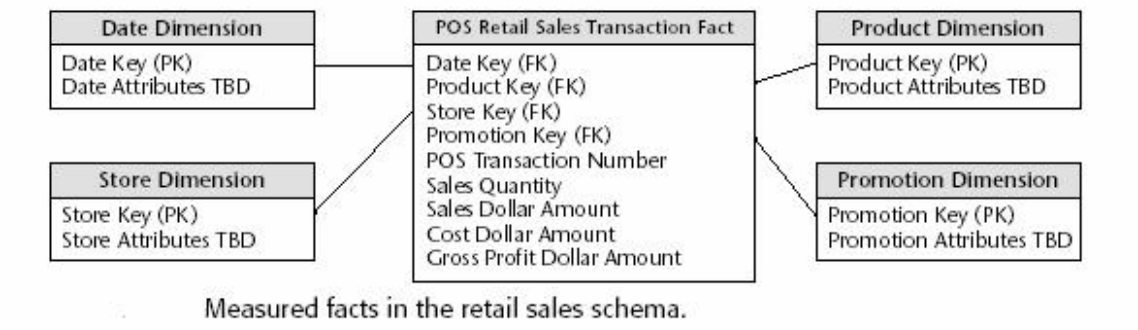


Ilustración 9: Ejemplo de una tabla de hechos. (Kimball 2002:36)

"TDB = To be determinined
POS= Point of sale"

2.2.2.4 Filosofía

Kimball, en su primer capitulo, nos menciona algunos puntos que considera metas para un Data Warehouse, estos son:

- Hacer que la información sea accesible.
- Presentar a la organización información consistente.
- Ser adaptable a los cambios.
- Proteger la información.
- Servir como fundamento para la mejora de la toma de decisiones.

2.3 Comparación

Similitudes
La más importante similitud entre los modelos de Inmon y Kimball son el uso del tiempo (*time-stamped data*), y los procesos de extracción transformación y carga (ETL, *extract transform load*). Aunque la ejecución de estos dos elementos difiere entre ambos modelos, los atributos de los datos y los resultados de las consultas son muy similares.

Similitud en el uso del elemento tiempo (time-stamped data)
Los sistemas operacionales de bases de datos generalmente guardan detalles de los datos ya sea desde una semana hasta dos años. En contraste, el almacenamiento de datos del DW es por 5 a 10 años. El atributo de tiempo es posiblemente el más importante definiendo características de los datos. Esto es porque el atributo de tiempo permite soportar las decisiones, por ejemplo, para el análisis de comparaciones de ventas de un producto X en un año, sería imprescindible saber si fueron vendidas en un fin de semana o en vacaciones; por lo que el atributo de tiempo es indispensable para controlar este tipo de análisis.

Kimball llama a este atributo como "*date dimention*", mientras que Inmon lo llama "time element". En el ejemplo de Kimball la dimensión de la fecha enseña el rango de posibilidad para el atributo de fecha para el Data Mart de la venta al público, la llave de la fecha es una llave artificial que define como se conforma la dimension. En el ejemplo de Inmon, los mismos atributos serán también contenidos con severas diferencias, tablas más normalizadas o simplemente calculadas en el tiempo que el usuario las consulte. La opción de almacenar versus calcular en el modelo de Inmon será guiado por las consideraciones de la presentación.

Similitud en el proceso de extracción transformación y carga (ETL, *extract transform load*)
El ambiente de DW comienza con un proceso de ETL. Los datos son extraídos desde la base de datos operacional, transformados para los estándares del DW y cargados.El proceso de extracción es la primer parte del ETL, involucra mover los datos desde el sistema operacional al área persistente de almacenamiento. Los tiempos de extracción son importantes en este proceso ya que los diferentes sistemas pueden dar distintos tiempos dependiendo de los datos que estén disponibles.

El proceso de transformación es cuando un dato ya extraído en el área de almacén, está listo para tener una modificación, como ser renombrado en el caso de que dos sistemas operacionales nombren de diferente forma un mismo dato. En general, se realizan todas las modificaciones que eviten redundancia e inconsistencias en los datos. Existen varios métodos para la transformación de los datos incluyendo el mapeo de campos y algoritmos de comparación.
El proceso de carga es el paso final en ETL y es cuando se cargan los datos en el DW atómico, en el modelo de Inmon o en Data Marts, con el modelo de Kimball. Este proceso involucra poner físicamente los datos.

El proceso de ETL es esencial para la viabilidad del DW, en él se pretende cuidar la integridad de los datos dentro el DW. Obviamente si dos usuarios realizan consultas en el mismo período de tiempo y se les arrojan diferentes resultados, esto afecta la credibilidad de los datos y el DW estará erróneo a los ojos del usuario. Por esto, el ETL es considerado como la actividad del DW más intensa y que soporta el análisis y la toma de decisiones.

Diferencias
Las diferencias esenciales entre ambos modelos se encuentran en las áreas de desarrollo de metodología, el modelado de datos, y la arquitectura del DW. La siguiente tabla las enuncia:

	Inmon	**Kimball**
Metodología y arquitectura		
Enfoque general	Arriba-Abajo	Abajo- Arriba
Arquitectura estructurada	El DW alimenta las bases de datos departamentales	Modelo de Data Mart un singular proceso del negocio; consistencia de la empresa alcanzada por datos dimensiones
Complejidad en el método	Poco compleja	muy simple
Comparación con metodologías establecidas desarrolladas	Derivada de la metodología en espiral	Proceso en 4 pasos, salida desde métodos de RDBMS
Discusión sobre el diseño físico	Muy complicado	Muy ligero
Modelado de datos		
Orientación de los datos	Orientado al objeto	Orientado al proceso
Herramientas	Tradicionales(ERD DIS)	Modelo dimensional, a salida desde el modelo relacional
Acceso del usuario final	Bajo	Alto
Filosofía		
Audiencia primaria	Especialistas de IT	Usuarios finales
Lugar en la organización	Parte integral de Corporate Information Factory (CIF)	Transformador y retenedor de datos operativos
Objetivo	proveer de una buena solución	Proveer una solución que hace

	técnica basadas en metodologías y tecnologías para bases de datos	fácil al usuario final la consulta directa de los datos y la habilidad de obtener tiempos de respuesta razonables.

(Business Intelligence Journal 2002:23)

2.2.2.5 Diferencias en metodologías de desarrollo y arquitecturas

Para tener un DW atómico, en el modelo de Inmon, deben ser hechos desarrollados de arriba-abajo. El DW atómico debe servir para la empresa entera y todas las bases de datos departamentales obtendrán sus datos a través del DW atómico. El esfuerzo del desarrollo de arriba-abajo tiene un inevitable grado de complejidad, la metodología de Inmon no es la excepción, aunque claramente presenta una ayuda que lo hace menos compleja.

La metodología y la orientación arquitectónica son técnicas, cuyo interés primario es asegurar que una solución técnica funcione. El objetivo de esta solución técnica es optimizar entradas y salidas. Por esto, la audiencia de Inmon está claramente formada por profesionales de IT. Pocos lectores de negocios tienen un conocimiento previo para entender la propuesta de desarrollo de Inmon, ya que enfatiza aspectos técnicos y es necesario el entendimiento del desarrollo en espiral con el que se basa. El énfasis en los aspectos técnicos del desarrollo implica que los miembros del equipo del DW sentirán mejor grado de propiedad sobre el DW a diferencia de los usuarios finales. En contraste, Kimball desarrolla su metodología en 4 pasos y es muy accesible para el usuario final. El usuario puede entender algunos conceptos técnicos sobre el bus de datos y las dimensiones que lo conforman sin un extensivo estudio, en contraste con aprender a interpretar los ERDs.

Por definición, la propuesta de abajo hacia arriba (bottom-up) involucra pocos elementos de datos en comparación con un desarrollo top-down. Incluso si los usuarios no están familiarizados con el concepto de procesos de negocio, el pequeño alcance de un Data Mart es más accesible para los usuarios finales. La metodología que propone la cual llama The meth2, antes mencionada, ayuda a hacer que el alcance de las enterprisewide sea menos desalentador, pero el alcance de las Data Marts sea de todas formas considerablemente fácil para usuarios físicos

2.2.2.6 Diferencias en modelado de datos

Las diferencias en el modelado entre Inmon y Kimball son la orientación hacia los datos y modelar las reglas y las técnicas, Inmon toma la propuesta de orientación a temas o el manejo de dato como modelo. Esto quiere decir que la naturaleza de los datos dirige el proceso de modelado de datos. Los datos tradicionales que Inmon nos presenta en el modelo como herramientas son los *Entity Relationship* ERD y los *Data Item Set* DIS. Los miembros del equipo de IT del DW tendrán la responsabilidad principal de modelar los datos porque las herramientas y los procesos que implican requieren un fondo técnico para usarlos efectivamente. Los usuarios finales pueden asistir a presentaciones de la revisión, pero pocos podrían entender ERDs o DISs sin ayuda, a menos de que recibieran un entrenamiento especial bastante extenso.

En contraste, Kimball toma una orientación de proceso, queriendo decir que el modelado de datos se hace una tentativa de definir la interacción de datos a través de un proceso de negocio (como ventas al público o el inventario). Por esa naturaleza, cada proceso del negocio usualmente cruza líneas departamentales. Esto encaja muy bien con el nuevo modelado de datos enfocado al modelado dimensional de los datos, en el cual los procesos determinan qué métricas (hechos) y atributos (dimensiones) son importantes para reclamar un lugar en el DW. Las herramientas del modelado dimensional permiten al usuario tomar un papel activo dentro del proceso del modelado de datos.

2.2.2.7 Diferencias en filosofía

Inmon cree que la representación de un DW completo podrá maximizarse mediante un aseguramiento técnico del proceso de desarrollo. Mientras que Kimball ve al usuario final y a los profesionales de IT compartir tareas similares. La participación activa del usuario durante el proceso de desarrollo aumentará la probabilidad de aceptación del usuario sobre el DW.

Por supuesto, ambos expertos son muy conscientes de que un DW que no implique a los usuarios en todos los puntos de su ciclo de vida es igual de probable que fallen como un mal que lleva a cabo para los usuarios. Lo que los dos no están de acuerdo es cuales de estas puntos debe ser considerado como el más importante.

CAPITULO III HERRAMIENTAS PARA LA CONSTRUCCIÓN DE UN DW O UN DM.

En este capítulo veremos características de algunas herramientas que nos pueden facilitar la tarea de la construcción de un Data Warehouse o bien de un Data Mart.
El objetivo de este apartado es mostrar un pequeño resumen de herramientas que nos pueden ayudar en la construcción de un Data Warehouse (DW) o un Data Mart (DM), exponiendo algunas de sus características.

3.1 Herramientas de Software libre u *OpenSource*

Software Libre se refiere a la libertad de los usuarios para ejecutar, copiar, distribuir, estudiar, cambiar y mejorar el software. De modo más preciso, se refiere a cuatro libertades de los usuarios del software: [Web: 06]

- La libertad de usar el programa, con cualquier propósito (libertad 0).
- La libertad de estudiar cómo funciona el programa, y adaptarlo a ciertas necesidades (libertad 1). El acceso al código fuente es una condición previa para esto.
- La libertad de distribuir copias, con lo que se puede ayudar a otros (libertad 2).
- La libertad de mejorar el programa y hacer públicas las mejoras a los demás, de modo que toda la comunidad se beneficie. (libertad 3). El acceso al código fuente es un requisito previo para esto.

Para dar solución a diversos problemas de volúmenes de información, existen varias herramientas que agrupan un conjunto de aplicaciones como son la integridad de los datos, limpieza de datos, creación y mantenimiento de una Data Warehouse entre otras.

A continuación mencionaré algunas herramientas de *Open Source* que hay en la actualidad y que cubren las necesidades de volúmenes de información dando una solución. Las herramientas de las cuales mencionaré algunas de sus características son las siguientes:

- La plataforma Pentaho Open Source Business Intelligence.
- JasperSoft Open Source Business Intelligence Suite.

La plataforma Pentaho Open Source Business Intelligence (BI)

Pentaho se define a sí mismo como una plataforma de BI “orientada a la solución” y “centrada en procesos” que incluye todos los principales componentes requeridos para

implementar soluciones basadas en procesos y ha sido concebida desde el principio de esa manera.

Las soluciones que Pentaho pretende ofrecer se componen fundamentalmente de una infraestructura de herramientas de análisis e informes integrados con un motor de *workflow* de procesos de negocio. La plataforma será capaz de ejecutar las reglas de negocio necesarias, expresadas en forma de procesos y actividades, y de presentar y entregar la información adecuada en el momento adecuado. Pentaho está construido en torno al servidor de aplicaciones J2EE JBoss y Jboss Portal, habilitando que toda la información sea accesible mediante un *browser* en la intranet de la empresa

Pentaho presenta informes en los formatos habituales (html, excel, pdf, etc.) mediante JfreeReport, proyecto incorporado recientemente a Pentaho junto con su responsable Thomas Morgner, u otras plataformas como BIRT o JasperReports. Para la generación de PDFs utilizan, como podría ser previsible, el conocido Apache FOP. Asimismo incorpora la librería JPivot, gracias a la cual podemos ver tablas OLAP a través de un browser y realizar las aplicaciones típicas de análisis OLAP (*drill down* o *slice and dice*).
Recientemente se anunció Pentaho Report Design Wizard, una herramienta de diseño de informes, que facilita el trabajo con JfreeReport y supera sus limitaciones. Es de suponerse que algo tiene que ver JFreeDesigner, el diseñador de informes para JFreeReport de jfree.org, ya que Thomas Morgner es también su responsable.

Los módulos de la plataforma Pentaho BI son:

- Pentaho Reporting – es un modulo de informes que ofrece la solución adecuada a las necesidades de los usuarios. Es una solución basada en el proyecto JFreeReport y permite generar informes de manera ágil y de gran capacidad. Además, permite la distribución de los resultados del análisis en múltiples formatos - todos los informes incluyen la opción de imprimir o exportar a formato PDF, XLS, HTML y texto. Los reportes de Pentaho permiten también programación de tareas y ejecución automática de informes con una determinada periodicidad.
- Pentaho Analysis - suministra a los usuarios un sistema avanzado de análisis de información con uso de las tablas dinámicas (*pivot tables* y *crosstabs*), generadas por Mondrian y Jpivot. El usuario puede navegar y ajustar la visión de los datos, los filtros de visualización, añadiendo o quitando los campos de agregación. Los datos pueden ser representados en forma de SVG o Flash, como *dashboards widgets*, o también integrados con los sistemas de minería de datos y los portales web (*portlets*). Además, con el Microsoft Excel Analysis Services, se puede analizar los datos dinámicos en Microsoft Excel (usando la conexión a OLAP server Mondrian).
- Dashboards - todos los componentes del modulo Pentaho Reporting y Pentaho Análisis pueden formar parte de un tablero de mando. En Pentaho Dashboards es muy fácil incorporar una gran variedad en tipos de gráficos, tablas y velocímetros (dashboard widgets) e integrarlos con los Portlets JSP, en donde se puede visualizar informes, gráficos y análisis OLAP.
- Data Mining – el análisis en Pentaho se realiza con la herramienta WeKa.
- Integración de Datos - se realiza con una herramienta Kettle ETL (Pentaho Data Integration) que permite implementar el proceso ETL.

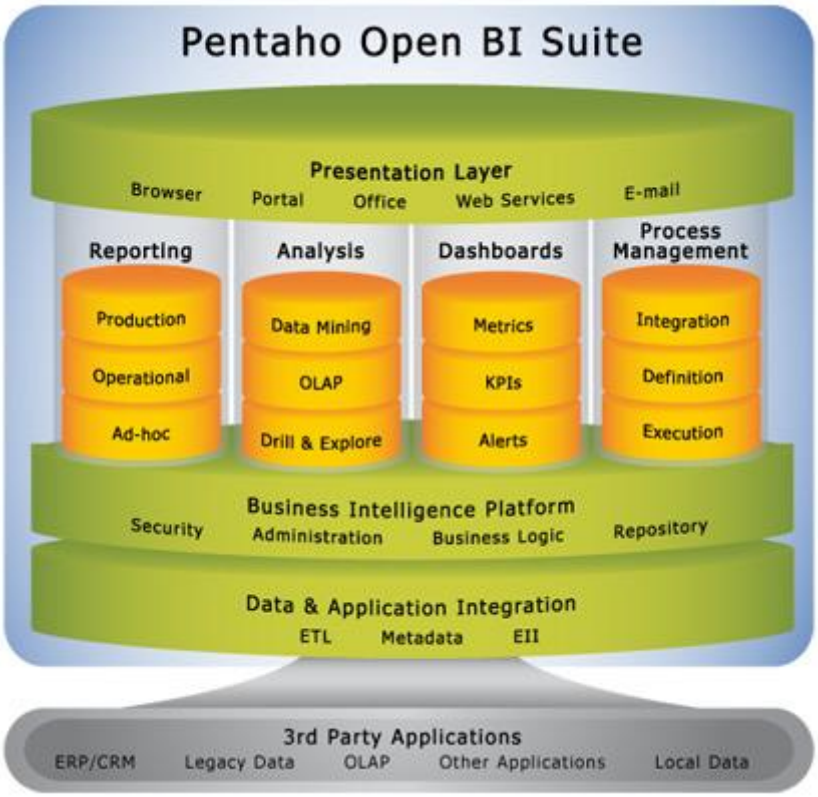


Ilustración 1: Arquitectura de Pentaho [Web:02]

Kettle ETL es un proyecto belga que incluye un conjunto de herramientas para realizar ETL. Uno de sus objetivos es que el proyecto ETL sea fácil de generar, mantener y desplegar.

Se compone de 4 herramientas:

- SPOON: permite diseñar de forma gráfica la transformación ETL.
- PAN ejecuta las transformaciones diseñadas con SPOON.
- CHEF permite, mediante una interfaz gráfica, diseñar la carga de datos incluyendo un control de estado de los trabajos.
- KITCHEN permite ejecutar los trabajos batch diseñados con CHEF.

JasperSoft Open Source Bisiness Intelligence Suite [Web:02]

Cuando usamos alguna de las parte de JasperSoft Open Source Bisiness Intelligence Suite, JasperReports, JasperServer o JasperAnalysis, las organizaciones pueden desarrollar, admnistrar y documentar los procesos de la integración de datos con más precisión, mayor entendimiento de reportes y procesos analíticos en línea (OLAP).

Dentro de JasperSoft Open Source Bisiness Intelligence Suite, también encontramos JasperETL. Ésta es una plataforma completa y lista para correr integración de datos para organizaciones de todos los tamaños, provee de una alta representación de capacidades de extracción transformación y carga de datos. JasperETL es apropiado para todas las necesidades de análisis e integración operacional de datos, a pesar de la

complejidad de la herramienta. JasperETL puede sin embargo ponerse en marcha para proveer capacidades comprensibles de ETL para otras aplicaciones y sistemas.

Esta herramienta simplifica y automatiza la integración de datos. Algunos de sus beneficios son:

- Crea, maneja y mantiene fácilmente procesos de integración de datos.
- Brinda interfaces avanzadas y gráficos intuitivos para usuarios.
- Puede ser usado para cualquier empresa sin importar el tamaño.
- Es más rápida que las herramientas de ETL que se encuentran en el mercado.
- Se trata de un *open source* comercial que abarca el rango total de necesidades del mercado.

Reportes y análisis de datos

JasperETL simplifica y estandariza la experiencia de un usuario final con un Data Warehouse o Data Mart. JasperETL puede rutinariamente extraer, transformar y cargar datos del sistema operacional rutinariamente dentro de "un esquema de estrella", donde puede ser seguro y rápidamente accedido por usuarios interactivos o usuarios finales que reporten y analicen.

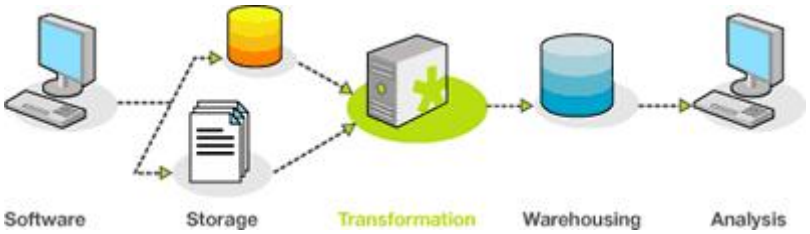


Ilustración 2: Diagrama de Acceso [Web:02]

Costo beneficio

Pocas de las compañías en el mundo usan herramientas de ETL para conocer sus necesidades de integración de datos y el resto de las compañías usan scripts escritos en Cobol, Java, Python, Bash, Perl, y otros lenguajes. Estos scripts representan una amenaza real para la integridad de los sistemas de información, presentando problemas con ciertas operaciones, mantenimiento, actualizaciones y robustez. JasperETL provee una plataforma completa para integración, que encuentra todas las necesidades de integración de datos operacionales, como consolidación de datos, duplicación, sincronización, calidad, migración y cambios de datos capturados.

Componentes clave y características importantes

Job designer- provee de un editor gráfico y una vista funcional del proceso de ETL.

Transformation Mapper- provee de un editor grafico y una vista del mapeo complejo y las transformaciones.

Ilustración 3: Editor gráfico de Job Designer. [Web:02]

Ilustración 4: Editor gráfico de Transformation Mapper. [Web:02]

Debugging en tiempo real- permite un seguimiento de las estadísticas de ETL y rastrear el proceso completo de transformación en tiempo real.

Business Modeler- proporciona una vista gráfica no técnica del *workflow* de la información del negocio.

3.2 Herramientas de Software Comercial o propietarias.

Las herramientas que mencionaremos en este apartado son:

- Oracle Business Intelligence que a su vez agrupa las siguientes herramientas:
  - Oracle Warehouse Builder.
  - Oracle DataMart Suite.
- SAS.
- Gnexux.

Oracle Warehouse Builder

Oracle Warehouse Builder es una herramienta para todos los aspectos de la administración de datos. Usa la Base de datos de Oracle para transformar datos en una información de alta calidad. Provee calidad de datos, auditoria de datos, integración completa al modelo relacional o al dimensional, y maneja un completo ciclo de vida de los datos y metadatos. También permite crear un Data Warehouses y migrar datos desde sistemas heredados. Además, brinda la posibilidad de consolidar datos desde distintos recursos, limpiar y transformar los datos para proveer la información de calidad y el mantenimiento de metadatos corporativos.

Oracle Warehouse Builder admite, de manera funcional, la extracción, transformación y carga de los datos desde recursos heterogéneos. Es posible almacenar datos en formatos relacionales, multidimencionales o en formatos de archivos planos.
Transformar los datos para lograr una alta calidad de información requiere de:

- Acceder a una amplia variedad de recursos de datos.
- Habilidad para transformar y limpiar los datos.
- Habilidad para implementar el diseño de diversas aplicaciones.

Warehouse Builder Core Funcinality esta incluido en las ediciones de Oracle Database por lo cual no cubre un costo adicional para diseñar, desarrollar y manejar un Data Warehouse básica con Oracle, si se requieren procesos avanzados para la extracción de datos o un Data Warehouse más robusto se requiere del kit de Oracle Warehouse Builder que incluye:

- Warehouse Builder Enterprise ETL Option
- Warehouse Builder Data Quality Option
- Warehouse Builder Connector E-Business Suite
- Warehouse Builder Connector People Soft
- Warehouse Builder Connector SAP/ R3 Connector

Oracle Warehouse Builder se compone de un conjunto de interfaces gráficas que facilitan las tareas en la aplicación de diseños de sistemas de datos complejos. Sus diseños se guardan como Metadatos en un repositorio centralizado. El repositorio centralizado, conocido como *Warehouse Builder Repository* es almacenado en la base de datos de Oracle. El centro de diseño es la interfaz que provee la representación visual del repositorio centralizado, se usa para importar fuentes de objetos como tablas e índices y diseñar el proceso de ETL.
El mapeo es el objeto con el cual se define el flujo de datos desde las fuentes. Basado en el diseño de mapeo, el Warehouse Builder genera el código requerido para implementar el ETL lógico. Deployment es el proceso de copiar los metadatos relevantes y el código generado en el centro de diseño para el esquema meta. Éste está definido como la base de datos que ejecuta el ETL lógico que ha proyectado el centro de diseño

La siguiente figura muestra los componentes de BW:

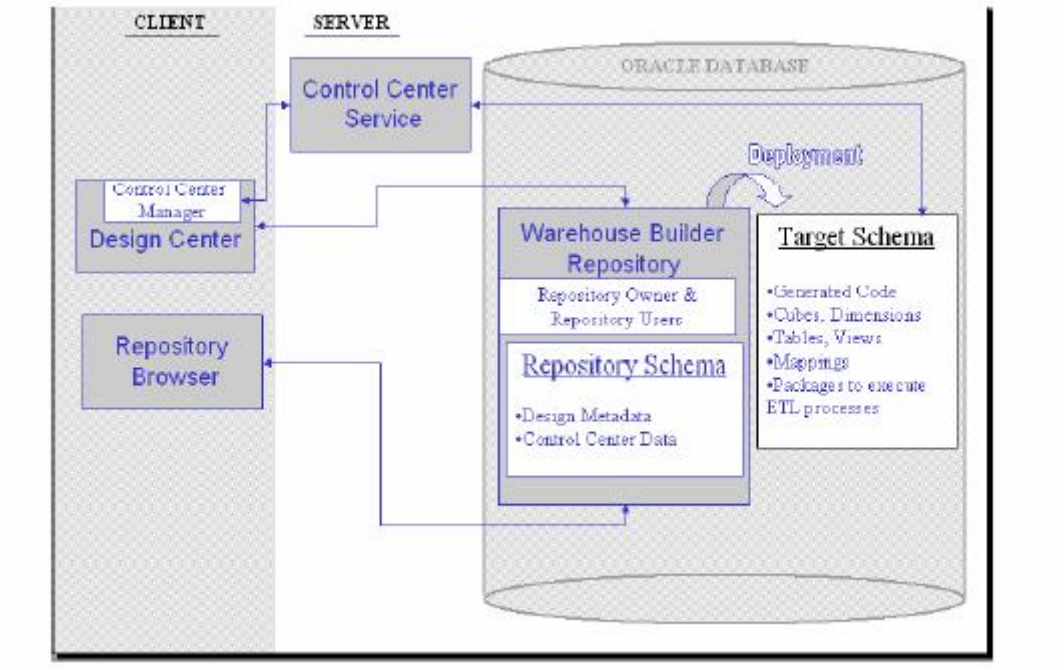


Ilustración 5:Componentes del builder warehouse(Oracle Data Mart Suite 1999:2-2)

Oracle DataMart Suite

Oracle Data Mart Suite es construido en Oracle8 database server, la bases de datos para data warehousing más ampliamente usada en la industria. La suite integra herramientas visuales para diseño de un Data Mart, herramientas gráficas fáciles de usar para la extracción de información desde sistemas operacionales, una alta representación, bases de datos escalables que sirve como un repositorio compartido para datos y metadatos, un servidor web para el acceso a Internet para el Data Mart, una nueva generación de consultas, reportes, herramientas de análisis y documentación. Todos estos componentes se pueden encontrar al instalar Oracle Data Mart Suite.

Los componentes que incluye Oracle Data Mart Suite son los siguientes:

- Oracle Data Mart Designer
- Oracle8 Enterprise Edition
- Oracle Enterprise Manager
- Oracle Data Mart Builder
- Oracle Discoverer
- Oracle Reports
- Oracle Web Application Server

The SAS Intelligence Value Chain

SAS Intelligence Value Chain representa los enlaces requeridos para construir un sistema de *business intelligence*. Cada enlace en la cadena, excepto el de planeación, corresponde a los productos de SAS. Esto significa que se puede crear una completa solución usando el Software de SAS.
Existen 5 componentes clave en la cadena de SAS que a continuación se muestra:

Ilustración 6: Componentes claves de SAS (SAS Instituet 2004: 6)

SAS ETL Studio es el principal producto que es asociado con el enlace ETL, la etapa en donde se crea el Data Warehouse o Data Mart mediante la integración de datos desde los recursos de datos operacionales existentes, como los son los SAS data set, tablas del manejador de sistemas de base de datos, y las tablas de las aplicaciones de la empresa. Los componentes del sotware en este enlace cumplen con las siguientes tareas:

- Extraer datos desde los recursos de datos, a pesar de la plataforma en donde los recursos de los datos residan o sus formatos.
- Transformar los datos antes de escribirlos en la tabla final.
- Carga de las tablas finales ya transformadas en el Data Warehouse o Data Mart.

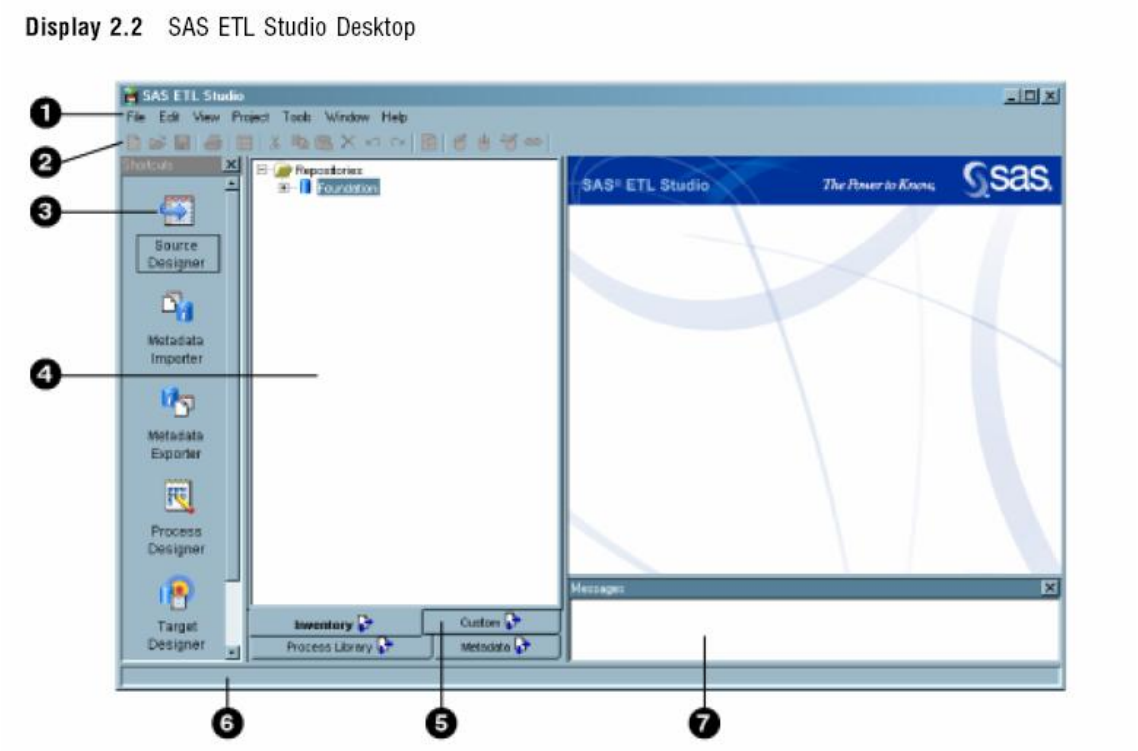


Ilustración 7: Escritorio de SAS ETL Studio. (SAS ETL Studio 2004:16)

El escritorio de SAS ETL Studio consiste en los siguientes componentes (Ilustración 19):

1. Barra de Menú.- Se usa para acceder a los menús desplegables. La lista de opciones activa varias de ellas de acuerdo a lo que está en el área de trabajo y el tipo de objetos seleccionados. Los menús inactivos están escondidos o deshabilitados.
2. Barra de Herramientas.- Contiene métodos abreviados de los temas en la barra de menús.
3. Barra de Métodos abreviados (*Shortcut bar*).- Despliega el panel de tareas en forma de iconos en el lado izquierdo de la aplicación. Cada icono despliega una ventana, un *wizard* o unas ventanas de *wizards*.
4. Vistas de árbol.- Despliega la metadata, la cual es asociada con el actual repositorio de metadatos
5. Árboles.- Varios árboles despliegan su contenido en repositorios de metadatos de varias formas. El proceso de librería de árbol puede usarse para arrastrar y soltar plantillas transformadas en procesos de flujo de diagramas para el trabajo.
6. Línea de estado.- Ubicada al final, nos sirve para desplegar información de errores.
7. Ventana de mensajes.- Despliega mensajes.

Genexus

Dentro de Genexus se encuentran herramientas para Data Warehouse como GXplorer y Gxquery. GXplorer es parte de la suite de Business Intelligence que permite a los

clientes de GeneXus dar soporte para que el proceso de toma de decisiones sea eficiente y rentable. Además, permite construir y mantener una Data Warehouse fundado en el conocimiento del negocio capturado en sus bases de conocimiento, evitando las herramientas costosas y el largo tiempo de la puesta en práctica.
Genexus y GXplorer soportan el ciclo completo de una Data Warehouse, desde la implementación hasta el análisis. GXplorer OLAP es un cliente de uso fácil de OLAP que permite a usuarios finales definir sus propios informes
.

GXplorer Web Access permite realiza consultas *ad hoc* por parte del usuario, especificadas en su lenguaje (en términos del negocio) sobre un Data Warehouse, además tiene la inteligencia de guiarlo en el análisis de la información, no permitiendo la realización de consultas inválidas.
Los grandes conceptos sobre los cuales se formalizó el análisis del negocio son los denominados: dimensiones e indicadores.

Los indicadores son aquello que se quiere analizar. Las dimensiones nos muestran las perspectivas bajo las cuales se quieren analizar los indicadores. La siguiente imagen muestra como se selecciona un indicador y las dimensiones dentro de GXplorer.

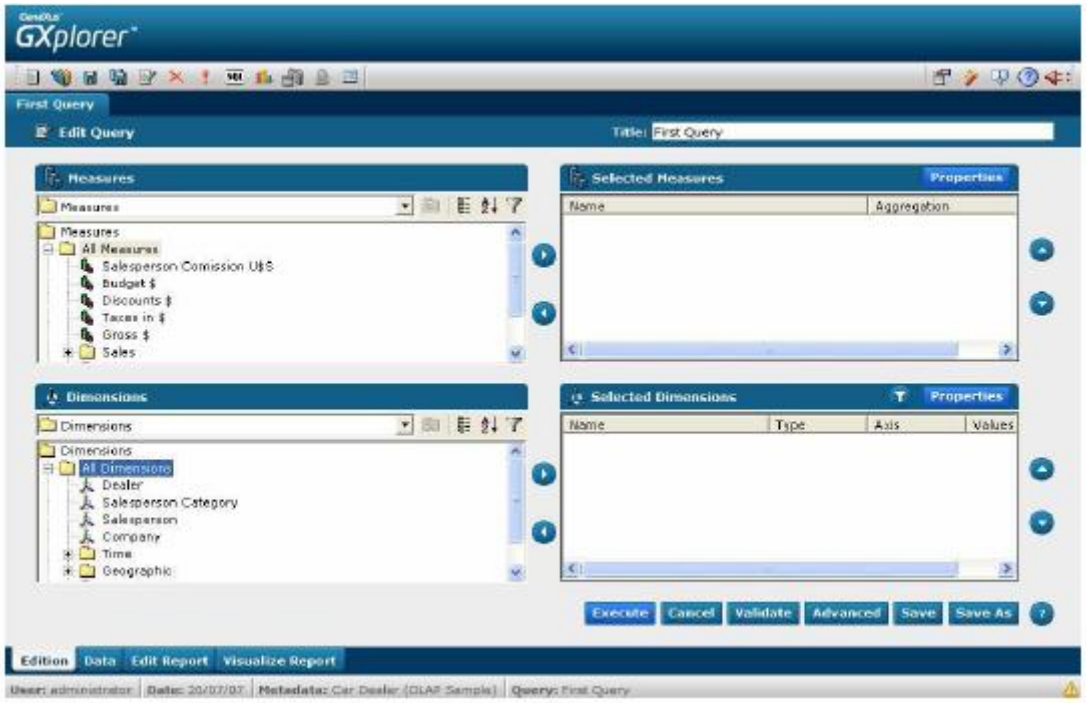


Ilustración 8: Escritorio de GXplorer. (GeneXus Gxplorer 2007:5)

GXquery tiene los siguientes componentes:
GXquery Manager.- Obtiene automáticamente toda la información de una base de conocimiento o de una base de datos que necesita GXquery para operar (Metadata). Debe ejecutarse cada vez que se efectúa un cambio en la base de conocimiento o en la base de datos asociadas.
GXquery Settings.- Se encarga de configurar las opciones de seguridad de la herramienta, las propiedades de los atributos, conexión con el DBMS, entre otras cosas. Es una herramienta para el administrador.

GXquery para MSExcel.- Es una interfaz de Excel para efectuar consultas dinámicas sobre una base de datos operacional. Es una herramienta para el usuario final.
GXquery Web Access.- Es una interfaz HTML para efectuar consultas dinámicas mediante un *browser* sobre una base de datos operacional. Es una herramienta para el usuario final
GXquery Services.- Es una interfaz que a través de Web Services permite interactuar con la metadata y sus elementos (usuarios, atributos, consultas, etc.) Es una herramienta para desarrolladores que desean integrar la potencia de GXquery en sus aplicaciones.

Conclusiones de las herramientas

Hay una gran diversidad de herramientas tanto de software libre como software comercial para el desarrollo de una Data Warehouse o un Data Mart, por lo que las herramientas que se presentaron en este capitulo no contienen un indicador por el cual se puedan apreciar su participación en el mercado pero considero que son de gran utilidad para un profesionista que comienza a introducirse en estos temas.
Las herramientas anteriormente mencionadas son de gran utilidad dentro del proceso de ETL. Como vimos en el capitulo dos el proceso de ETL es una similitud que tienen las metodologías de Inmon y Kimball por lo que podría decirse que comparten en parte las metodologías.
Mi experiencia en el uso de estas herramientas no es profunda, las herramientas que he usado en general son las comerciales básicamente porque en las organizaciones donde he laborado cuentan con las licencias y para estas le es conveniente porque a diferencia de algunas herramientas de software libre no tienen soporte técnico por medio de un contrato como las de software comercial.

CAPITULO IV BREVE GUÍA PARA LA CONSTRUCCIÓN DE UN DM O UN DW.

4.1 Construcción de un Data Mart .

A continuación se muestra las etapas para construir un Data Mart según el manual de Oracle(Oracle Data Mart Suite:1999):

- Diseño
- Construcción
- Población
- Acceso
- Administración

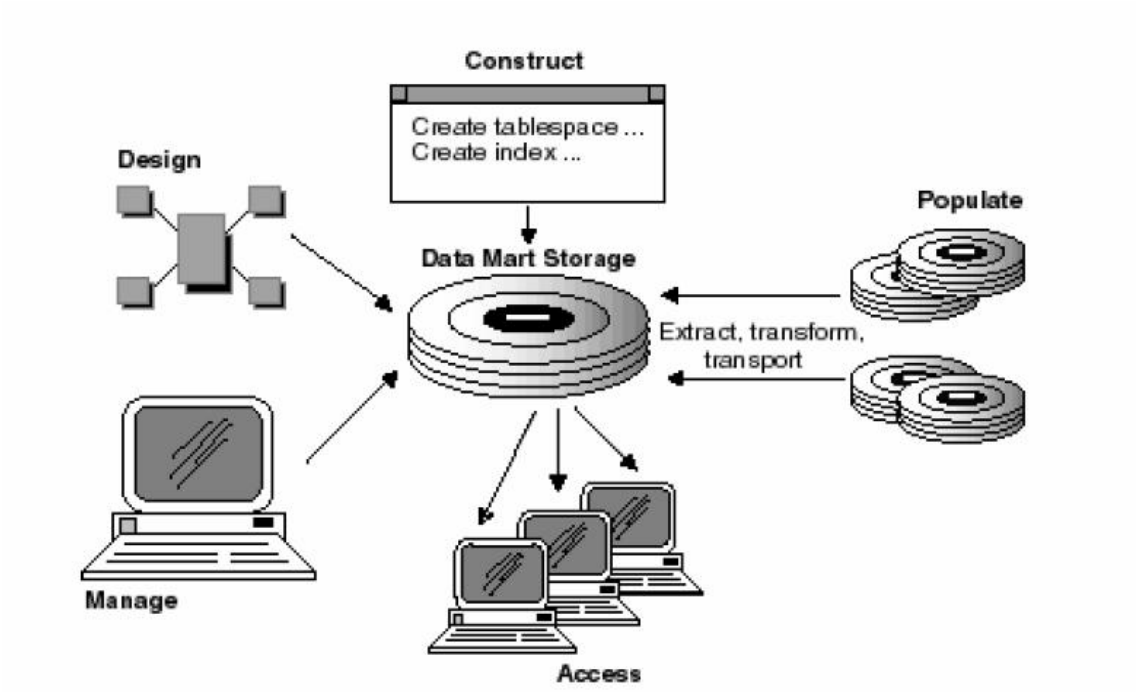


Ilustración 1: Etapas para construir un Data Mart (Oracle Data Mart Suite 1999:26)

El primer paso es el diseño del proceso de Data Mart. Este paso cubre todas las tareas que van desde iniciar la solicitud del Data Mart y la recopilación de la información sobre el requerimiento, hasta el desarrollo del diseño lógico y físico del Data Mart. Este paso contiene las siguientes tareas:

- Recopilar los requerimientos técnicos y del negocio.
- Identificar los recursos de los datos
- Seleccionar un adecuado subconjunto de datos
- Diseñar la estructura lógica y física del Data Mart

El segundo paso es la construcción, que involucra la creación de la base de datos física y de las estructuras lógicas asociadas con el Data Mart para proveer un fácil acceso a los datos. Las tareas que se realizan son

- Creación de la base de datos física y las estructuras de almacenamiento como el espacio de las tablas relacionadas al Data Mart
- Creación del esquema de objetos como tablas e índices definidos previamente en el paso anterior
- Determinar la mejor forma de crear las tablas y las estructuras de acceso, tales como
  índices de mapa de bits, para la óptima ejecución de consultas

Para este paso se requiere de un manejador de bases de datos ya que en esta paso se realizan varias funciones de almacenamiento y manejo de datos como la creación de tablas, eliminar, insertar y modificar los datos. Además se proveen los mecanismos de consulta y protección de datos para que no sean modificados los datos por un usuario no permitido.

La etapa de llenado cubre todas las tareas requeridas para obtener los datos de los recursos, limpiarlos, modificarlos en el formato correcto y nivel de detalle para moverlos al Data Mart. Más formalmente las tareas de esta etapa son:

- Mapear la fuente de datos para orientarlos.
- Extracción de datos.
- Limpiar y Transformar los datos.
- Carga y almacenamiento de metadatos.

Para la ejecución de esta etapa se puede usar cualquier herramienta de ETL como las que se mencionaron en el capitulo anterior.

La etapa de Acceso se refiere a poner los datos en uso para consultarlos, analizarlos, crear reportes, cartas y gráficos para publicarlos. El usuario final comúnmente usa una herramienta grafica que le permite realizar consultas de la base de datos. En esta etapa se realizan las siguientes tareas:

- Construcción de capas intermedias para el usuario final. Esta capa transfiere las estructuras de bases de datos y los nombres de los objetos en términos del negocio, por lo que el usuario puede interactuar con el Data Mart usando los términos relacionados con la función del negocio.
- Mantener y manejar las interfaces del negocio.
- Crear y mantener estructuras de bases de datos, como resumir las tablas que ayudan a los usuarios a la resolución de consultas rápidamente.

Según los expertos antes de pensar en lo robusto que puede llegar a ser un Data Warehouse es importante dimensionarlo en pequeña escala y después dejar abierta la posibilidad de que vaya creciendo el almacenamiento de información, a continuación menciono algunas ideas tomadas de los autores del libro…

Dependiendo del tamaño de datos se puede determinar si requiere de un modelado ya que existen DM muy pequeños e informales que no lo requieren. (Ken, Buss y Ryan1998:85)

A continuación mencionare algunas generalidades importantes sobre los DM basándome en (Ken, Buss y Ryan.1998:82-87)

La carga de datos en un Data Mart.
Es propiamente desde una DW. Algunos factores en la carga de programas incluye: la agenda de carga, la frecuencia con que se ejecuta un programa, los totales o parciales de una actualización, si las tablas de Data Mart deben ser refrescadas en su integridad o solo añadidas, la personalización de los datos del DW, selección de funciones de agregación, combinación y resumen de datos, eficiencia en ejecución (que tan rápido puede ser la carga completada), integridad de datos y producción de metadatos.

Metadata en los Data Marts.
Metadata es uno de los componentes más importantes del Data Mart. Metadata es una parte integral del ambiente de DM, sirve de igual forma que en un DW. La metadata del Data Mart permite que el análisis del DM encuentre dónde están los datos en el proceso para su descubrimiento y exploración. Contiene los siguientes componentes la metadata: identificación de los recursos de los datos, descripción de la personalización que ha ocurrido como el paso de datos desde el Data Warehouse al Data Mart y simple información descriptiva sobre el Data Mart incluyendo tablas, atributos, relaciones y definiciones,
La metadata del DM es creada por la actualización desde la carga de programas que mueve datos en el DM. Se necesita que estén ligados entre la metadata encontrada en el Data Mart y la encontrada en el Data Warehouse.

El modelado del Data Mart.
El modelado de un Data Mart se requiere dependiendo del tamaño de éste, ya que varios DM son pequeños e informales, para estos no es necesario un realizar un modelo. Otros son grandes y formales, en estos es normal hacer algunos procesos repetitivos y predecibles, para estos es necesario construir un modelo formal.

Limpieza de un Data Mart.
Como en una Data Warehouse, periódicamente el Data Mart requiere ser limpiado, esto comienza leyendo algunos datos siendo seleccionados y removidos para ser purgados, archivados y condensados. El criterio de limpieza puede ser basado en la fecha y el tiempo o puede ser basado en cualquier otro criterio.

Contenido de un Data Mart.
El contenido de un Data Mart son datos independientes que son necesarios para la toma soportar el proceso de toma de decisiones departamentales. Contiene el detalle de los datos y resúmenes de los mismos.

Estructura dentro de un Data Mart.
Los datos son estructurados dentro el Data Mart a lo largo de lineas *star-joins* y tablas normalizadas.

4.2. Descripción de la metodología de desarrollo de un sistema aplicado a un DW.

Como cualquier proyecto o sistema requiere de una metodología o un ciclo de vida, para el desarrollo de un Data Warehouse como sugerencia podemos tomar el siclo de vida que se propone en el libro"*The Data Warehouse lifecycle toolkit* en el capitulo *The business dimencional lifecycle*" que consta de las siguientes etapas (Cf Kimball,Reeves,Ross,Thornthwaite 1998:33-37);

- Plantación del proyecto
- Definición de los requerimientos del negocio
- Modelo dimensional
- Diseño físico
- Diseño y desarrollo del área de almacenamiento
- Diseño técnico de la arquitectura
- Selección e instalación del producto
- Especificación de la aplicación para el usuario final
- Desarrollo de la aplicación para el usuario final
- Despliegue
- Mantenimiento
- Manejo del proyecto

El ciclo de vida comienza con la planeación del proyecto con las expectativas que se quieren alcanzar dentro del proyecto. Se define un objetivo del proyecto de DW que se pretende lograr, se desarrolla un plan de trabajo que incluye toda la evolución del proyecto que se va a realizar desde la preparación, justificación del negocio, el costo asociado, las tareas criticas, recursos, requerimientos de logística, asignaciones de tareas, duración y secuencias.

La definición del requerimiento del negocio consiste en entender el negocio los usuarios y sus necesidades, esto impacta en todos los aspectos del ciclo. Los diseñadores de DW deben entender los factores clave manejando el negocio efectivamente para determinar los requerimientos y trasladarlos en un considerable diseño. Los requerimientos del negocio establecen la fundación de 3 pistas paralelas enfocadas en tecnología, datos y aplicaciones para el usuario final. Es necesario, además, determinar las necesidades de datos para dirigir los requerimientos analíticos de usuarios del negocio
La definición de los requerimientos del negocio es una determinante para el diseño del modelado de datos. Se comienza construyendo una matriz que representa los procesos clave del negocio y sus dimensiones, ya que nos sirve para determinar si es extensible el DW alrededor de la organización a lo largo del tiempo. Este modelo debe identificar la principal tabla de hechos, las dimensiones asociadas, los atributos, caminos jerárquicos y hechos. El final, el diseño lógico de una base de datos esta completo con las apropiadas estructuras de tablas y relación de llaves primarias y foráneas.

El diseño físico se centra en definir físicamente las estructuras necesarias para soportar el diseño lógico de las bases de datos. Los elementos primarios de este proceso incluyen definir estándares y poner a punto el ambiente de la base de datos. Primeramente se realiza la estrategia de indexar y particionar.

El desarrollo y diseño del almacenamiento de datos comprende tres subprocesos estos son extracción, transformación y carga de datos que ya han explicado en los capítulos anteriores.

El diseño de la arquitectura técnico establece una visión y el marco global de la arquitectura, para ello, se requiere de considerar los requerimientos del negocio, el actual ambiente técnico y la plantación estratégica de la dirección técnica

La selección del producto y la instalación se logra usando el diseño técnico de la arquitectura como marco, se requiere que los componentes específicos como la plataforma de hardware, los sistemas manejadores de bases de datos, herramientas de almacenamiento o las herramientas de acceso de los datos se requieren que sean evaluados y seleccionados. El proceso estandarizado de evaluación técnica es definido a lo largo con factores específicos por cada componente. Cada producto que es evaluado y seleccionado es luego instalado y probado a fondo para asegurar una integración con el ambiente del DW.

Para las especificaciones de la aplicación de usuario se recomienda definir el conjunto de las aplicaciones estándar para el usuario final ya que no todos los usuarios del negocio necesitan acceso *ad hoc* al DW. Aquí se describe una plantilla de reporte, los parámetros manejados por el usuario y el cálculo de requerimientos.

El desarrollo de la aplicación para el usuario final involucra la configuración de la herramienta del metadata y construcción de los reportes especificados. Esta aplicación es construida usando una herramienta de acceso de datos que provee productividad para el equipo de desarrollo de la aplicación; además, debe ofrecer un mecanismo para los usuarios del negocio que facilite la fácil modificación de los reportes de las plantillas existentes.

El despliegue representa la convergencia de tecnología, datos, y las aplicaciones finales del usuario final accesibles desde su escritorio. Adicionalmente el soporte y al comunicación de los usuarios en el marco de estrategias deben ser establecidas antes de que cualquier usuario del negocio tenga acceso al DW.

El mantenimiento y crecimiento es necesario para continuar enfocando en los usuarios del negocio con soporte y capacitación, también se debe poner atención en el *back room*, la aseguración que los procesos y procedimientos están en el lugar correcto para la operación eficaz del *warehouse*. El plan de mantenimiento debe incluir la estrategia de comunicación de amplio alcance.
Los procesos de ordenación deben ser establecidos para tratar con la demanda del usuario del negocio para la evolución y el crecimiento del DW, cuando el DW está creciendo y evolucionando después de identificar las prioridades se requiere comenzar desde el principio del ciclo nuevamente construyendo sobre lo que ya se ha establecido en el ambiente de DW.

La administración del proyecto comienza en paralelo con la etapa de Requerimientos del negocio y continua en paralelo de las demás etapas, su función principal es verificar que todas las actividades de cada etapa alcancen los objetivos y resultados propuestos

dentro de la etapa de plantación del proyecto, valida que el tiempo y la calidad sean los óptimos.

El siguiente cuadro ilustra las etapas que se han mencionado anteriormente

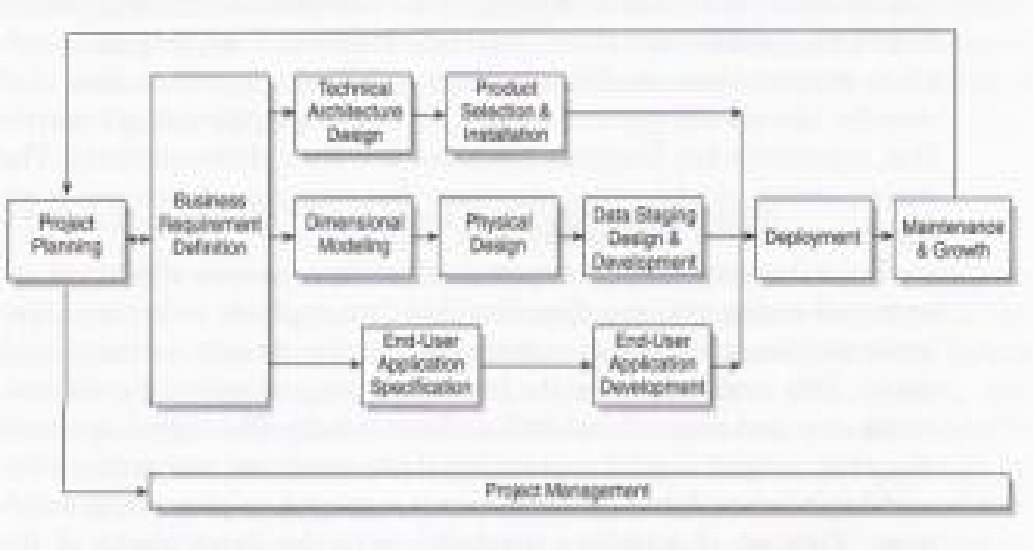


Ilustración 2: Diagrama del ciclo dimensional del negocio. (Kimball, Reeves, Ross, Thornthwaite 1998:33)

Para desarrollar un buen proyecto de DW se requiere seguir los pasos antes mencionados, pero en esta ocasión para comprensión del tema a continuación detallaré más los pasos de modelado, diseño y de arquitectura porque considero son los fundamentales para comprender el proceso.

Comencemos por el modelado, el modelado puede ser de dos tipos modelado de entidad relación o modelado dimensional.

El modelo entidad relación es una técnica de diseño lógico que busca eliminar la redundancia de datos. Es usado en la fase de administración de datos para la construcción de un DW.

El modelado dimensional es la técnica de diseño lógico que busca presentar los datos en un marco estándar que es intuitivo y permite alto funcionamiento de acceso. Cada modelo dimensional esta compuesto por una tabla con llaves de multipartes llamadas tablas de hechos, y un conjunto de pequeñas tablas llamadas tablas de dimensión.

En las siguientes imágenes se puede observar un ejemplo de cada uno.

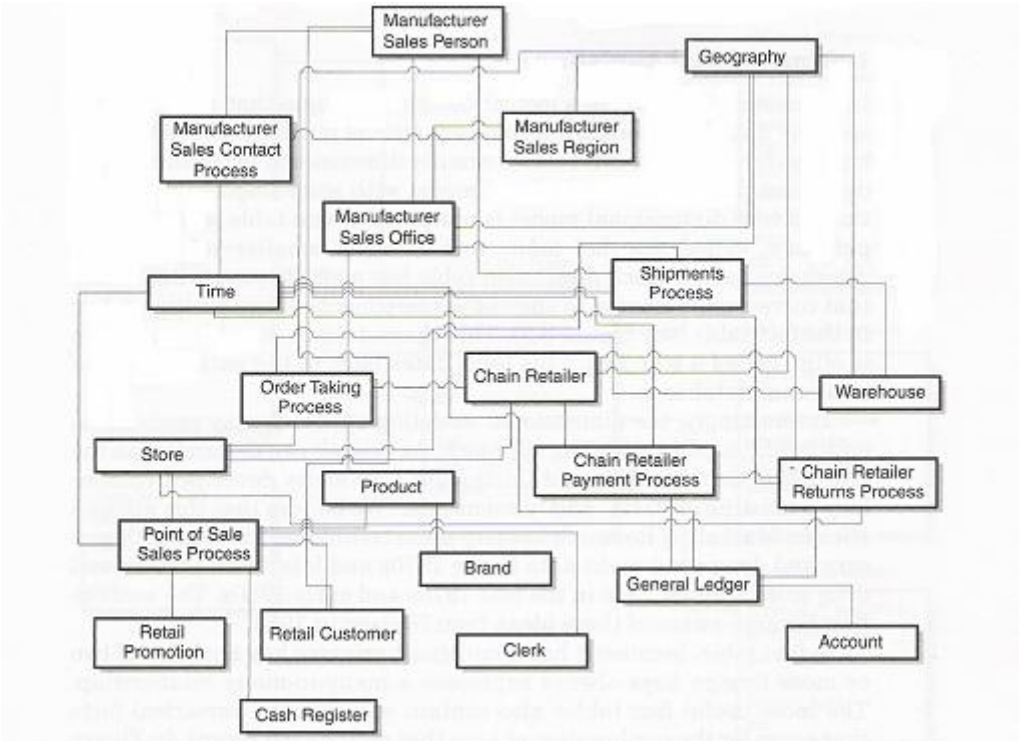


Ilustración 3: Modelo entidad relación (Kimball, Reeves, Ross, Thornthwaite 1999:143)

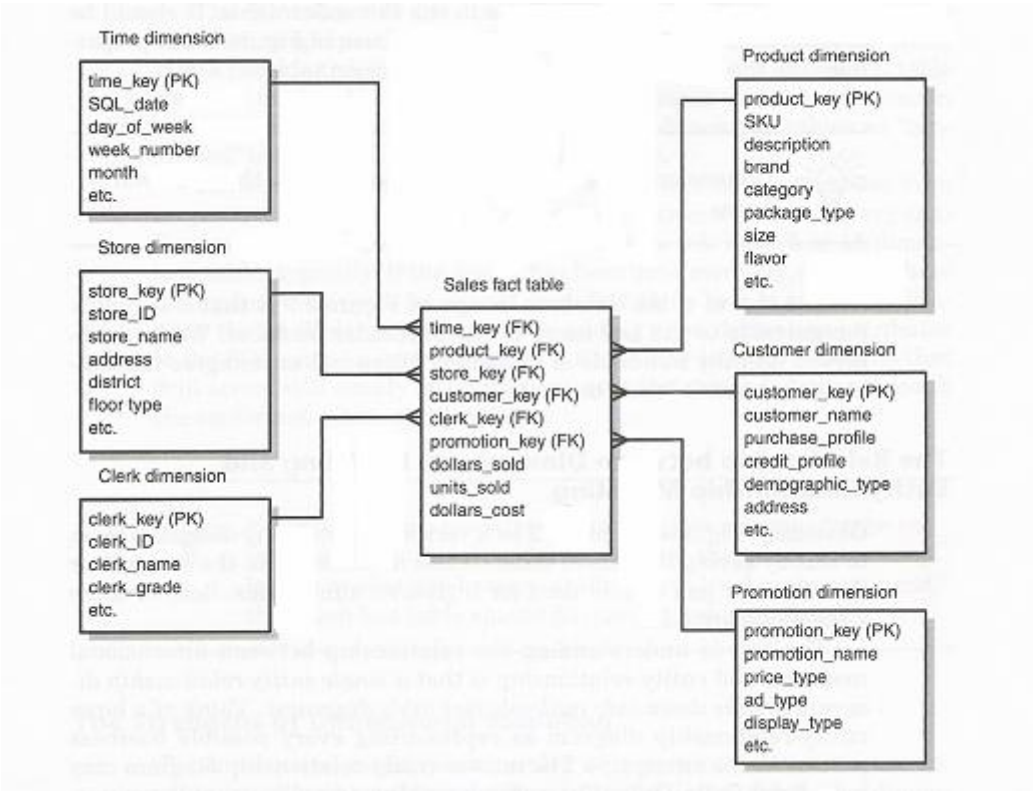


Ilustración 4: Modelo dimensional (Kimball, Reeves, Ross, Thornthwaite 1999:145)

La relación entre ambos modelos es que un diagrama singular entidad relación se divide en múltiples diagramas de tablas de hechos.

Un diagrama maestro de entidad relación puede contener cualquier proceso del negocio, por ejemplo ventas, órdenes, pagos, devoluciones todos en un mismo diagrama, es decir, en un solo diagrama se representan múltiples procesos que nunca coexisten en un único conjunto de datos en un punto consistente de tiempo haciendo que no sea demasiado complejo.
El primer paso para convertir de un diagrama entidad relación a un diagrama de modelo dimensional es separando el diagrama de entidad relación en procesos discretos del negocio y modelar cada uno separadamente.
El segundo paso es seleccionar las relaciones de muchos a muchos en el modelo entidad relación que contiene hechos numéricos y aditivos no llaves, y designar estos a las tablas de hechos.
El tercer paso es normalizar las tablas restantes en tablas planas con llaves únicas que conecten directamente a la tabla de hechos, las cuales se convertirán en tablas de dimensiones.

El modelo dimensional tiene ventajas importantes de las que el modelo de entidad relación carece. Primeramente, el modelo dimensional es un marco estándar y fiable. Los reportes, las herramientas de consulta y las interfaces de usuario pueden hacer fuertes suposiciones sobre el modelo dimensional para hacer que las interfaces de usuario sean entendibles y los procedimientos sean más efectivos. Cada dimensión es equivalente, es decir, todas las dimensiones pueden ser pensadas como simétricamente puntos de entrada dentro de las tablas de hechos. El diseño lógico puede ser casi independiente del patrón; las interfaces son simétricas, y las estrategias de consulta son simétricas. El modelo dimensional es extensible para acomodar elementos inesperados de datos nuevos y nuevos diseños de decisiones. Esto significa que todas las tablas pueden cambiar en lugar por simple adición de nuevos registros de datos en las tablas o ejecutando el comando SQL ALTER TABLE. A pesar de ello, los datos no necesitan ser cargados nuevamente. Extensible significa que las herramientas de consulta no requieren ser reprogramadas para acomodarse al cambio y también significa que todas las viejas aplicaciones sigan corriendo sin tener resultados diferentes.

La idea fundamental del modelado dimensional es que casi cada tipo de dato del negocio puede ser representado como una clase de cubo de datos, donde las celdas del cubo contienen valores moderados y los bordes del cubo define la dimensión natural de los datos. Por supuesto que se permiten más de tres dimensiones a este cubo se le llama hipercubo, aunque el término cubo y cubo de datos es usado indistintamente. Las dimensiones reales de los modelos de los negocios del mundo pueden contener entre 4 a 15 dimensiones. Modelos de 2 y 3 dimensiones son raros, los modelos que contienen más de 20 dimensiones puede ser injustificados.

El modelo dimensional se distingue entre los hechos y los atributos. Varios hechos en el mundo del negocio pueden ser numéricos o en pocas ocasiones pueden ser valores de texto. El diseñador debe sospechar que cualquier valor numérico en un campo de datos, especialmente si el valor numérico es de punto flotante, es probablemente un hecho y no un atributo.
Los atributos en cambio son usados frecuentemente en campos de texto, y ellos usualmente describen una característica tangible de algo. El más evidente atributo es la descripción de un producto. Los atributos textuales que describen cosas están organizados en dimensiones. En una base de datos de ventas al menudeo, por ejemplo,

tenemos la dimensión del producto, la dimensión del almacén, la dimensión del cliente, la dimensión de promociones y la dimensión del tiempo. Una dimensión es una colección de texto como atributos que están altamente relacionados con cada uno. Por ejemplo en la dimensión del producto y en la dimensión del almacén podemos combinar ambas obteniendo una relación de producto-almacén. En el modelo dimensional es de gran importancia las dimensiones y los atributos ya que la calidad de un DW es medida por la calidad de los atributos dimensiónales.

Para realizar el modelado dimensional se elabora el diseño lógico el cual incluye 4 pasos. Estos se mencionaron anteriormente en el capitulo dos. Para la construcción de un modelado dimensional se sugieren los siguientes puntos:

- El encargado de llevar a cabo el modelado y el administrador de las bases de datos deben ir un paso adelante para conducir los esfuerzos del modelado.
- El modelo dimensional debe ser revisado por los usuarios del negocio y por el equipo principal del proyecto. Esto permite la ventaja de realizar modificaciones al alcance.
- El equipo que organiza los datos está implicado en identificar las fuentes de datos necesarias para la población del modelo dimensional.
- La base de datos y los vendedores de las herramientas de acceso de datos también deberían estar involucradas en proveer los principios para el diseño de las bases de datos para optimizar sus productos.
- Se estima que puede llevar el proceso de modelado dimensional entre 2 semanas a 12 meses dependiendo de la complejidad de los datos, la industria, las llaves del negocio para tomar decisiones y el alcance del proyecto.

Diseño Físico.

El comienzo para diseñar el modelo físico es el modelo lógico. El modelo físico debe copiar el modelo lógico lo más posible, aunque se realicen varios cambios en la estructura de las tablas y columnas ya que el modelo físico incluye almacenamiento y mantenimiento de tablas que usualmente el modelo lógico no incluye. La mayor diferencia entra ambos modelos es el cuidado y detalle de la especificación física de las características de la base de datos, comenzando con los tipos de datos y continuando por la segmentación de la tabla, y los parámetros de la tabla
Para diseñar una estructura de datos física se recomienda:

- Comenzar por el diseño lógico, algunas herramientas pueden ayudar a la elaboración de estos diseños.
- Estandarizar los nombres Es importante seguir ciertos estándares que a continuación menciono el estándar para nombrar objetos de la base de datos. Esencialmente hay tres aspectos básicos que componen la definición de los elementos de los datos lógicos y físicos estos son;
    - Palabras principales describe el área de almacenamiento de los elementos de los datos. Responde a la pregunta ¿qué objeto es? Algunos ejemplos: cliente, producto, cuenta, ciudad, estado, código postal y región.
    - Palabras clases Las palabras clase describen la mayor clasificación de los datos asociados con los elementos de los datos. Responde a la pregunta ¿Qué tipo de objeto es? Algunos ejemplos: total, promedio, cuenta, código, fecha, bandera, ID, nombre, descripción, monto y numero.

- Calificadores. Son elementos opcionales que pueden definir o describir más a las principales o palabras clase. Estos pueden ser estrella, final, principio, secundario.

Se recomienda que los nombres lógicos y físicos sean idénticas y lo más descritos posibles.

- Determinar cuál columna tendrá el tipo de dato que difiere desde la representación dentro de la fuente del sistema. Por ejemplo en el código postal puede ser almacenado como entero en la fuente del sistema y como CHAR en el almacén de datos
- Determinar el tipo de dato para las columnas llave.
- Las llaves de fechas deben ser reducidas para ser eficiente las llaves sustitutas a pesar de la tentación de dejarlas como fechas. Aunque generalmente en las consultas de los datos se involucra una condición de tiempo.
- Determinar y especificar las columnas que permitan valores nulos
- Especificar las llaves primarias y foráneas relacionadas en el modelo de datos.
- Decidir si hay que explícitamente declarar llaves primarias/foráneas en la base de datos.
- Confirmar que todas las tablas y vistas del modelo estén incluidas en tablas.
- Representar todas las tablas índices en el modelo físico.

A Continuación se muestra un ejemplo de un modelo físico.

Table/column name	*Data type*	*Permit nulls?*	*Prim. Key*	*Comment*
calendar				**Calendar or period dimension table**
date_key	integer	n	1	Surrogate key
day_date	date	n		Date, can be used for date arithmetic
day_of_week_name	varchar(9)	n		Weekday, e.g., "Monday"
week_begin_date_key	integer	n		Key of this week's Monday
week_begin_date	date	n		Date of this week's Monday
calendar_week_num	smallint	n		Takes values 1..53. Week 1 begins first Mon in year
calendar_month_num	smallint	n		Takes values 1..12
calendar_month_name	varchar(9)	n		Month, e.g., "January"
calendar_quarter_num	smallint	n		Takes values 1..4.
calendar_year_num	integer	n		Calendar year carried as a number
year_month_num	integer	n		Year and month, carried as a number, e.g., 199801
fiscal_week_num	integer	n		Takes values 1..53. Week 1 begins first Monday in fiscal year
fiscal_month_num	integer	n		Takes values 1..12
fiscal_quarter_num	integer	n		Takes values 1..4.
fiscal_year_num	integer	n		Fiscal year carried as a number
weekday_ind	char(8)	n		Takes values "weekday" or "weekend"
hour				**Hour dimension**
hour_key	integer	n	1	Integer, 0..23, corresponds to hour in which purchase occurred
hour_time	time	n		Corresponding time, can be used for time arithmetic
am_pm_ind	char(2)	n		Takes values am/pm
peak_period_ind	char(8)	n		Takes values peak/off-peak. "peak" when hour btwn 15-20 (3p-8p)
product				**Product dimension**
product_key	integer	n	1	Surrogate key
brand_key	integer	n		Surrogate key, may be used for building aggregates
manufacturer_key	integer	n		Surrogate key, may be used for building aggregates
sub_category_key	integer	n		Surrogate key, may be used for building aggregates
category_key	integer	n		Surrogate key, may be used for building aggregates
product_name	varchar(15)	n		Short product name, use as column headings
brand_name	varchar(15)	n		Brand name, aggregates to manufacturer and sub-category
manufacturer_name	varchar(20)	n		Manufacturer name
sub_category_name	varchar(25)	n		Sub-category, aggregates to category
category_name	varchar(25)	n		Product category name
product_descr	varchar(125)	n		Long product name
package_size_amt	number(11,2)	y		Package size as a number
pkg_size_unit_name	varchar(15)	y		Units of package size, e.g., "ounce" or "quart"
package_size_group	varchar(15)	y		Package group, e.g., "family"
flavor_name	varchar(25)	y		Flavor, e.g., "chocolate"
store_market				**Store/market dimension**
store_key	integer	n	1	Surrogate key
market_key	integer	n		Key for markets, may be used to build aggregate tables

FIGURE 15.4 Beverage chain case study physical model.

Ilustración 5: Ejemplo de diseño físico. (Kimball, Reeves, Ross, Thornthwaite 1999:581)

Modelo de arquitectura técnico de alto nivel.

Una efectiva arquitectura requiere de la flexibilidad de los sistemas, facilidad de aprendizaje y mejoramiento de la productividad Para comprender mejor esto, hablaremos un poco sobre el marco de la arquitectura.

La arquitectura de datos define la granularidad, el área de arquitectura contiene el área de arquitectura de datos, el de arquitectura técnica y el de infraestructura, que a continuación explicaré.

El área de arquitectura de datos. Incluye el contenido del DW, este se refiere a los datos, una lista de datos más importantes para el negocio, el almacenamiento de datos que constituye el medio ambiente del almacén general y los recursos que son alimentados. Incluye el diseño físico y lógico de modelos de datos, agregaciones, jerarquías, etc. Todo volumen y el tiempo de datos en varios puntos del almacén.

El área de arquitectura técnica cubre los procesos y herramientas que se aplican a los datos. Esta área responde a la pregunta "¿Cómo?" De manera más precisa podemos plantearlo así:
¿Cómo llevar los datos en los recursos, ponerlos en un formato que reúna los requisitos del negocio y moverlos a un lugar para fácil acceso? En la arquitectura técnica hay dos principales subconjuntos los cuales tienen suficientes requerimientos diferentes para garantizar ser considerados de forma independiente. Estos son el *back room* y el *front room*. El primero, es la parte del almacén responsable de reunir y preparar los datos, se le puede llamar también adquisición de datos. El segundo, es la parte responsable de entregar los datos a los usuarios, también se le puede llamar de acceso a los datos. Cada uno tiene sus servicios y sus componentes de almacenamiento de datos.
La arquitectura técnica para todo el almacén es la combinación de código cliente, las utilerías hechas en casa, y las herramientas disponibles.

El área de infraestructura tiene que ver con las plataformas que reciben los datos y los procesan. La infraestructura es el plano físico de un DW, es decir, los tubos y plataformas que sostienen y transportan los datos y soportan las aplicaciones.

Parte de las tres áreas de la arquitectura en su conjunto involucran la creación de muchos dibujos y documentos en varios niveles. A estos niveles se les llama modelos, son el medio para la comunicación de la arquitectura. A continuación explicare brevemente estos niveles de arquitectura.

El nivel de requerimientos de usuario es explícitamente no técnico. Los planeadores del sistema deben ser disciplinados y no buscar soluciones técnicas en este nivel, pero si tienen que entender la especialización de las fuerzas del negocio y las condiciones divisorias que afectan el proyecto de DW.

El nivel de modelo de arquitectura es el primer nivel que responde a los requerimientos; procesa los componentes mayores de la arquitectura que deben estar disponibles para enfocarse a los requerimientos. En este nivel la perspectiva del sistema se enfoca en

si varios componentes técnicos pueden comunicar el uno con el otro, si las suposiciones administrativas que rodean el empleo de las tecnologías son razonables y si la organización tiene recursos que soporten las tecnologías.

En el nivel de modelo a detalle se realizan las especificaciones funcionales de cada uno de los componentes de la arquitectura. Debe incluir suficiente información para que sirva como guía de una confiable implementación confiable para el grupo de trabajo. Este modelo debe ser bastante completo para crear un contrato legal de modo que cuando el trabajo sea hecho pueda ser soportado a la especificación funcional cuando la implementación esta completa. El nivel de detalle es muy útil para el establecimiento de las expectativas y de comunicación de la intención

El nivel de implementación es el resultado del modelado de detalle. Para el software entregable es en sí mismo el código. Para el área de datos es la definición del lenguaje de los datos usados para construir la base de datos, y en varias formas, los datos en si mismos. Para varios componentes técnicos, la implementación esta expresada en metadatos, si son tareas agendadas, especificaciones extraídas, o un conjunto de herramientas con parámetros de consultas.

A continuación se comentará sobre el modelado de arquitectura técnico. Como lo mencione anteriormente la arquitectura técnica se divide en dos aplicaciones el *back room* y el *front room*, ambas aplicaciones interactúan de varias formas como se muestra en el siguiente diagrama.

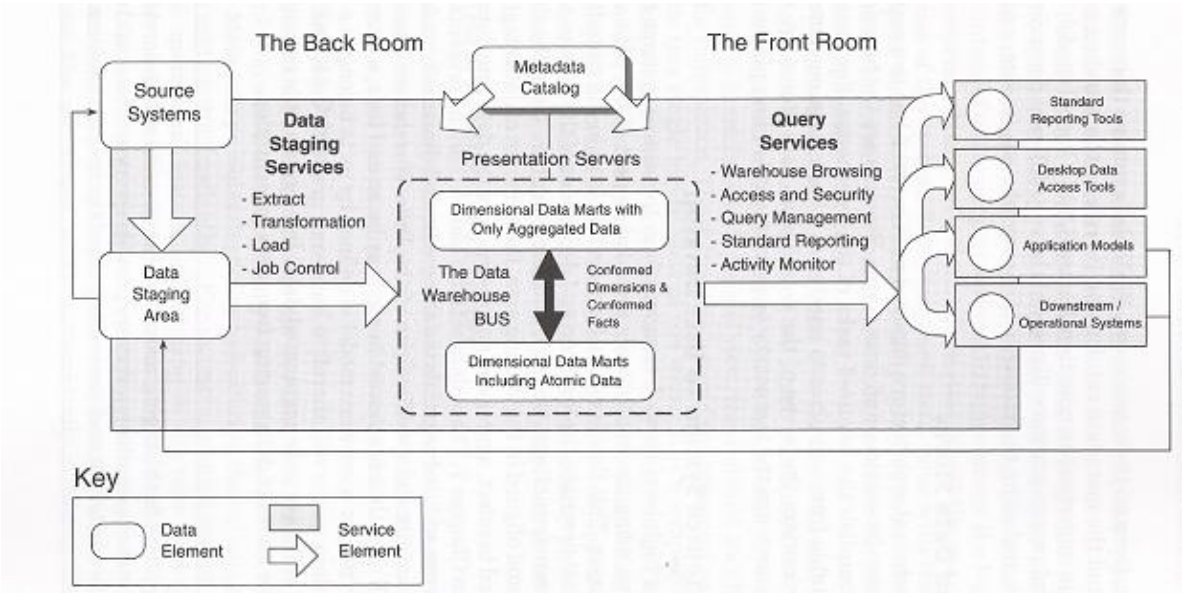


Ilustración 6: Diagrama Alto nivel del modelo técnico de arquitectura (Kimball, Reeves, Ross, Thornthwaite 1999:329)

Mientras que los requerimientos responden a la pregunta ¿qué se necesita hacer?, la arquitectura técnica responde a la pregunta ¿cómo lo vamos a realizar? La arquitectura técnica describe el flujo de datos desde los recursos del sistema para la toma de decisiones y la transformación de los datos, especifica las herramientas y tecnologías que se requiere para hacer más fácil el trabajo.

En el diagrama anterior (Ilustración 26) se muestra los elementos de un nivel superior de arquitectura técnica, ésta tiene dos tipos principales de componentes que son los servicios y el almacenamiento de datos. Los servicios son las funciones que se necesitan para lograr las tareas requeridas en el almacén de datos. Por ejemplo, copiar una tabla desde un lugar a otro es un servicio básico de movimiento de datos. Los almacenes de datos son el lugar de aterrizaje ya sea permanente o temporal de los datos.
Los datos se mueven desde los sistemas fuente al área de almacenamiento usando las aplicaciones provistas como parte de los servicios de la capa de almacenamiento de datos. Este flujo es conducido por metadatos los cuales están contenidos en catálogos que describen las locaciones, definen los recursos y objetivos, las transformaciones, los tiempos y dependencias de los datos. Una vez que los datos son combinados y alineados en esta área la misma cantidad de servicios de datos en almacenes son usados para seleccionar, agregar y ser reestructurados dentro de los conjuntos de datos. Estos, son cargados dentro de las plataformas del servidor y unidos vía las dimensiones y hechos conformados que son las especificaciones del *Data Warehouse bus*. Todos los datos que son accedidos en cualquier camino por el usuario final, por las herramientas de consulta, por un generador de reportes o por algún otro un software modelo son definidos para pertenecer a un Data Mart.
El flujo desde el sistema fuente hasta el escritorio del usuario es soportado por metadatos desde el catalogo. Finalmente, los usuarios pueden acceder a los datos con herramientas de acceso de datos de escritorio. Estas son por lo general una combinación de herramientas de *front-end* creadas con diversos productos.
Como ya hemos comentado sobre los flujos de datos, continuemos con el servidor de presentación, desde el punto de vista del usuario, solo debe haber un lugar par ir por la información. La línea punteada que se puede apreciar en el diagrama representa la vista lógica del mundo; es una simple capa que hace visibles todos los DM para los usuarios del la comunidad. Esta capa puede ser creada de varias formas como por ejemplo por medio de aplicaciones de servidor, metadatos, *gateways*, tablas físicas o una combinación de estas.

El catalogo de metadatos juega un papel muy importante en la arquitectura como *metadata driven*. Provee parámetros e información que permite a las aplicaciones realizar sus tareas.
El catalogo de metadatos en este punto es un concepto lógico, un cambio en este puede reflejarse en todas las partes de la arquitectura. En casi todos los casos, no es practico llevarse toda la información en un lugar común, el metadata viven en varias herramientas, programas y utilerías que hacen que el DW funcione.

La infraestructura y metadata proveen los fundamentos de todos los elementos de la arquitectura. La infraestructura para el Data Warehouse incluye el hardware, la red y funciones de bajo nivel como la seguridad. La metadata provee el mismo tipo de soporte para el *back room* y *front room* que la infraestructura.

Muchos son los factores que determinan una apropiada infraestructura para una buena implementación y varios de ellos no necesariamente deben ser técnicos.
El primer determinante son los requerimientos del negocio, el negocio debe determinar el apropiado nivel de detalle que el almacén necesita para llevar a través de lapsos de tiempo los datos, esto nos dice cuantos datos necesita administrar la infraestructura. Otro requerimiento del negocio determina la frecuencia de carga de los datos así como la complejidad de las reglas del negocio que son necesarias durante la transformación

del proceso. Otro determinante tiene que ver con las habilidades especificas y la experiencia de los implementadotes del Data Warehouse. Las políticas y otros temas organizacionales juegan un papel determinante en la infraestructura.

El principio básico cuando se considera plataformas de hardware es recordar que el almacén crecerá rápidamente en los primeros 18 meses en términos de los datos y su uso.
El primer paso para el proceso de selección de la plataforma es entender los requerimientos. No es suficiente tener entendido qué plataforma debe hacerse y cómo debe representarse sino también es necesario saber los requerimientos del negocio: el tamaño de los datos, la volatilidad, el número de usuarios, el número de procesos del negocio, la naturaleza del uso, preparación técnica disponibilidad de software, recursos financieros. etc.
La velocidad del disco y la memoria son especialmente importantes para el Data Warehouse porque las consultas pueden ser de datos intensivos. Una solicitud de un sistema de transacción típicamente recupera un solo registro desde la tabla optimizada. Una consulta de Data Warehouse puede requerir la agrupación de miles de registros a través de varias tablas.

Existen muchas plataformas para soportar el data warehouese con casos de éxito en la implementación, algunos Data Warehouse son implementados en productos de main-frames y otros en bases de datos multidimensionales llamadas *multidimensional on line analytical processing* (MOLAP)

También es de vital importancia en la arquitectura técnica elaborar un plan de respaldo y de seguridad para el Data Warehouse ya que varios de los datos contenidos en este pueden ser de carácter financiero, legales, o de alto impacto para la organización. El administrador del Data Warehouse debe estar en el entendido de todos los temas de seguridad que están en torno a éste, para que tenga la facultad de supervisar y contratar al experto de seguridad el cual dedicará esfuerzos al Data Warehouse. Un administrador de Data Warehouse debe también entender Internet porque casi cada componente de éste esta siendo readecuado para trabajar en un entorno web.
Algunas vulnerabilidades que puede estar involucrado el Data Warehouse: robo por medio de que un ente desactivo los permisos; destrucción intencional como un golpe con un instrumento pesado, prender fuego, derramar líquidos; desastres naturales como inundaciones, terremotos, humedad; descargas eléctricas e interferencias eléctricas; perdidas de información por descuido, activos secuestrados, descubrimiento de código, descubrimiento de información sensible entre muchas otras.
Una de las principales razones por las que el equipo encargado del Data Warehouse no agrega seguridad es por el sentimiento de que ese tema no les corresponde pero también podemos decir que es por el hecho de que no son especialistas en el tema de seguridad, sin embargo es muy importante tener en cuenta la seguridad más aún si los datos son confidenciales y de alto impacto para la organización.
Hay varias formas de tener seguridad en nuestro Data Warehouse una de ellas es poner un firewall en el servidor para la transferencia de datos. Otra forma es encriptando el código secreto, sólo el administrador de los códigos del Data Warehouse debe tener acceso a las contraseñas de la encriptación ya sea por medio de la encriptación de llaves simétricas, de llaves publicas o certificados.
Un administrador del Data Warehouse debe proteger la información con un programa de seguridad que contemple los siguientes puntos;

- Conciencia. La necesidad de seguridad debe estar continuamente reforzada a través de un constante proceso educativo.
- Soporte ejecutivo. El administrador ejecutivo debe tener los conocimientos sobre la importancia de la seguridad y de los elementos principales de la seguridad.
- Políticas. La seguridad debe estar implementada a través de un conjunto de políticas comprensibles que sean visibles y procesables,
- Vigilancia. Una efectiva seguridad envuelve una continua vigilancia
- Sospecha. Una actitud de sospecha en el administrador del Data Warehouse es necesaria para permitir la disminución de las vulnerabilidades.
- Renovación. La seguridad debe ser dinámica y continua.

Algunas recomendaciones sobre seguridad en la Data Warehouse son las siguientes:

- Instalar un antivirus para revisar todo el software sobre todo en las PC de los usuarios finales.
- Quita los discos flexibles del ambiente. Si se requiere que un usuario lea un disco flexible se debe revisar manualmente para que no contenga un virus.
- Quita el modem local del ambiente. Prohíbe el uso de *dial-out modems* dentro de las facilidades corporativas desde cualquier usuario final.
- Controla todo el software instalado en maquinas internas
- Asigna a usuarios contraseñas y *passwords* las cuales deberán memorizar y usar.
- Quita todos los servicios no usados del servidor.
- Implementa un programa de capacitación para la seguridad.
- Implementar un programa de auditoria de la seguridad

Ejemplo de plan de arquitectura con cuadros

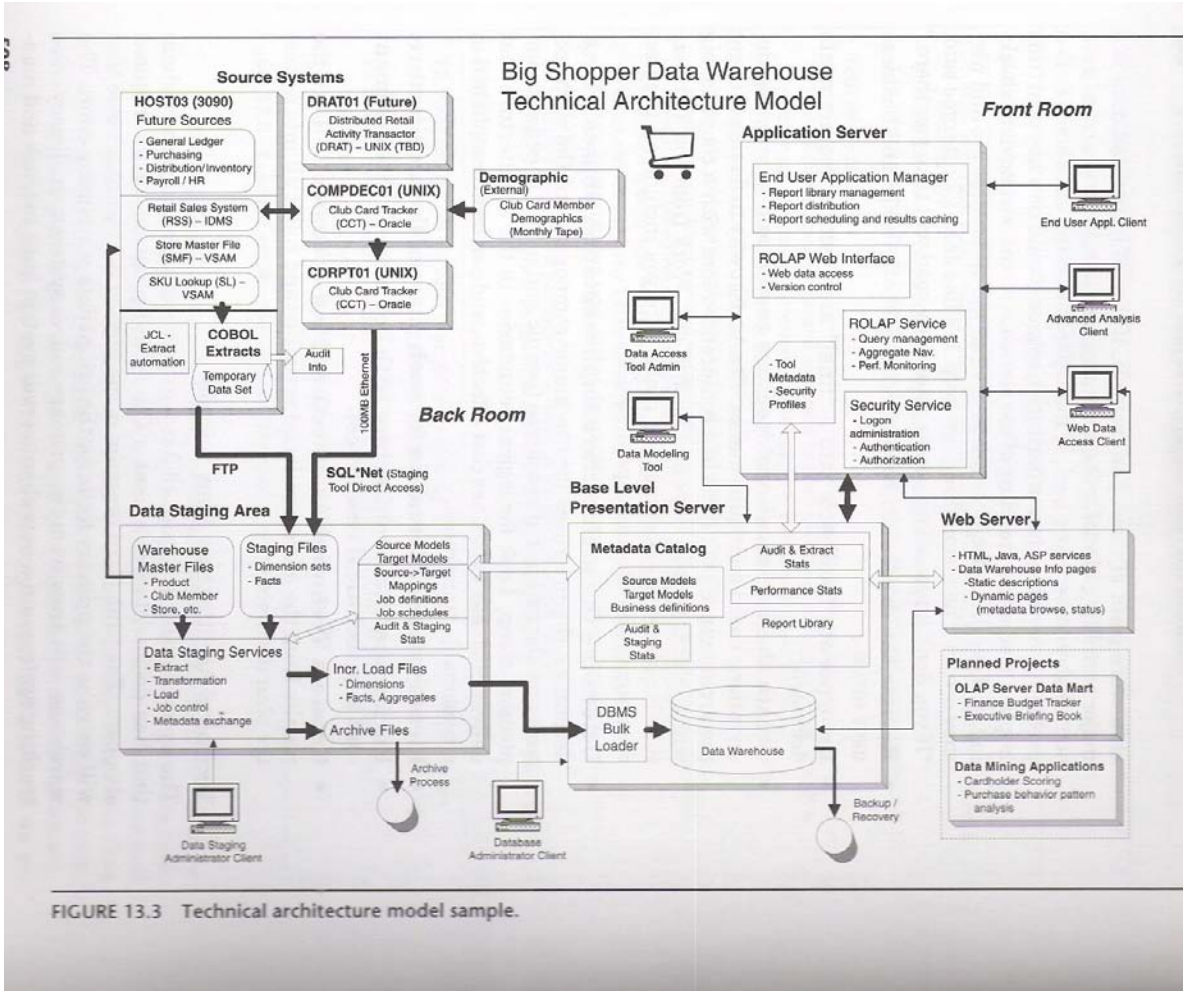


FIGURE 13.3 Technical architecture model sample.

Ilustración 7: Diagrama de arquitectura técnico (Kimball, Reeves, Ross, Thornthwaite 2002:508)

CAPITULO V DISEÑO DE UN DATA MART PARA LA COORDINACION DE INFORMATICA EN LA FACULTAD DE CONTADURIA Y ADMINISTRACION.

En este capitulo se examinará un caso practico de estudio dentro de la Coordinación de Informática de la Facultad de Contaduría y Administración siguiendo los pasos comentados en el capítulo anterior lo más apegado posible para poder ejemplificar el diseño de un Data Mart. Siguiendo la metodología que hemos estado estudiando podremos entender cómo analizar un determinado problema que su solución esté orientada al diseño de un Data Mart o un Data Warehouse para después formar un equipo de trabajo que lleve a cabo la implementación.
Lo que se quiere lograr en este apartado es un ejemplo de diseño de Data Mart aplicado en un caso real que sirva como diagnostico o propuesta para que en un futuro se pueda retomar para que alguien logre implementarlo, sólo se pretende aportar en esta tesis el diseño con sus respectivos diagramas y finalmente la propuesta para la Coordinación de Informática porque el tema es muy amplio y requiere de la coordinación de un equipo de trabajo para el desarrollo e implementación del mismo.

Introducción

Antecedentes de la FCA (WEB:03)

La Facultad de Contaduría Administración (FCA) e Informática es como su nombre lo indica la escuela que imparte dichas carreras a los estudiantes que han cumplido con el requisito de termino de los estudios de media superior con la visión de ser un modelo educativo en un plano internacional tanto para la investigación como para la formación de profesionistas.
Algunas fechas importantes de la FCA son:
En 1929 como parte de la ley de autonomía de la Universidad Nacional se crea la Facultad de comercio y administración formada por la Escuela Superior de Comercio y Administración y la Escuela Superior de Administración Publica.
En 1957 se crea la licenciatura en Administración como resultado de las condiciones del sistema económico mexicano.
En 1972 Se crea la Asociación Nacional de Facultades y Escuelas de Contaduría y Administración, ANFECA; por el compromiso que adquiere la FCA con la enseñanza de sus disciplinas en el nivel nacional y en el internacional,
En el año de 1985 la FCA dio inicio a la carrera de informática a causa de las necesidades que tenía el país causa del crecimiento tecnológico que comenzó a desarrollarse en esos momentos.

Su población estudiantil es de 14,318 estudiantes, de los cuales 12,698 pertenecen al sistema escolarizado y 1,620 al sistema abierto. Del total de alumnos inscritos al sistema escolarizado, 6,857 pertenecen a la licenciatura en Contaduría, 5,096 a la licenciatura en

Administración y 745 a la licenciatura en Informática. Por otra parte, del total de alumnos inscritos en el sistema abierto, 781 pertenecen a la licenciatura en Contaduría,713 a la licenciatura en Administración y 126 a la licenciatura de informática.

Coordinación de Informática.

El objetivo principal de las coordinaciones es apoyar a las jefaturas de carreras para cumplir con la meta de controlar a los profesores, materiales e instalaciones para que el alumno pueda tener un desarrollo integral académicamente en su área de conocimiento específico, es decir, por ejemplo la de informática solo atiende problemas que se detectan dentro de las materias asignadas a ésta.

Operativamente, cada Jefatura tiene a cargo varias coordinaciones, con el fin de distribuir las áreas de conocimiento de manera más uniforme:
La Jefatura de carrera de Administración tiene a su cargo las siguientes coordinaciones:

- Recursos humanos
- Metodología (Investigación y Ética)
- Operaciones
- Administración básica
- Administración avanzada

La Jefatura de la carrera de Contaduría tiene a su cargo las siguientes coordinaciones:

- Contabilidad Básica y Avanzada
- Derecho
- Finanzas
- Fiscal
- Auditoria
- Costos y presupuestos

La Jefatura de la carrera de Informática tiene a su cargo las siguientes coordinaciones:

- Informática
- Matemáticas
- Economía

Las coordinaciones trabajan como enlace entre los profesores, los alumnos, las dependencias de la FCA y las dependencias externas como se muestra en la figura.

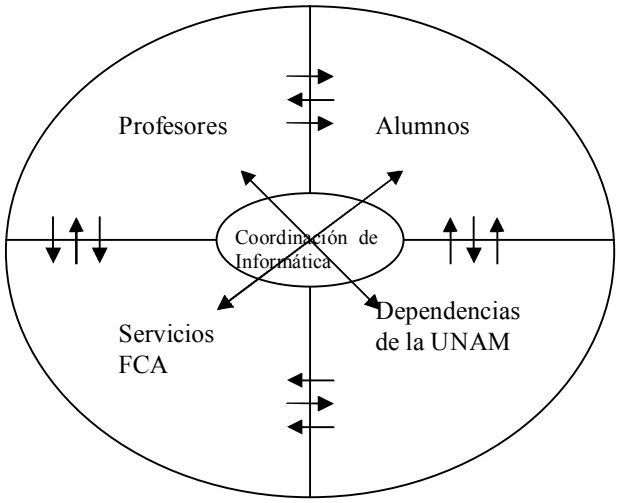


Ilustración 1: Relaciones de la coordinación de informática

Las coordinaciones se encargan de un número determinado de materias y grupos al semestre y varia entre cada coordinación. La coordinación de informática cuenta aproximadamente con 50 grupos para la licenciatura en informática con 12 materias, 25 grupos de informática básica para la licenciatura en contaduría y 25 grupos de para la licenciatura en administración, sólo en el semestre non, pues en el semestre par ésta no se imparte.

Las coordinaciones pueden relacionarse a través de actividades conjuntas en eventos cuando la temática de éste tiene que ver con materias de varias coordinaciones.

Existen muchas actividades que realizan las coordinaciones pero no existe ningún tipo de manual que las describa ya que estas actividades son muy específicas y dependen de las necesidades de tres factores los profesores, los alumnos y los materiales. Por ejemplo:

- Si un profesor solicita una clase adicional y requiere del uso de un laboratorio, se coordina con el centro de informática la utilización de éstos, de acuerdo a la disponibilidad de aulas, tanto para el tipo de laboratorio, como para el uso que le dé, las fechas y horas de uso.
- Si un profesor requiere de los servicios bibliotecarios y no cuenta con la credencial de servicio, por ejemplo, antes de comenzar el semestre, se coordina con la jefatura de servicios al público, extendiéndose una carta con la información del grupo asignado, para que sea la biblioteca quien autorice el servicio.
- Si un profesor requiere los servicios del correo electrónico institucional, se canaliza con el centro de informática CIFCA, para que se le abra una cuenta, previa solicitud del servicio.
- Cuando hay un evento académico relativo a cualquier asignatura de la carrera en que se encuentre inscrito el alumno, se coordina la utilización del espacio físico,

como auditorios, laboratorios, audiovisuales, salones, etc., se genera las constancias de asistencia y participación al evento, así como los demás insumos que se requieran para la celebración de la actividad, como sonido, equipo audiovisual, etc.

- Si un alumno tiene problemas con alguna calificación o se presentan problemas de comunicación con los profesores en el periodo de evaluación, se hace el enlace con el profesor y la coordinación actúa como mediador, si el caso lo requiere. En todo caso se establecerán las condiciones para que no salgan perjudicados alumnos con las decisiones de los profesores, o inclusive profesores con la actuación de los alumnos.
- Si un alumno o profesor desea participar en la elaboración de algún artículo técnico, o relacionado con algún tema de las áreas de conocimiento de los planes de estudio, y quieren hacer un artículo para alguna de las revistas de la Facultad, se les orienta y canaliza con el área de publicaciones de la FCA, para obtener el apoyo necesario.
- Si un alumno tiene interés en irse al extranjero como intercambio académico, ya sea como opción de titulación o para cursar alguna asignatura en otra escuela, se le canaliza con la secretaría de relaciones y extensión universitaria, dando la orientación necesaria al respecto.

No hay que olvidar que las coordinaciones, no disponen de los medios ni la infraestructura para dar todos los servicios que requieren los alumnos y profesores, pero sirven de enlace y contacto para que las áreas o departamentos de la Facultad, que si pueden llevar a cabo estos servicios, puedan realizar su trabajo en beneficio de los interesados.

Planteamiento del proyecto

Objetivo
Elaborar los diagramas y la propuesta para satisfacer las necesidades de información de la Coordinación de Informática de la Facultad de Contaduría y Administración para determinar la asignación de profesores a los grupos, por medio de un DM o un DW que le permita al usuario tener la información más ágil, centralizada y sencilla para su explotación.

Alcance
Se analizaran las necesidades de información siguiendo los pasos mencionados en el capitulo anterior.

Plantación
Se determinaran el objetivo, los alcances y la situación actual de la información requerida de la Coordinación de Informática.

Requerimientos del negocio.
Se obtendrá el resumen de los criterios a considerar para la obtención de la información resumida para la toma de decisión en la asignación de los profesores y el análisis del requerimiento.

Modelado Dimensional.
Se determinaran la tabla de hechos principal y las tablas de dimensiones con sus atributos y se mostraran en un diagrama de modelo dimensional.

Modelado Físico.
Se describirán los atributos de las tablas obtenidas del modelo dimensional.

Arquitectura.
Se elaborará un diagrama que muestre como puede estar estructurada la arquitectura a nivel básico.

Desarrollo de la aplicación para el usuario final
Se ejemplificara con imágenes el resultado que podría esperar el usuario final de una solución integral de Data Mart.

Esta tesis no incluye los siguientes pasos del ciclo mencionado en el capitulo anterior:

- Diseño y desarrollo del área de almacenamiento
- Selección e instalación del producto
- Especificación de la aplicación para el usuario final
- Desarrollo de la aplicación para el usuario final
- Despliegue
- Mantenimiento

La razón de que no se va a llevar a cabo los puntos anteriores es porque se planteó desde un principio que es esta tesis sólo se llevaría a cabo la parte de diseño, ya que para llegar a la implementación se requieren de varias personas que realicen el proyecto.

Las demás coordinaciones, en términos generales, llevan a cabo los mismos procesos y por lo tanto el levantamiento de información sería redundante, respetando los estilos de dirección y lineamentos particulares de cada coordinación y jefatura de carrera respectiva

Situación actual

1. El coordinador requiere de comunicación para obtener información, sus fuentes son manuales (impresos), escritas (e-mail) y documentos electrónicos.
Algunos de los problemas que se podrían presentar son los siguientes:

- Podrían existir diferentes versiones de la información
- La veracidad de la información podría tener inconsistencias al ser tomadas de diversas fuentes.
- Podrían generarse dependencias en la obtención de la información por ser las fuentes de diversos sistemas.
- Podría constar de muchos formatos la información consultada, ya que no existen prototipos y estándares del manejo de información.

2. Existen varios sistemas aislados.
Algunos de los problemas que se podrían presentar son los siguientes:

- Falta de oportunidad en la información.
- Varios dueños de la información.
- Falta de centralización.
- Dependencias de terceras personas para acceder a la información.

3. La información sobre los profesores se agrupa en Microsoft Excel por el coordinador
Algunos de los problemas que se podrían presentar al utilizar este software en lugar de una base de datos son los siguientes:

- Los datos podrían no estar normalizados.
- La generación de consultas puede ser tardada.
- El usuario tienen que alimenta el archivo de Excel por el mismo.
- El manejo de datos históricos podría llegar a ser difícil.
- Redundancia.
- Problemas de acceso para usuarios indirectos, porque solo el coordinador es propietario del archivo de Excel.

Descripción de las fuentes actuales

La Coordinación de Informática cuenta con varias fuentes de datos, unos son sistemas y otros son archivos manuales, el siguiente diagrama muestra en su núcleo la Coordinación de informática y alrededor están las fuentes de información que requiere para tomar la decisión de asignación de profesores.

Las coordinaciones están ligadas entre ellas porque requieren de datos en común, como por ejemplo las materias de las diversas coordinaciones pueden ser impartidas por un mismo profesor.
El área de sistemas de Personal docente maneja las bases de datos de Evaluaciones y asistencia. Por medio de un proceso las encuestas que se aplican cada semestre a los alumnos de la facultad son cargadas al sistema. En cuanto a la de la base de datos de asistencias se cargan los datos por medio de un proceso el cual se ingresan datos de la firma autógrafa, cabe aclarar que actualmente ya esta desarrollado y puesto a prueba un sistema de huella digital. El coordinador de cada área solicita la información que necesita cada semestre a esta área, esta área realiza las consultas necesarias a las bases y le devuelve un archivo de Excel con los datos pedidos.
El área de CIFCA tiene la base de datos de Asignaciones en el cual se registran a los profesores, grupos y materias. La participación de la coordinación en este sistema es dar de alta, baja y modificar lo referente a profesores en cuanto a su asignación dentro de las materias, así como consultar los horarios disponibles. A su vez la coordinación solicita información de la base para obtener datos de los grupos y materias Actualmente en CIFCA se están desarrollando varios sistemas uno es el de publicaciones y otro es el de CV de profesores. Actualmente se toman archivos manuales de cada área para actualizar la información.

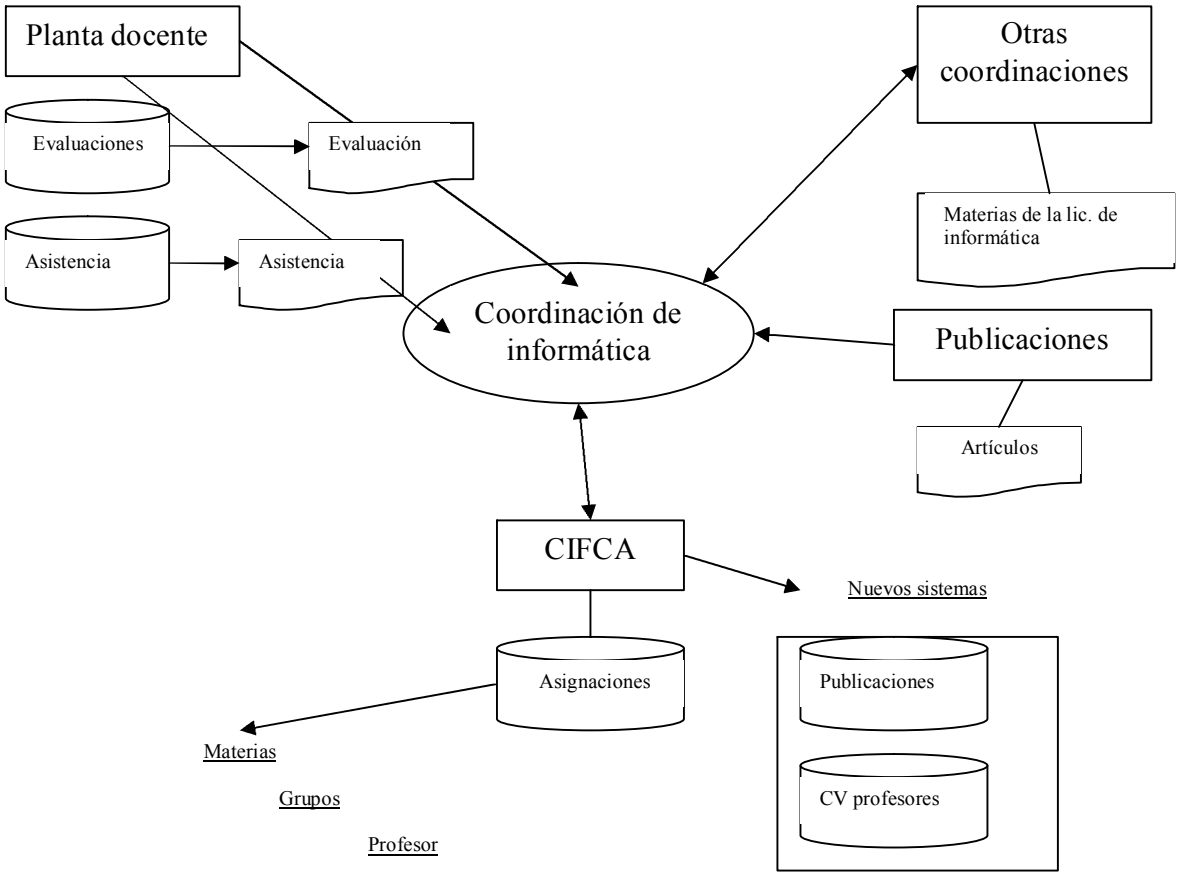


Ilustración 2: Fuentes de Información requerida

Requerimientos de la FCA

Obtener información precisa, oportuna, clara, veraz que permita la toma de decisión para la asignación de profesores a grupos dentro de las materias de informática, así como conocer la distribución de los laboratorios, salones y audiovisuales en cada semestre, de acuerdo a los siguientes criterios:

1- Asignación de profesores.
- Evaluaciones semestrales anteriores (Sistema).
- Área de especialidad docente (CV archivos manuales o impreso).
- Participación académica: conferencias, publicaciones y artículos.
- Experiencia docente: materias impartidas, número de grupos.
- Asistencias.
- Disponibilidad de profesores.

2- . Consulta de laboratorios y salones asignados.

Asignación de Salones
• EL AREA DE PERSONAL DOCENTE
Asignación de laboratorios
• CIFCA Centro de Informática de la Facultad de Contaduría y Administración

3- Usuarios que lo solicitan.

Usuarios directos
Coordinador de informática
Usuarios indirectos
Jefe de la división de informática Jefe de exámenes profesionales Otros

Análisis de los requerimientos.

Descripción del requerimiento.

- Extracción de información

  - Bases de datos, fuentes de información.

  Las fuentes de información como antes comenté son archivos manuales y en ocasiones sistemas ya desarrollados conectados a una base de datos, los propietarios de estas bases de datos son diferentes, entre ellos están CIFCA, Personal docente y Administración escolar. En general el manejador en común para la mayoría de las bases de datos de los sistemas de las dependencias antes mencionadas es PostgresSQL Estas bases de datos son controladas por los arquitectos de ellas, el líder de proyecto o el creador del software.
  Los datos que se manejan de un profesor básicamente son los siguientes: datos personales, adscripción, consultoría, distinciones, niveles/grados académicos, experiencia laboral, idiomas, libros, reportes técnicos, reseñas, tesis dirigidas, etc., docencia, artículos publicados, y participación en congresos.

- Calculo de métricas

  - Promedios, porcentaje y, totales.

  Los datos que se quieren obtener deben ser lo más abstracto posibles porque esto simplifica al usuario su búsqueda de la información.
  Para la Coordinación es necesario saber el promedio de evaluación por materia y semestre de cada profesor, el porcentaje de asistencia en el semestre en un determinado grupo y materia de cada profesor, el número total de publicaciones ya sean libros de texto o artículos, número de participaciones como ponentes en congresos y cursos o congresos a los que asistieron y el numero de grupos y materias que un profesor dio en el semestre.

- Resultados esperados.

  - Consultas

  Las consultas que realizaría el usuario deberán ser lo mas sencillas posibles y en un ambiente agradable que pueda entender y ser fácil de usar; el usuario espera que el tiempo de su búsqueda de información sea oportuna, es decir, que el tiempo de obtención de la consulta sea el adecuado, por ejemplo en el caso del promedio de las evaluaciones requiere obtener antes de que comience el siguiente periodo escolar para que tenga un uso para poder tomar la decisión de la asignación del profesor.

Modelo dimensional y relacional

El siguiente diagrama muestra la tabla principal, (la tabla de hechos) que es la de profesor y las tablas dimensionales de evaluación, disponibilidad, asistencia, tiempo, nivel académico, materia, grupo, especialidad, publicación, conferencia y estado; cada una con sus atributos.

Diagrama Multidimensional para la tabla de hecho Profesor

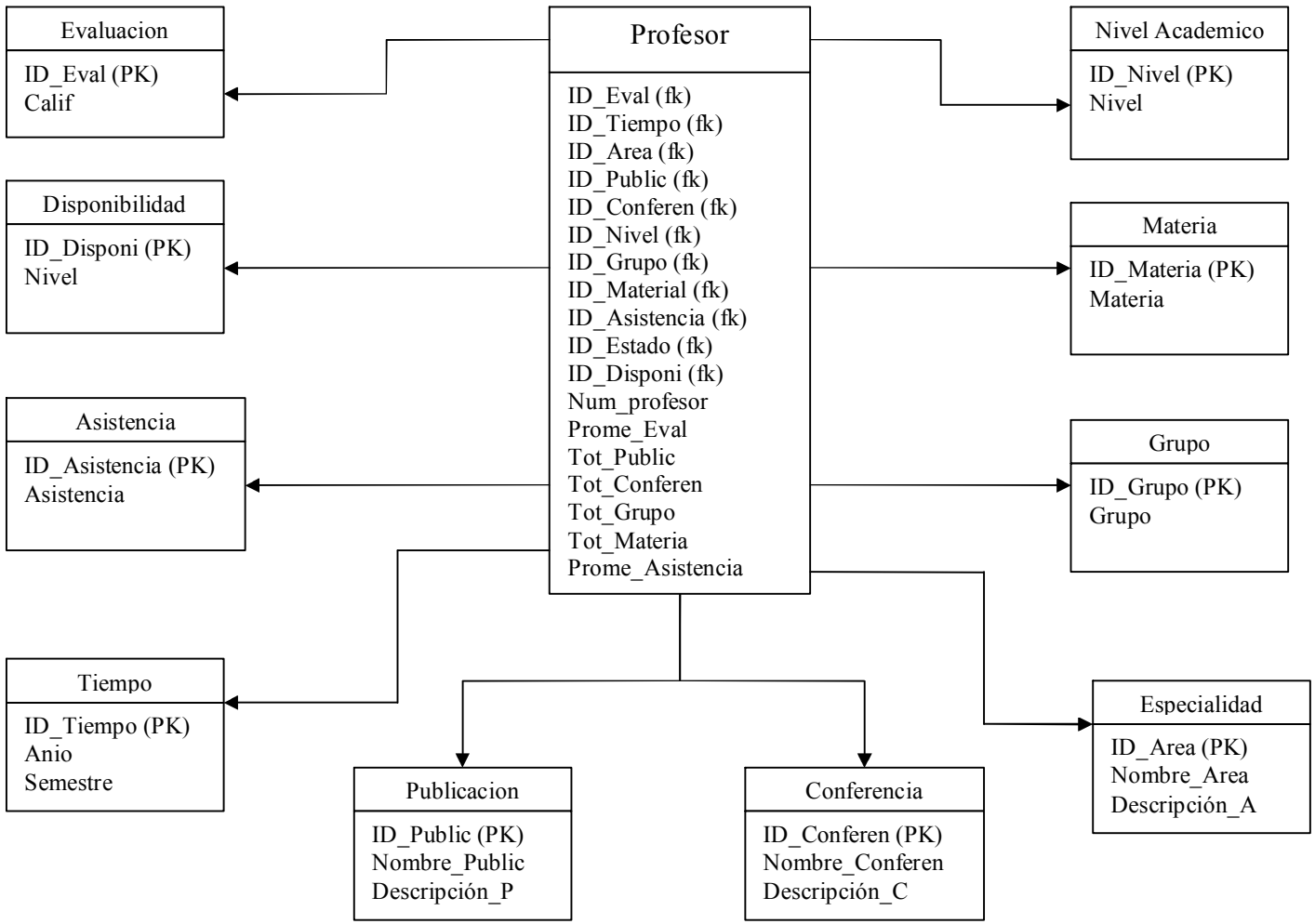


Diagrama físico

El diagrama físico como antes mencioné se basa en el diagrama lógico, estas tablas son las mismas que se enseñaron en el modelo dimensional, a cada una en el siguiente cuadro se muestra sus principales características: los nombres de los atributos, los tipos de datos, sus llaves ya sean PK o FK y si permiten los datos ser nulos o no.

Nombre tabla/columna	Tipo de dato	Permite nulos	Llave primaria/foranea	Observaciones
PROFESOR				
ID_Eval	Integer	N	Fk	Id Evaluacion
ID_Tiempo	Intrger	N	Fk	Id Tiempo
ID_Area	Intrger	N	Fk	Id Area
ID_Public	Intrger	N	Fk	Id Publicaciones
ID_Conferen	Intrger	N	Fk	Id Conferencias
ID_Nivel	Intrger	N	Fk	Id Nivel academico
ID_Grupo	Intrger	N	Fk	Id Grupo
ID_Materia	Intrger	N	Fk	Id Materia
ID_Asistencia	Intrger	N	Fk	Id Asistencia
ID_Estado	Intrger	N	Fk	Id Estado
ID_Disponi	Intrger	N	Fk	Id Disponibilidad
Num_Profesor	Intrger	N		Numero de profesor
Prome_Eval	Flota	n		Promedio de evaluacion
Tot_Public	Intrger	n		Total de publicaciones
Tot_Conferen	Intrger	n		Total de conferencias
Tot_grupo	Intrger	n		Total de grupos
Tot_Materia	Intrger	n		Total de materias
Prome_Asostencia	Float	n		Promedio de asistencia
EVALUACION				
ID_Eval	Intrger	n	PK	Id Evaluacion
Calif	Flota	n		Calificacion
DISPONIBILIDAD				
ID_Disponi	Intrger	n	PK	Id Disponibilidad
Nivel	Varchar	n		Nivel academico
ASISTENCIA				
ID_Asistencia	Intrger	n	PK	Id Asistencia
Asistencia	Flota			
TIEMPO				
ID_Tiempo	Intrger	n	PK	Id Tiempo
Anio	Char(8)	n		Año
Semestre	Char(2)	n		semestre
PUBLICACION				
ID_Public	Intrger	n	Pk	Id Publicacion
Num_Public	Intrger	n		Numero de publicacion
Nombre_Public	Varchar (50)	n		Nombre de publicacion
Descripción	Varchar (200)			descripcion de la publicacion

CONFERENCIA				
ID_Conferen	Intrger	n	PK	Id Conferencia
Num_Conferen	Intrger	n		Numero de conferencia
Nombre_Conferen	Varchar (50)	n		nombre de la conferencia
Descripción	Varchar (200)			Descripcion de la conferencia
ESPECIALIDAD				
ID_Area	Intrger	n	Pk	Id area de especialidad
Nombre_Area	Varchar (15)	n		nombre del area
Descripcion_A	Varchar (150)			Descripcion
GRUPO				
ID_Grupo	Intrger	n	PK	Id grupo
Num_grupo	Intrger	n		Numero de grupo
MATERIA				
ID_Materia	Intrger	n	Pk	ID materia
Num_materia	Intrger	n		Numero de materia
Nombre_Materia	Varchar (15)	n		Nombre de materia
NIVEL ACADEMICO				
ID_Nivel	Intrger	n	Pk	Id nivel academico
Nombre_Nivel	Varchar (15)	n		Nombre del nivel
ESTADO				
ID_Estado	Intrger	n	Pk	Id estado
Nombre_Estado	Varchar (15)	n		Nombre del estado

Definición de arquitectura

El siguiente diagrama muestra la arquitectura general que podemos abstraer para comprender la estructura que podría llegar a tener en caso de una futura implementación

En primera instancia tenemos las fuentes de datos, estas pueden ser bases de datos o archivos manuales, las bases de datos en esta arquitectura son Evaluaciones y Planta Docente mientras que tenemos los archivos manuales de las asistencias, las publicaciones (libros y artículos de revistas) y el Currículo Vite de los profesores. En el segundo apartado se encuentra el área de almacenamiento donde se lleva a cabo dos de los más importantes procesos para el desarrollo de un Data Mart que es la conexión de las bases de datos y el proceso de ETL (Extracción, Transformación y Carga). El tercer recuadro se ejemplifica lo que sería el área del servidor el cual contiene el Data Mart o en caso de mayor numero de datos el Data Warehouse. A continuación se encuentra el área de aplicación donde se realiza el cálculo de las métricas dando como resultado los cubos de información resumidos.

Hasta este punto se ha hablado del *back end*, que es la parte transparente para el usuario ya que el usuario final no puede tener los permisos de tablas como el dueño de la información, para el usuario final el *back end* es una caja negra. Para finalizar, el *front end* será todo aquello con lo que el usuario puede interactuar, por ejemplo todos aquellos servicios Web. El usuario puede realizar varias formas de extracción de las consultas y análisis de la información que necesita, como por ejemplo por herramientas que permitan el *Business Intelligence* o la minería de datos.

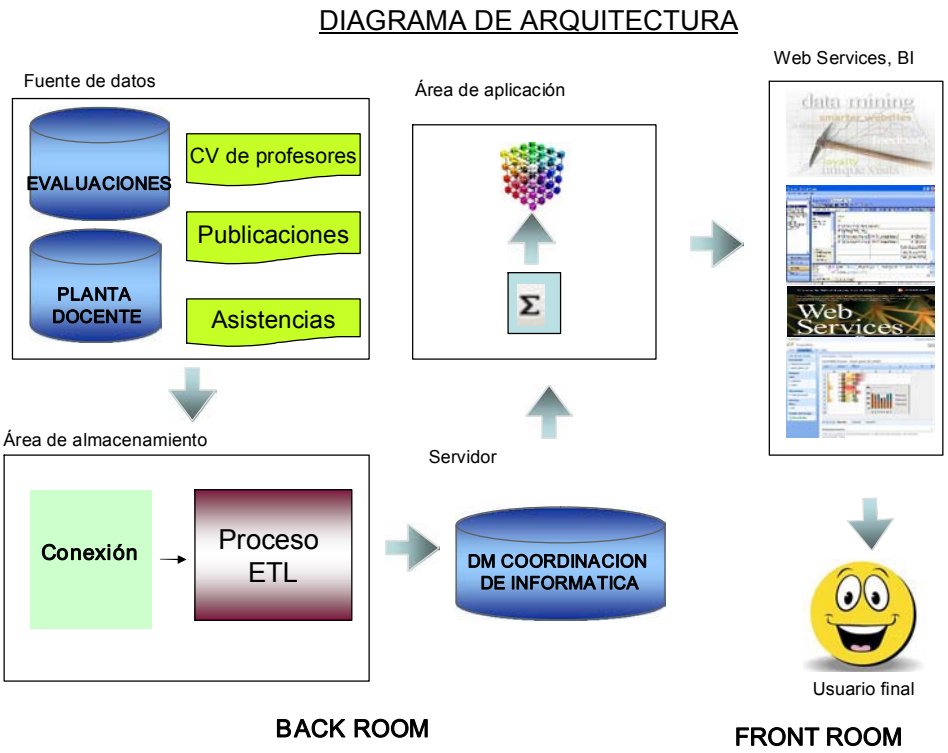
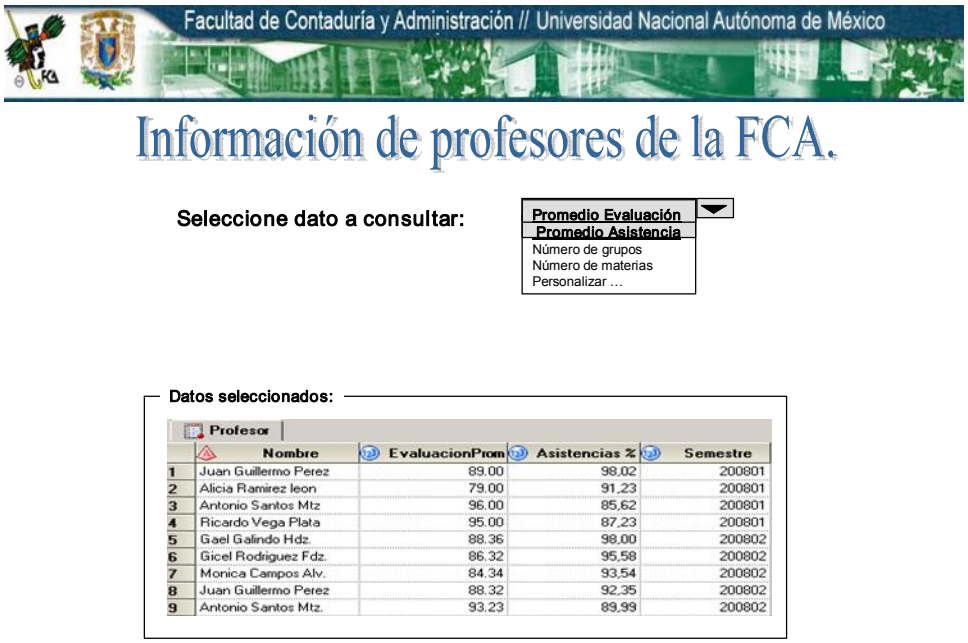


Ilustración 3 Diagrama de Arquitectura.

Desarrollo de la aplicación para el usuario final (propuesta para la FCA)

A continuación se muestran tres figuras, la primera indica el tipo de búsqueda de información que queremos obtener, la segunda muestra el resultado obtenido de la consulta y por ultimo la tercera nos muestra las estadísticas que se pueden obtener de esa consulta.
Estas consultas le proporcionarían al usuario información exacta para la toma de decisiones en este ejemplo se observa que la consulta se realizó por los semestres 2008-01 y 2008-02, para obtener el promedio de evaluaciones y el promedio de asistencias por semestre de los profesores, las estadísticas nos permiten ver el porcentaje del promedio de asistencias y evaluaciones del lado izquierdo y del lado derecho podemos ver la asistencia y evaluación por semestre.
Este podría ser la representación de un sistema de Web Service conectado a un Data Mart o Data Warehouse.

Ilustración 4Búsqueda de información

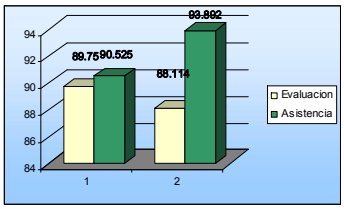


	Nombre	EvaluacionProm	Asistencias %	Semestre
1	Juan Guillermo Perez	89.00	98,02	200801
2	Alicia Ramirez leon	79.00	91,23	200801
3	Antonio Santos Mtz	96.00	85,62	200801
4	Ricardo Vega Plata	95.00	87,23	200801
5	Gael Galindo Hdz.	88.36	98,00	200802
6	Gicel Rodriguez Fdz.	86.32	95,58	200802
7	Monica Campos Alv.	84.34	93,54	200802
8	Juan Guillermo Perez	88.32	92,35	200802
9	Antonio Santos Mtz.	93.23	89,99	200802

Ilustración 5 Resultado de la búsqueda

Información de profesores de la FCA.

Promedio de asistencias y evaluación por semestre.



Información del profesor.

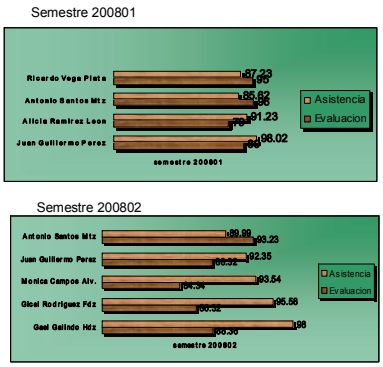


Ilustración 6 Estadísticas

CONCLUSIONES

En síntesis, en esta tesis vimos los principales conceptos de lo que es un Data Warehouse, un Data Mart así como sus principales características y componentes; se vieron las biografías de los padres del DW que son Kimball e Inmon, su filosofía sobre el DW y las metodologías, así como sus diferencias y similitudes entre éstas. También vimos algunas de las herramientas que pueden ayudarnos en la construcción de un DW, una breve guía de cómo se debe construir un DM y un DW y por ultimo se aplicó de lo aprendido a un caso dentro de la FCA con un problema que sugiere una solución de DM o DW.

Como objetivos cumplidos tenemos que en esta tesis se explicó las diferencias entre un DW y un DM, también los pasos de cómo construir un DM y un DW, Propuse un caso dentro de la Facultad de Contaduría y Administración oteniendo como resultado un diseño de DW básico como ejemplo de los pasos en la guía que menciono.
Determiné con base a una pequeña investigación y en el conocimiento que tengo algunas de las herramientas usadas en ese tema y comprendí que tanto hay herramientas de software libre como lo hay de software comercial, cada una con sus ventajas y desventajas pero eso lo dejo abierto para el que quiera profundizar más en el tema.
Se compararon las metodologías de ambos padres del DW: Kimball e Inmon, llegando a la conclusión de que ambas metodologías son buenas y que para adoptar una depende básicamente de las necesidades de la organización ya sea privada o pública de cualquier sector, inclusive puede llegar a suceder que se haga una mezcla de ambas. En lo personal a esta tesis le di el enfoque de la metodología de Inmon por ser a mi forma de verlo más simple para una solución de este tipo.
Comprendí la importancia de un DW en cualquier tipo de negocio y la utilidad que tiene para las organizaciones y lo grande que puede llegar a ser un proyecto de DW.
En cuanto al principal objetivo realicé una guía para la construcción de un DW que sirve también como apuntes para la capacitación del tema para en un futuro proponerla como materia optativa para los compañeros de informática de las generaciones futuras.

Contestando a las hipótesis sugeridas

Si Existen diferencias entre un Data Mart y un Data Warehouse como se ve en el capítulo 1 donde hay un cuadro comparativo que lo explica brevemente.
Si hay varias diferencias entre INMON Y KIMBALL, pero las dos metodologías pueden ser aplicadas dependiendo de las necesidades de la empresa u organización
Si existen varios software libre y comercial que nos apoya en las etapas del diseño e implementación como se vio en el capitulo 3.

REFERENCIAS

- Kimball, Ralph 2002 *The Data Warehouse toolkit: the complete guide to dimensional modeling* 2nd ed. R Elliott R Ipsen. United States of America.. Published by John Wiley & Sons Inc

- Inmon,William H. 2002. *Building the Data Warehouse* Thierd edition. R Ipsen. United States of America.. Published by John Wiley & Sons, Inc

- Mallach Efrem 2000 *Decision support and Data Warehouse systems* McGraw-Hill Cap 12 Data Warehouse and Executive Information System Fundamentals

- Ken Rudin, Christopher K. Buss, Ryan Sousa. 1999. *Data Warehouse Perforomance*. Wiley. United States of America. Published by John Wiley & Sons, Inc.

- Ralph Kimball, Laura Reeves, Margy Ross, Warren Thornthwaite 1998 *The Data Warehouse Lifecycle Toolkit.* United States of America. Published by John Wiley & Sons, Inc.

- Paul Westerman 2001. *Data Warehousing.* United States of America. Morgan Kaufmann publishers

- Jill Dyché. 2001 *E-Data*. Transformando datos en información con Data Warehousing. Buenos Aires, Agentina. Person Education.

- Christopher Adamson, Michel Venerable. 1998. *Data Warehouse Design Solutions*. United States of America. Published by John Wiley & Sons, Inc

- Claudia Imhoff, Nicholas Galemmo, Jonathan G. Geiger. 2003. *Mastering Data Warehouse Design Relational and Dimensional Techniques* United States of America. Published by John Wiley & Sons, Inc

- John Poole, Dan Chang, Douglas M. Tolbert, David Mellor.2002. *Common Warehouse Metamodel United States of America.* Published by John Wiley & Sons, Inc

- Robert Wrembel, Christian Koncilia 2007. *Data Warehouses and OLAP: Concepts, Architectures and Solutions.* Published in the United States of America by IRM Press

- A Min Tioa, Juan Trujillo. 2006. *Data Warehouse Knowledge* Discover. Berlin. Springer-Verlang

- Matthias Jarke, Maurizio Lenzenni, Yannis Vassiliou, Panos Vassiliadis. 2000. *Fundamentals of Data Warehouse*. Berlin. Springer-Verlang

Revistas y manuales

- Business Intelligence Journal Volume 8 Number 4 Fall 2003 TDWI. ESTADOS UNIDOS,Seattle

- Business Intelligence Journal Volume 9 Number 1 Winter 2004 TDWI. ESTADOS UNIDOS,Seattle

- Business Intelligence Journal Volume 11 Number 4 4th Quarter 2006 TDWI. ESTADOS UNIDOS,Seattle

- GeneXus GXquery Manual de Instalación Copyright Artech Consultores S.R.L. 1998-2007

- GeneXus GXplorer Manual de Instalación Copyright Artech Consultores S.R.L. 1998-2007

- Oracle® Data Mart Suite *The Oracle Data Mart Suite Cookbook* Release 2.6 August 1999 Part No. A75671-01

- Oracle® Warehouse *Builder User's Guide 10g* Release 2 (10.2.0.2) B28223-03 November 2006

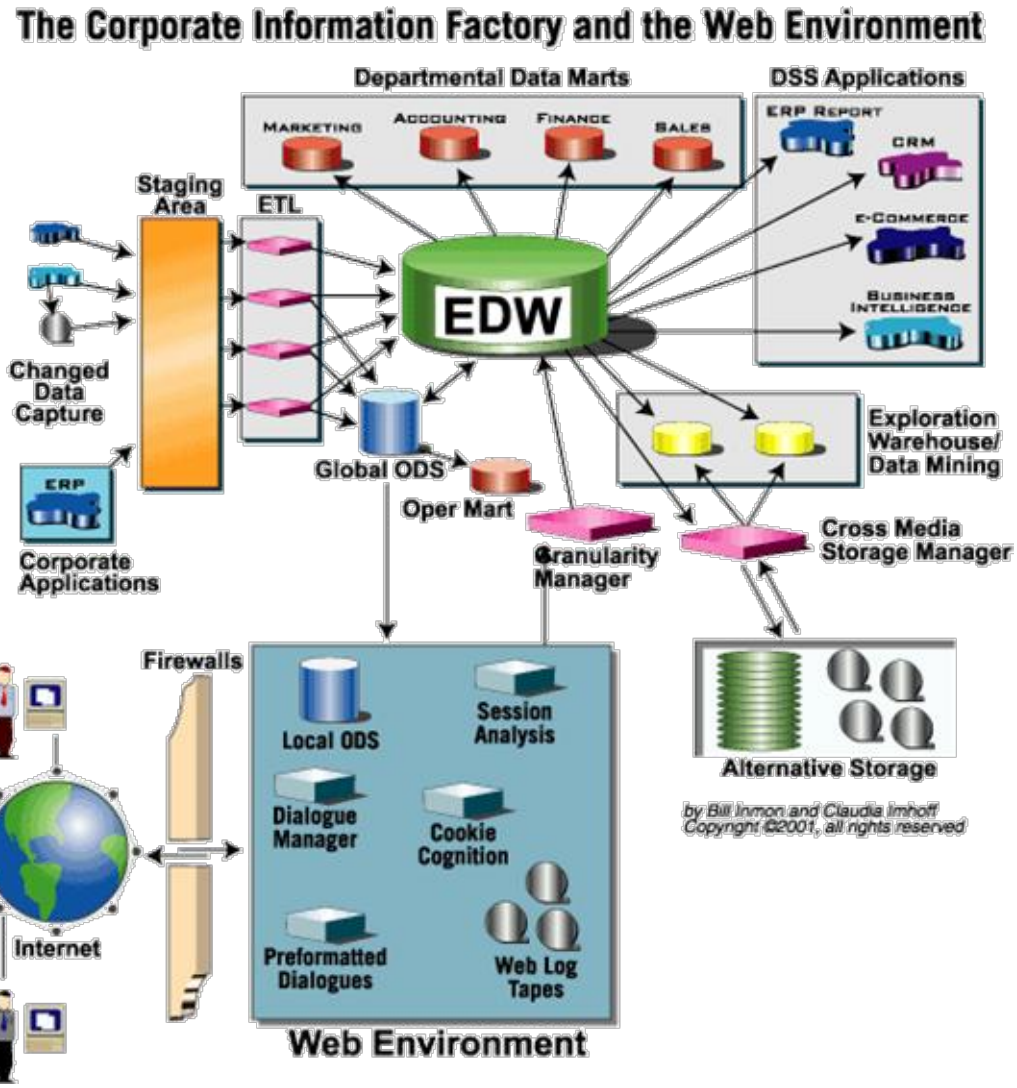
- SAS® *Data Integration Studio 3.4 User's Guide* SAS Institute Inc., Cary, NC, USA 2007

Fuentes de Internet.

- WEB:01 Corporate Information Factory About Bill, http://www.inmoncif.com/, visitada el 30 de agosto de 2007.
- WEB02: Kimball Group, http://www.kimballgroup.com/html/about.html, visitada el 17 de septiembre de 2007
- WEB:03 Facultad de Contaduría y Administración www.fca.unam.mx visitada el 26 de mayo de2008.
- WEB:04. Pentaho.http://pentaho.almacen-datos.com/ visitado el 30 de Septiembre de 2007.
- WEB:05 Todo BI.http://todobi.blogspot.com/2006/04/kettle-etl-para-pentaho.html vistado el 30 de Septiembre de 2007.
- Web:06 Proyecto GNU, en http://www.gnu.org/philosophy/free-sw.es.html. Visitada el 30/11/07.

APÉNDICES

Apéndice 1 The Corporate Information Factory



Brief Description

Operational Systems are the internal and external core systems that support the day-to-day business operations. They are accessed through application program interfaces (APIs) and are the source of data for the Data Warehouse and operational data store. (Encompasses all operational systems including ERP, relational and legacy.)

Data Acquisition is the set of processes that capture, integrate, trans-form, cleanse, reengineer and load source data into the Data Warehouse and operational data store. Data reengineering is the process of investigating, standardizing and providing clean consolidated data.

The Data Warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data used to support the strategic decision-making process for the enterprise. It is the central point of data integration for business intelligence and is the source of data for the Data Marts, delivering a common view of enterprise data.

Primary Storage Management consists of the processes that manage data within and across the Data Warehouse and operational data store. It includes processes for backup and recovery, partitioning, summarization, aggregation, and archival and retrieval of data to and from alternative storage.

Alternative Storage is the set of devices used to cost-effectively store Data Warehouse and exploration warehouse data that is needed but not frequently accessed. These devices are less expensive than disks and still provide adequate performance when the data is needed.

Data Delivery is the set of processes that enable end users and their supporting IS group to build and manage views of the Data Warehouse within their Data Marts. It involves a three-step process consisting of filtering, formatting and delivering data from the Data Warehouse to the Data Marts.

The Data Mart is customized and/or summarized data derived from the Data Warehouse and tailored to support the specific analytical requirements of a business unit or function. It utilizes a common enterprise view of strategic data and provides business units more flexibility, control and responsibility. The Data Mart may or may not be on the same server or location as the Data Warehouse.

The Operational Data Store (ODS) is a subject-oriented, integrated, current, volatile collection of data used to support the tactical decision-making process for the enterprise. It is the central point of data integration for business management, delivering a common view of enterprise data.

Meta Data Management is the process for managing information needed to promote data legibility, use and administration. Contents are described in terms of data about data, activity and knowledge.

The Exploration Warehouse is a DSS architectural structure whose purpose is to provide a safe haven for exploratory and ad hoc processing. An exploration warehouse utilizes data compression to provide fast response times with the ability to access the entire database.

The Data Mining Warehouse is an environment created so analysts may test their hypotheses, assertions and assumptions developed in the exploration warehouse. Specialized data mining tools containing intelligent agents are used to perform these tasks.

Activities are the events captured by the enterprise legacy and/or ERP systems as well as external transactions such as Internet interactions.

Statistical Applications are set up to perform complex, difficult statistical analyses such as exception, means, average and pattern analyses. The Data Warehouse is the source of data for these analyses. These applications analyze massive amounts of detailed data and require a reasonably performing environment.

Analytic Applications are pre-designed, ready-to-install, decision sup-port applications. They generally require some customization to fit the specific requirements of the enterprise. The source of data is the Data Warehouse. Examples of these applications are risk analysis, database marketing (CRM) analyses, vertical industry "Data Marts in a box," etc.

External Data is any data outside the normal data collected through an enterprise's internal applications. There can be any number of sources of external data such as demographic, credit, competitor and financial information. Generally, external data is purchased by the enterprise from a vendor of such information.

Apéndice 2 Entrevista

INFORMACIÓN SOLICITADA PARA LA ELABORACIÓN DE UN DATAMART

	PROFESORES	INVESTIGADORES
¿Qué tipo de información maneja?	▪ Datos personales ▪ Adscripción ▪ Consultoría ▪ Distinciones ▪ Niveles / grados académicos ▪ Experiencia laboral ▪ Idiomas ▪ Libros, reportes técnicos, reseñas, tesis dirigidas, etc. ▪ Docencia ▪ Artículos publicados ▪ Participación en congresos	▪ Datos personales ▪ Adscripción ▪ Consultoría ▪ Distinciones ▪ Niveles / grados académicos ▪ Experiencia laboral ▪ Idiomas ▪ Libros, reportes técnicos, reseñas, tesis dirigidas, etc. ▪ Docencia ▪ Estancias de investigación ▪ Apoyos CONACYT ▪ Desarrollos tecnológicos ▪ Grupos de investigación ▪ Proyectos de investigación ▪ Participación en congresos ▪ Artículos publicados
¿En qué manejadores de datos se encuentran las bases de datos?	PostgreSQL 8.1	PostgreSQL 8.1
¿Cuántas bases por manejador hay?	La idea es manejar una sola base integral. Actualmente solo es una base de datos para todos los sistemas. En el futuro queremos establecer bases de datos distribuidas.	La idea es manejar una sola base integral. Actualmente solo es una base de datos para todos los sistemas. En el futuro queremos establecer bases de datos distribuidas.
¿Quién las controla?	Los líderes de los proyectos, los arquitectos y los creadores de software. NO HAY DBA.	Los líderes de los proyectos, los arquitectos y los creadores de software. NO HAY DBA.
¿De qué departamento son?	El área de Proyectos Institucionales (PI) y el área de Sistemas (SI). NO HAY DBA.	El área de Proyectos Institucionales (PI) y el área de Sistemas (SI). NO HAY DBA.