CIBERTEC
con el respaldo de la UPC

Inteligencia de Negocios

Teoría

Índice

Presentación			5
Red de contenidos			6
Sesiones de aprendizaje			
SEMANA 1	:	Data Warehouse: Conceptos básicos. Data Warehousing: Conceptos básicos.	7
SEMANA 2	:	Indicadores de gestión - Conceptos Caso Práctico.	17
SEMANA 3	:	La necesidad de una arquitectura. La arquitectura de referencia de Zachman Evaluación Continua	23
SEMANA 4	:	La estrategia de Data Warehouse. Construcción de un Data Warehouse: Una metodología I.	37
SEMANA 5	:	Construcción de un Data Warehouse: Una metodología II. Análisis de los requerimientos empresariales.	45
SEMANA 6	:	Planificación de un proyecto Datawarehouse Identificación de requerimientos de negocio Evaluación Continua	55
SEMANA 7	:	Semana de Exámenes Parciales de Teoría	
SEMANA 8	:	Modelamiento de datos en el Data Warehouse Modelamiento dimensional: Conceptos.	61
SEMANA 9		Modelamiento dimensional: Casos prácticos. Modelamiento dimensional: Conceptos avanzados	69
SEMANA 10	:	Taller : Modelamiento Dimensional	79
SEMANA 11	:	Diseño de la base de datos de Data Warehouse Evaluación Continua	81
SEMANA 12	:	Poblando el Data Warehouse: Extracción, transformación y carga Poblando el Data Warehouse: Estandarización y limpieza de datos.	93
SEMANA 13	:	Poblando el Data Warehouse: Primera carga y procesos de actualización El acceso a los datos.	105
SEMANA 14	:	Disponibilidad de soluciones en el mercado. Consultas y reportes como herramientas de acceso a los datos. El proceso KDD Lenguaje de Consulta MDX I	117
SEMANA 15	:	Lenguaje de Consulta MDX II Minería de datos I. Evaluación Continua	131
SEMANA 16	:	Minería de datos II. Sesión de integración 2.	141
SEMANA 17		Examen final de Teoría.	

Presentación

En un mundo donde las Tecnologías de Información determinan la forma en que se hacen los negocios, las empresas necesitan explotar su mayor recurso: la información. Este análisis permitirá que se realicen análisis de tendencias y se obtengan parámetros que permita optimizar la toma de decisiones, tales como fusión de empresas, nuevas giros en el negocio, expansiones, etc.

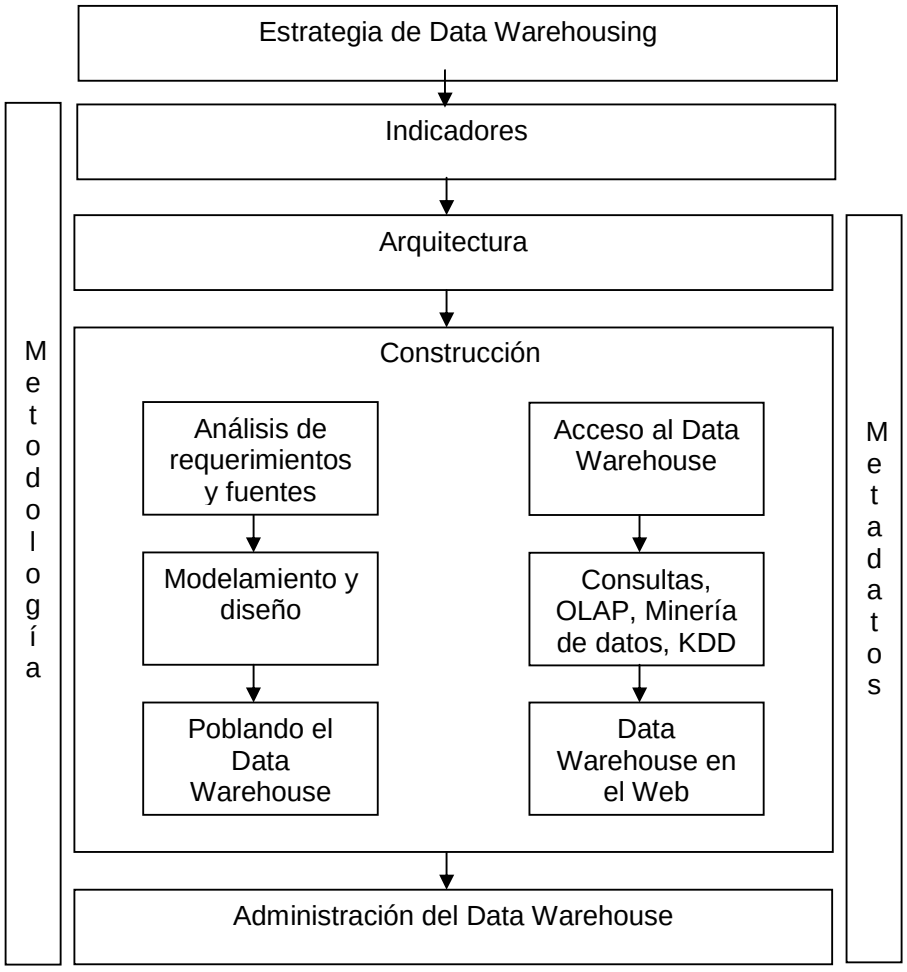
El presente manual tiene por objetivo brindar a los alumnos los conceptos básicos para el curso de Administración de Centro de Cómputo.

El manual esta desarrollado para complementar y afianzar lo expuesto en clase, con ejemplos y ejercicios prácticos se busca la comprensión de los temas.

El tema central del curso, Datawarehouse es abordado desde sus conceptos básicos, arquitectura, modelamiento dimensional, en la cual se incide de manera precisa y detallada, transformación de datos, MDX y minería de datos.

Red de contenidos

Data Warehouse



SEMANA 1

Data Warehouse - Data Warehousing: Conceptos básicos

OBJETIVOS ESPECÍFICOS

- Comprender los conceptos básicos de Data Warehouse.
- Comprender los conceptos básicos de Data Warehousing.

CONTENIDO

- Necesidad de un Data Warehouse
- Definiciones de Data Warehouse
- Componentes funcionales de un Data Warehouse
- Definir Data Warehousing
- Los componentes funcionales como proceso
- Infraestructuras

ACTIVIDADES

- Consolidar la definición de Data warehouse
- Entender los beneficios de un data warehouse

1. Necesidad de un Data Warehouse

Una de claves del éxito de las corporaciones modernas es el acceso a la información correcta, en el tiempo adecuado, en el lugar correcto y en la forma adecuada.

Es muy común escuchar a los ejecutivos decir las siguientes frases:

“Tenemos montañas de datos en esta compañía, pero no podemos acceder a ellos”

“Nada enloquece más a un gerente que tener dos personas que le presenten el mismo resultado de negocio, pero con diferentes cifras”.

“Sólo me interesa ver lo que es importante”.

“Todos sabemos qué datos no están bien”.

Estos problemas se presentan en la mayoría de las empresas, y pueden ser convertidos en oportunidades y transformados en requerimientos:

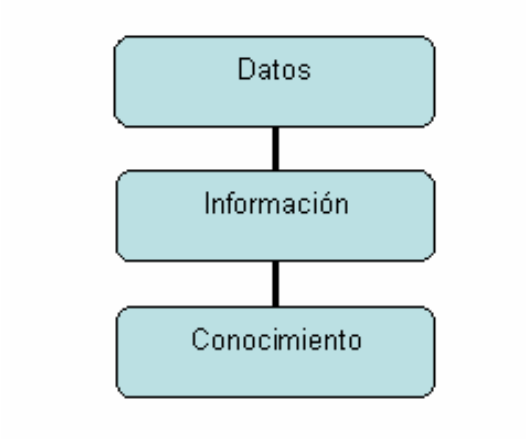
El Data Warehouse proporciona *acceso* a los datos corporativos u organizacionales.

Los datos en el Data Warehouse son *consistentes*.

El Data Warehouse no contiene solamente datos sino un conjunto de herramientas de *consulta, análisis y presentación* de la información.

La calidad de los datos en un Data Warehouse, *conducirá a una reingeniería* de las aplicaciones de negocio.

2. De los datos al conocimiento



Cuando los datos se ponen en un contexto, se convierten en información, y si luego esta información es sintetizada con la ayuda de la experiencia se llega al conocimiento.

3. Definiciones de Data Warehouse

Hay muchas definiciones de Data Warehouse en la literatura, de las cuales se presenta, las dos más representativas:

- William Inmon:

"***El Data Warehouse es una colección de datos, orientados a un tema, integrados, no volátiles, variantes en el tiempo, organizados para el apoyo a toma de decisiones***."

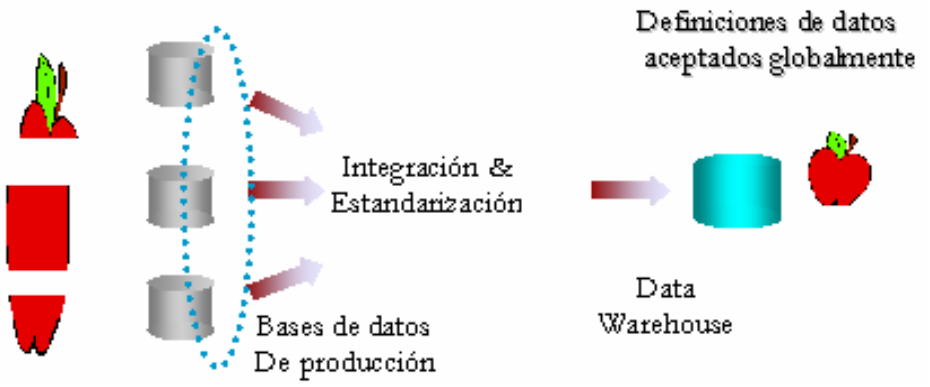
- Ralph Kimball:

"***Un Data Warehouse es una copia de los datos transaccionales, específicamente diseñada para realizar consultas y análisis*.**"

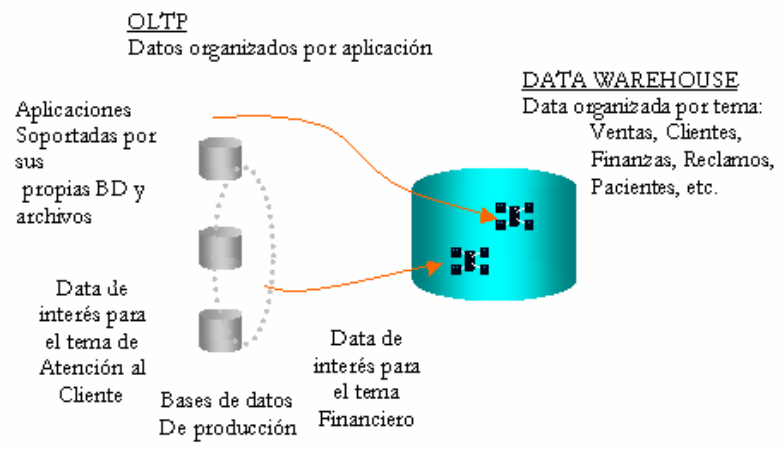
4. Análisis del concepto de Data Warehouse

- El Data Warehouse es una colección de datos que están almacenados en un lugar diferente a donde se almacenan los datos de las aplicaciones.

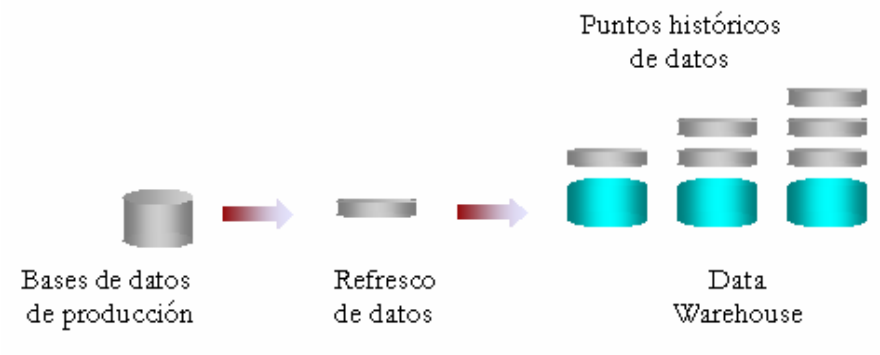
- Los datos en el Data Warehouse están *integrados*, lo que no sucede en los sistemas transaccionales debido a que estos solo almacenan información relevante al área usuaria y a la operatividad del sistema.



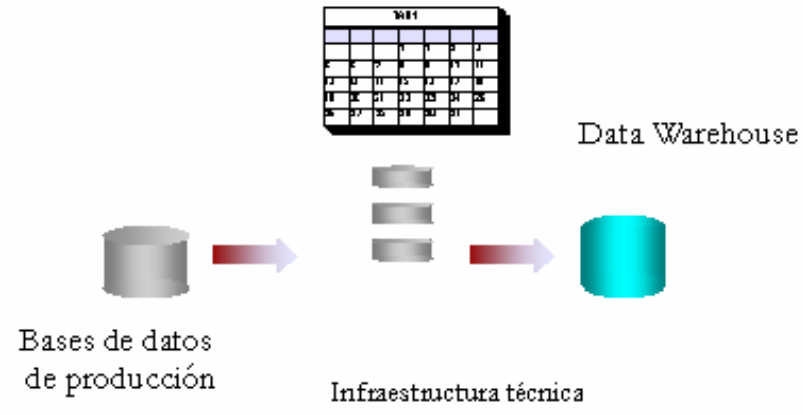
- Los datos en el Data Warehouse están orientados a un tema de negocio, se tienen modelos que representan las entidades del negocio.



- Los datos en el Data Warehouse son no volátiles, es decir que se guardan los datos históricos. Usualmente no se eliminan registros.



- Los datos en el Data Warehouse son variantes en el tiempo, es decir que se actualizan periódicamente. Se mantiene la historia.



5. Componentes funcionales de un Data Warehouse

Hay una serie de funciones que deben implementarse para el funcionamiento de un Data Warehouse

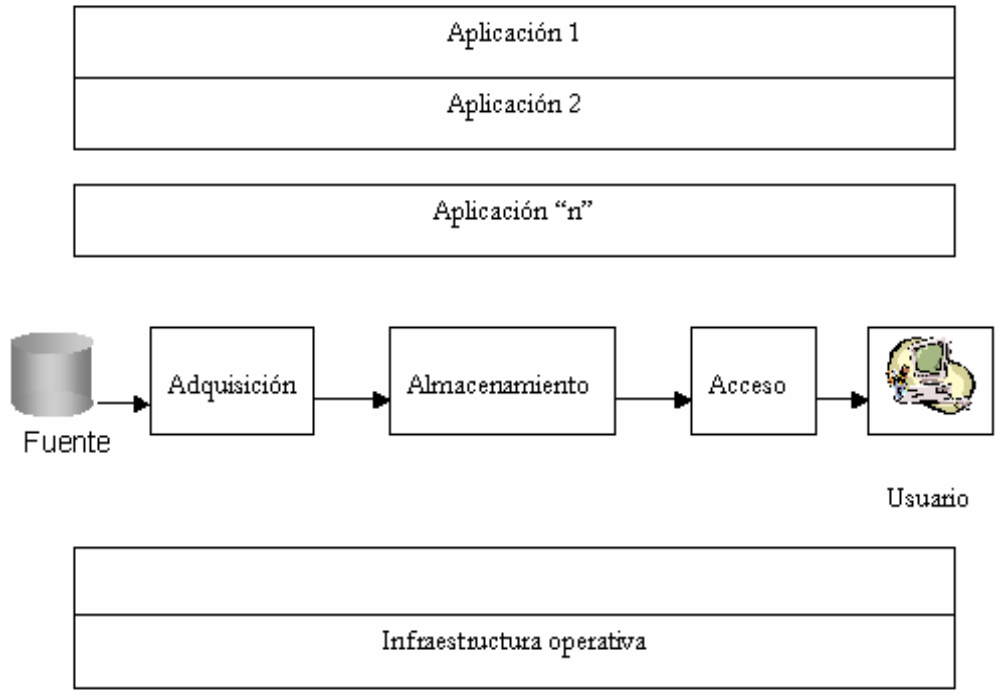
- Bloque de bases de datos operacionales, que capturan los datos y son la fuente de datos del Data Warehouse.
- Bloque de extracción, transporte, transformación, estandarización, limpieza y carga de los datos, que es el bloque responsable de poblar el Data Warehouse. Este bloque también es conocido como el "Staging area".
- Almacén o base de datos de Data Warehouse, es donde se almacena la información integrada, orientada al tema , histórica y actualizada.
- Bloque de explotación o acceso, que es donde se encuentran las aplicaciones que permiten el acceso, exploración y análisis de los datos.
- Metadatos, que es otra base de datos que contiene información acerca de los datos que hay en el Data Warehouse, acerca de los procesos y acerca del negocio.

6. Definir Data warehousing

Data Warehousing es el proceso de construir un Data Warehouse, el cual es un proceso continuo e incremental.
Transformar datos en conocimiento es un proceso complejo, en el que se transforman e integran los datos y puede ser sintetizado en las etapas representativas de un método industrial que se puede ilustrar con la siguiente frase:

"Ensamblar las materias primas(los datos de diferentes fuentes) según instrucciones específicas(modelo) para realizar un producto terminado(los datos listos para la consulta, exploración o análisis), guardado en un almacen de datos(el Data Warehouse) para que esté disponible a los clientes(usuarios finales)."

La figura siguiente ilustra el marco general de un Data Warehouse.



En este marco, se observan tres ámbitos: las aplicaciones, los componentes funcionales del Data Warehouse(adquisición, almacenamiento y acceso) y las infraestructuras(técnica y operativa).

Las aplicaciones analíticas.

Un Data Warehouse no se construye en una sola iteración. Cada tema tratado, se descompone en un conjunto de iniciativas (las aplicaciones).

Cada aplicación debe estar claramente definida(objetivos, actores, frecuencia y periodicidad del análisis).

Las aplicaciones deben ser controlables y proporcionar resultados “tangibles” en plazos menores a 6 meses, que corresponden al plazo medio de realización de una aplicación.

La descomposición en aplicaciones aporta numerosas ventajas, pero genera dificultades sobre ciertos temas, como los relacionados con la infraestructura técnica y organizativa que necesitan ser visualizados globalmente dentro de una arquitectura.

7. Los componentes funcionales como proceso

De los cinco componentes funcionales del Data Warehouse que se estudiaron en la primera sesión, se pueden visualizar tres desde el punto de vista de procesos: los componentes funcionales que son parte del proceso son la adquisición de datos, el almacenamiento y el acceso por parte de usuarios finales.

I. Adquisición

Consiste en recoger los datos útiles del sistema de producción. Se debe identificar los datos que sean necesarios para atender los requerimientos de información, luego planificar las extracciones con el fin de evitar saturación en la red, o afectar al sistema transaccional de producción.

Los procesos de extracción deben estar sincronizados con la finalidad de garantizar la integridad de la información. Los problemas que surgen al hacer esta sincronización puede ser muy complejos.

Después de extraer los datos del sistema transaccional, estos se deben “preparar” para adecuarlos a la forma del Data Warehouse. Esta “preparación” incluye la correspondencia de los formatos, la limpieza, la transformación y la agregación en muchos casos.

La carga es la última fase de la adquisición de datos, esta fase es particularmente importante sobre todo si se trata de volúmenes muy grandes.

II. Almacenamiento

El componente básico del soporte del almacenamiento es el DBMS (DataBase Manager System). El DBMS o motor de base de datos debe tener las características que le permitan responder eficientemente a las exigencias

de las consultas analíticas. Para lograrlo debe contar con diversos recursos como el paralelismo, la optimización del indexado con la finalidad de acelerar las consultas agregadas, ordenamientos y agrupaciones.

En relación con los tipos de datos, generalmente, se almacenan en formatos relacionales; sin embargo, frente a la gran cantidad de datos en forma de documentos, imágenes, audio y video, los DBMS están evolucionando en el sentido de permitir la gestión de estos tipos de datos. Esta evolución se ve reforzada aún más con la llegada de Internet.

III. Acceso

El acceso al Data Warehouse se da mediante herramientas o aplicaciones de tipo Cliente/servidor o herramientas que pueden utilizarse desde el Web. Hay una gran variedad de herramientas en el mercado y el número de aplicaciones de acceso que se pueden desarrollar es también muy grande.
Sea cual sea el tipo de herramienta, tendrá que adaptarse a las exigencias del usuario y su manera de trabajar. En el mundo de la decisión, el análisis es también un proceso iterativo y los resultados de la consulta actual influyen a menudo en la consulta siguiente. Esto se puede resumir en la siguiente frase: " Dame lo que te pido y luego podré decirte lo que realmente quiero ".

8. Infraestructuras

Para hacer frente a las necesidades de Data Warehouse, el papel de la informática es definir e integrar una arquitectura sobre la que implementará el Data Warehouse.

Se debe considerar dos niveles de infraestructura en un Data Warehouse: la infraestructura técnica o conjunto de componentes materiales y programas, y la infraestructura operativa o conjunto de procedimientos y servicios para administrar los datos, gestionar los usuarios y utilizar el sistema.

Por un lado, la infraestructura técnica se compone de productos que implementan las tecnologías elegidas, integrados en un conjunto coherente y homogéneo. Por otro lado la infraestructura operativa se compone de todos los procesos que permiten, a partir de los datos de producción, crear y gestionar el Data Warehouse.

Autoevaluación

1. Enumere al menos dos de los problemas que mencionan los ejecutivos y que se pueden resolver con el Data Warehouse.

2. Describa el proceso que se debe seguir para llegar de los datos al conocimiento.

3. En sus propias palabras, defina Data Warehouse.

4. ¿La base de datos del Data Warehouse está separada de las bases de datos transaccionales?

5. ¿Porqué se dice que los datos están integrados en un Data Warehouse?

6. ¿Por qué se dice que los datos están orientados a un tema de negocio en Data Warehouse?

7. ¿Por qué se dice que los datos son no-volátiles en Data Warehouse?

8. ¿Por qué se dice que los datos son variantes en el tiempo en Data Warehouse?

9. Enumere los componentes funcionales del Data Warehouse.

10. ¿Los problemas de calidad de datos pueden ser solucionados con un data Warehouse?

11. Defina Data Warehousing en sus propias palabras.

12. ¿Por qué se utiliza la analogía con un proceso industrial al definir Data Warehousing?

13. ¿Cuáles deben ser las características de una “aplicación” en el contexto de Data Warehousing?

14. Enumere los sub-procesos que se llevan a cabo dentro del proceso de adquisición.

15. ¿Cuál es el componente básico del proceso de almacenamiento?

16. Mencione las características de un DBMS adecuado para Data Warehouse.

17. ¿Cuáles son las infraestructuras básicas para la implementación de un Data Warehouse?.

Para recordar

1. El Data Warehouse nace con la finalidad de proporcionar el acceso a los datos, de la mejor manera, en el tiempo justo y de la forma más adecuada.

2. El Data Warehouse es un conjunto de datos separado de los datos transaccionales.

3. El Data Warehouse es un conjunto de datos integrados, orientados a un tema, no-volátiles y variantes en el tiempo.
4. El proceso de construcción de un Data Warehouse se denomina Data Warehousing.

5. Data Warehousing es un proceso continuo e incremental.

6. Se distinguen tres grandes procesos: adquisición, almacenamiento y acceso.

7. Hay dos niveles de infraestructura que soporte un Data Warehouse: la infraestructura técnica y la infraestructura operativa.

SEMANA

2

Indicadores de Gestión
Conceptos – Caso Práctico

OBJETIVOS ESPECÍFICOS

- Reconocer los indicadores
- Definir indicadores

CONTENIDO

- Definición de Indicador
- Importancia de un indicador
- Tipos de Indicadores

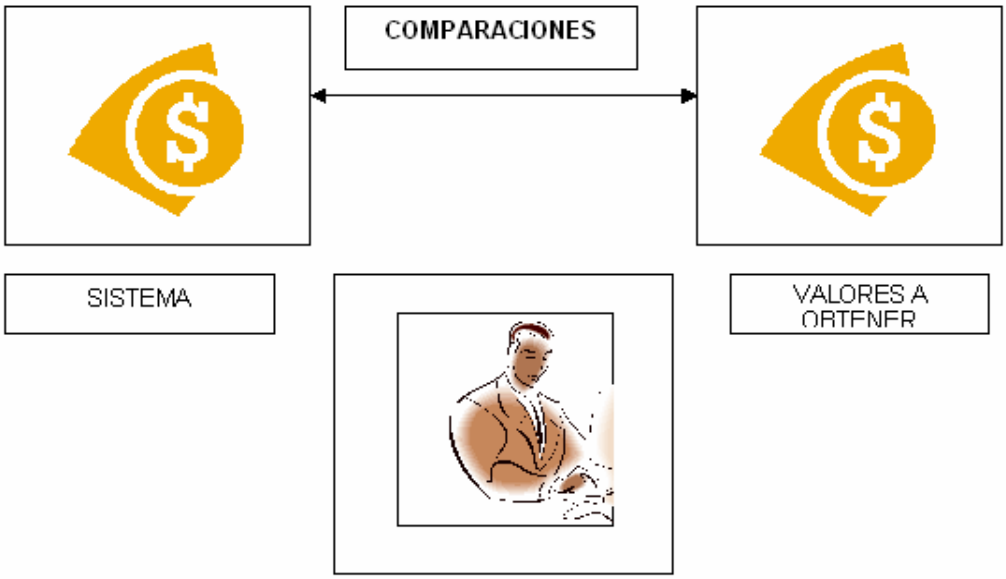
ACTIVIDADES

- Definir un conjunto de indicadores para un caso propuesto

1. Definición de indicador

Un sistema es definido como un conjunto de componentes que trabajan en conjunto los cuales tienen un objetivo específico. La importancia de la tarea de control radica en evaluar periódicamente si dicho sistema esta cumpliendo con lo esperado. Debemos recordar que aquello que no se puede medir, no se puede controlar y para poder medir debemos determinar ciertos valores de referencia. Dichos valores de referencia representan a los indicadores, es decir, la comparación entre el valor obtenido por el sistema vs. el valor del indicador nos revela el estado actual del sistema. Un indicador debemos entenderlo como la evaluación de un signo vital de una organización.

Ejemplo:
Número de latidos del corazón de un paciente, compararlo con los índices normales establecidos.



Se debe hacer comparaciones entre los valores esperados contra los valores producidos por el mismo sistema. Dicha métrica nos dará la información referente al éxito o fracaso del sistema.

Presentaremos algunas definiciones:

> "Normalmente, un indicador es una variable dimensional unitaria, expresada como un cuociente, que correlaciona dos variables cualesquiera. A través de este concepto es posible relacionar diversos variables presentes en los procesos de una empresa. Ej. (N° Ciclos/Hora), (Lts Comb./máquina), (Productos/día), etc. Un indicador es un indicador de gestión, cuando la correlación de estas dos variables permite conocer el funcionamiento de los procesos y recursos de mi empresa. Para ello el indicador puede tener distintos comportamientos, contenidos entre un valor mínimo y máximo. De este modo, un conjunto de indicadores seleccionados me permite conocer el comportamiento global de la empresa y controlar el normal funcionamiento de ella"
>
> FUENTE: TodoPymes
> Url : http://www.todopymes.cl/topicos_avanzados/gestion_avanzado.html#2

"Indicador de Gestión es una referencia que permite determinar en que medida la ejecución del plan lo acerca o lo aleja de los objetivos trazados en él. Como lo señala Serna (1994), los índices de gestión son unidades de medida gerencial que permiten evaluar el desempeño de una organización en relación a sus metas, objetivos y las res-ponsabilidades con los grupos de referencia."

- Expresar un resultado (de gestión)
- Ser Simple
- Ser Significativo
- Ser Coherente

Ser Relativo a un responsable

FUENTE: GepSea
Url : http://personales.com/venezuela/merida/gepsea/objetivos.htm

2. Importancia de un indicador

La importancia de un indicador radica en la particularidad de informar al usuario el estado actual del sistema.

Si se desea saber el estado actual de la presión en el ser humano, entonces se debería tomar dicha presión con los instrumentos debidos y comparar dichos valores obtenidos contra los valores esperados.

Un indicador además de revelar el estado del sistema nos permitirá tomar decisiones preventivas o correctivas de acuerdo a los resultados de las comparaciones entre el valor esperado y el valor obtenido del sistema.

Ejemplo :

Medir :

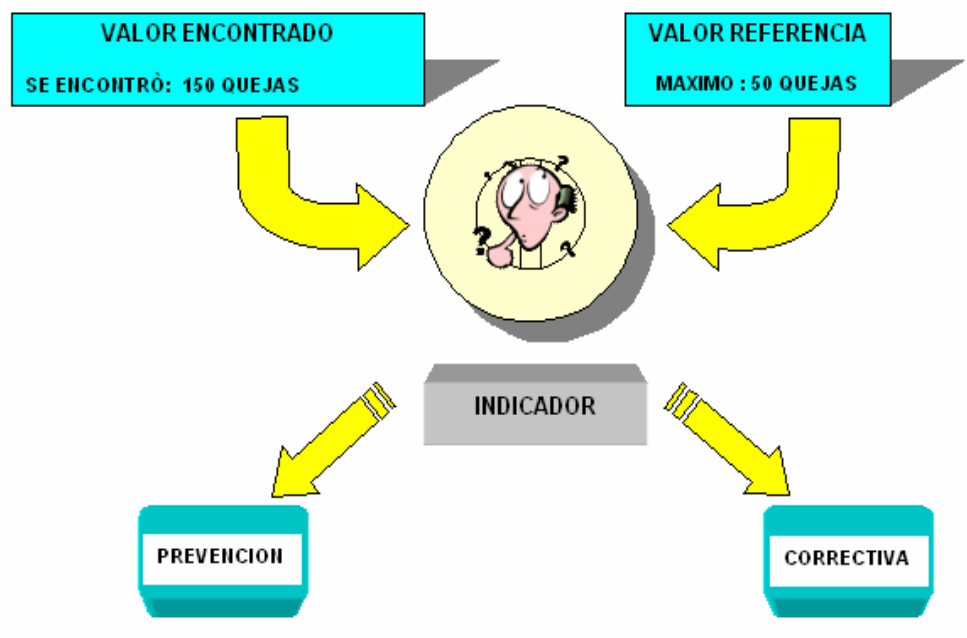
- "Grado de satisfacción del servicio al cliente"

Indicadores :

- "Número de quejas de los clientes"
- "Número de sugerencias con quejas de un determinado tipo"

Acciones :

- Si el número de quejas es muy alto, se deberían tomar acciones correctivas sobre el producto, o servicio.



3. Tipos de indicadores

Podemos clasificar a los indicadores en los siguientes tipos:

- **Indicadores de Cumplimiento**.- Indica el ratio de obtención de una tarea en particular. (Ejemplo : Cumplimiento de elaboración de reportes).
- **Indicadores de Evaluación**.- Indica el rendimiento en el desarrollo de una tarea.
- **Indicadores de Eficiencia**.- Indica el ratio relacionado con el tiempo invertido en el desarrollo de una tarea.
- **Indicadores de Eficacia.**- Indica la capacidad en el desarrollo de una tarea, es decir, el haberlo realizado de manera óptima.
- **Indicadores de Gestión.**- Indica la manera en que el proceso se está realizando. Mide la capacidad de administración con respecto a un proceso. Es vital para entender el día a día de la empresa.

Adaptado: Indicadores de gestión
Url : http://web.jet.es/amozarrain/gestion_indicadores.htm

4. Ejemplo de Indicadores

El siguiente ejemplo presenta 10 indicadores de una entidad educativa, en el área académica.

1. Promedio de Notas por: Alumno, Aula, Ciclo/Año, Global Ciclos/años, Carrera, y nota promedio global de la entidad
2. Promedio de Test de satisfacción a delegados de aula
3. Número de veces, en que se actualiza el Plan de estudios, en un año
4. Número de Matriculados
5. Número de traslados

6. Número de alumnos asistentes
7. Número de alumnos por profesor
8. % Desaprobados
9. % Repitentes
10. Número de Rematriculados - RETORNO

Extraído : 60 Indicadores de Gestión para Entidades
Autor : Ricardo Cuya Vera
Url: http://web.jet.es/amozarrain/gestion_indicadores.htm

5. Caso Práctico

Deberá identificar los indicadores en el caso que se propondrá en clase.

Caso : _____

Indicadores :

Autoevaluación

1. Proponga un caso e identifique los indicadores apropiadamente
2. ¿Por qué debemos encontrar indicadores?
3. ¿Por qué debemos medir?
4. ¿Qué acciones podemos tomar a partir de los resultados mostrados por los indicadores?

Para recordar

1. Un indicador representa la “marcha” del proceso con respecto a los resultados esperados.

2. Todo sistema debe tener un grupo de indicadores para controlar si los objetivos están siendo cumplidos.

SEMANA

3

La necesidad de una arquitectura – Arquitectura de Referencia de Zachman

OBJETIVOS ESPECÍFICOS

- Describir las razones de la necesidad de una arquitectura
- Identificar los componentes básicos de la arquitectura de Zachman

CONTENIDO

- La historia con visión de futuro
- La necesidad de una arquitectura
- La arquitectura de los sistemas de información y Datawarehouse
- Introducción a la arquitectura de Zachman
- Beneficios de la arquitectura de referencia
- Los bloques de construcción de la arquitectura de referencia

ACTIVIDADES

- Comprender la arquitectura de Zachman en el proceso Datawarehouse

1. La historia con visión de futuro

Las arquitecturas de los ambientes computacionales empresariales usualmente no proveen el nivel de acceso a los datos que las compañías modernas requieren. Dentro de la tercera generación de sistemas de información, las compañías han tenido éxito transformando datos en información, pero llegar al conocimiento aun parece complicado.

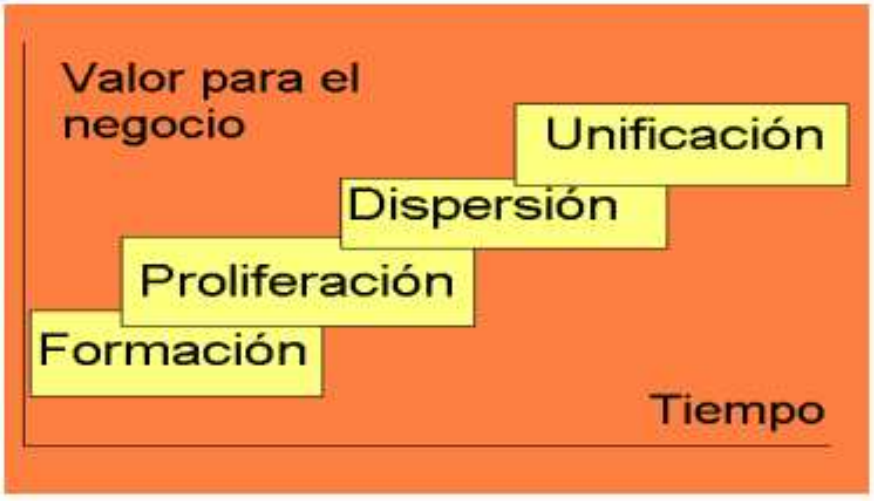
Durante la primera generación de los ambientes computacionales (1950 hasta 1970), el computador fue introducido y utilizado principalmente para mejorar la eficiencia de determinadas tareas.

En la segunda generación (1960 a1980), las aplicaciones de las computadoras en los negocios proliferan y los usuarios pueden interactuar con la computadora mediante terminales para mejorar la eficiencia y la efectividad.

En la tercera generación (1980 a 1990), los componentes computacionales se ven dispersos a lo largo y ancho de la compañía.

En la cuarta generación, las fuentes de información de la compañía se unifican desde la perspectiva del negocio permanecen dispersas desde la perspectiva tecnológica y física.

La evolución de las tecnologías de la información



(*)Bibliografía: Data Stores Data Warehousing and the Zachman Framework. W.H. Inmon, John A. Zachman, Jonathan G.Geiger

Evolución	Formación	Proliferación	Dispersión	Unificación
Tecnología	Compleja, componentes caros.	Compleja, componentes caros, terminales remotos.	Compleja, componentes baratos, servidores distribuidos.	Compleja, componentes baratos, servidores distribuidos.
Administración de datos	Tarjetas, cintas magnéticas. Forma: secuencial	Cintas y discos magnéticos. Forma: Jerárquica, relacional.	Cintas y discos magnéticos, discos ópticos. Forma: Jerárquica, relacional, O.O.	Cintas y discos magnéticos, discos ópticos. Forma: Relacional, O.O., Multidimensional.
Lenguajes de Programación	Máquina, assembler.	Procedural, assembler.	Procedural, gráfico.	Gráfico, intuitivo.
Metodología	Cascada.	Ingeniería de la información(CASE).	Ingeniería de la información(CASE), O.O.	Aproximación a las arquitecturales, O.O., Repositorios
Aplicaciones	Complejas, inflexibles, integración de datos y procesos.	Complejas, baratas, interfaces complejas.	Complejas, flexibles, interfaces complejas, O.O., Soporte a decisiones.	Muy flexibles, componentes reutilizables, Datos para DSS y Datos y objetos para OLTP.

2. La necesidad de una arquitectura.

Antes de entender la necesidad de una arquitectura, se debe entender qué es una arquitectura. La descripción y definición de arquitectura relacionada a los sistemas de información no es tan sencilla como para otras disciplinas. Aún así, es real e importante en el mundo de los datos, la información y los procesos.

Una manera de entender la arquitectura, es entendiendo qué hace una arquitectura.

Consideremos un constructor de pistas que no entiende una arquitectura. Un día pavimentará y asfaltará una pista. Pronto los carros empezarán a transitar por ella. Hasta que un día el tránsito se ve interrumpido porque la pista ha tenido que romperse con la finalidad de colocar una red de alta tensión por debajo de ella. Una vez colocada la línea de alta tensión, el tráfico vuelve a la normalidad pero sólo por unos días, porque es interrumpido nuevamente se ha tenido que romper la pista otra vez debido a que se necesita colocar una tubería de agua que atraviesa la misma. Así transcurre el tiempo y, después de muchas roturas, la pista queda desnivelada y con muchos parches.

Si el constructor de pistas, hubiera tenido una arquitectura, ésta le hubiera servido para saber cuál es el orden apropiado para hacer las cosas.

Desde una segunda perspectiva, una arquitectura define un patrón reconocido universalmente. Por ejemplo, una columna griega es reconocida en Grecia, en Inglaterra y en Francia, así haya sido una columna construida hace mil años por los griegos o sea una moderna columna de un nuevo edificio.

En conclusión, una arquitectura es necesaria porque nos dice el orden en el que se deben hacer las cosas y porque da a todos una visión de lo que será un sistema de información, gracias a que muestra un patrón reconocido universalmente.

9. La arquitectura de los sistemas de información

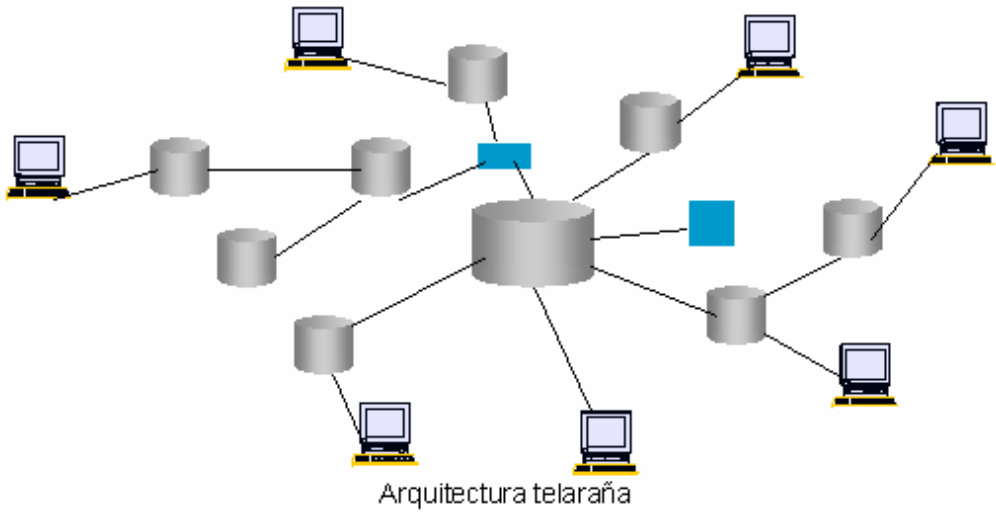
Los sistemas de una compañía evolucionan en el marco de una arquitectura.

En los primeros años los sistemas de información, se construyeron con la finalidad de automatizar las labores repetitivas, un ejemplo típico son los sistemas contables. Estos sistemas fueron construidos en una tecnología secuencial.

Con el advenimiento del procesamiento transaccional “on-line” la tecnología de la información se posiciona en el corazón del negocio. Por ejemplo los sistemas de reservas en las líneas aéreas, sistemas de caja en bancos, etc. Posteriormente, el éxito del procesamiento “on-line” hace que se multipliquen las aplicaciones y las bases de datos.

Surge, luego, la necesidad de contar con información estratégica y aparecen los sistemas de soporte a decisiones con los que los analistas se echaban a buscar la información en los sistemas transaccionales. Todo esto origina lo que se podría llamar una telaraña de sistemas y bases de datos. Esta telaraña es básicamente inestable y tiene algunas deficiencias como las siguientes:

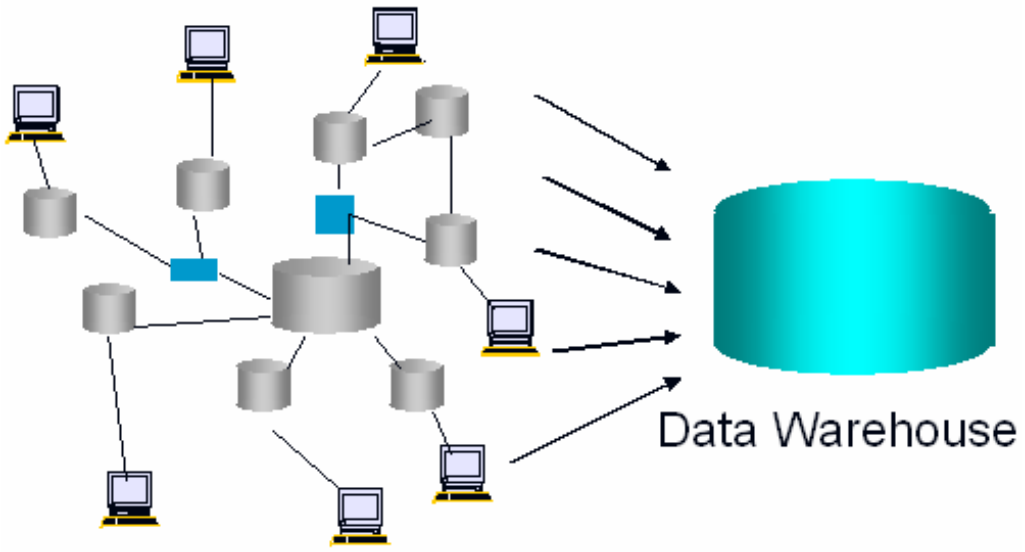
- Presenta dificultad para obtener resultados
- Hay un problema de consistencia en los resultados obtenidos
- La data no está integrada en esta red
- No hay información histórica
- Es complicado saber donde están los datos que se necesitan analizar



Esta realidad hace que sea necesario construir una estructura llamada Data Warehouse.

10. La arquitectura de un Data Warehouse

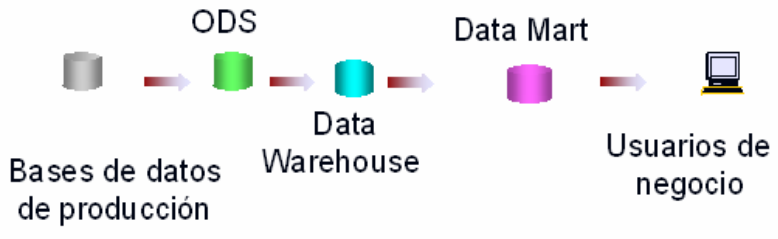
La arquitectura telaraña va a originar el Data Warehouse como un repositorio independiente con la finalidad de que las consultas no afecten el desempeño de los sistemas operacionales.



En un data warehouse, se tendrá información agregada, información archivada e información granular o detallada.

Uno de los problemas al tener niveles de agregación es que se complica la integración. Es por ello que surge el ODS (Operational Data Store), que contiene el mismo nivel de detalle que los sistemas operacionales y tiene la información integrada. Un ODS se utiliza para la toma de decisiones a nivel operacional.

Por otro lado, surge la necesidad de proporcionar información a determinados grupos de usuarios, para ello surgen los denominados DataMarts. Dentro de la arquitectura de un Data Warehouse, un DataMart debe alimentarse de un Data Warehouse. De lo contrario, corre el riesgo de ser un componente más en la arquitectura telaraña.



11. Introducción a la arquitectura de Zachman

La construcción de un Data Warehouse involucra tres tipos de técnicas. En primer lugar, las técnicas empresariales relacionadas con la comprensión del significado de los datos que contiene un Data Warehouse. En segundo lugar las técnicas relacionadas con la tecnología debido a la necesidad de interactuar con muchas tecnologías, distribuidores y usuarios finales. Finalmente las técnicas administrativas, que deben permitir administrar la diversidad de procesos, usuarios, temas de negocio, y tecnologías.

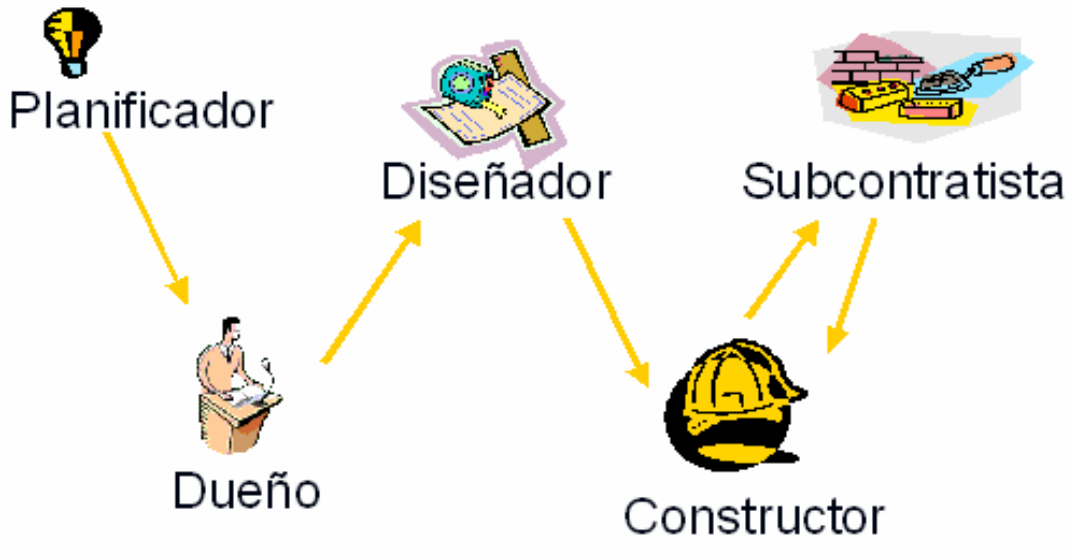
Para compartir una visión desde distintos puntos de vista, es necesario tener un diagrama. El mismo permite ver algo complicado y hacerlo inteligible mediante el uso de analogías que simplifican y ayudan a separar una solución compleja en componentes pequeños.

Los requerimientos de un Data Warehouse son tan variados y diversos como sus usuarios. Estos requerimientos se pueden analizar desde la perspectiva de cada usuario.

12. Las perspectivas de la arquitectura de Zachman

El diagrama de Zachman es una de las formas más eficaces de visualizar un sistema desde muchas perspectivas. En una compañía, las personas tienen diferentes roles y, por lo tanto, tienen diferentes perspectivas dependiendo de sus necesidades y usos de la información

Hay 5 roles básicos en la creación de un producto:



- Planificador: define parámetros básicos, especifica el alcance.
- El dueño (inversionista): proporciona información sobre el producto y su uso.
- Diseñador: especifica el producto, de manera que se cubran las expectativas del dueño.
- Constructor: administra el proceso de construcción y ensamblaje.
- Sub-contratista: construye cada componente especificado por el constructor.

Las perspectivas se caracterizan por:

Perspectiva	Propósito	Producto	Restricción
Planificador	Alcance	Definición alcance	Financiero y regulatorio
Dueño	Producto real	Modelo de negocio	Político y uso
Diseñador	Producto abstracto	Modelo de sistema	Arquitectura y físico
Constructor	Construye y ensambla	Modelo tecnológico	Construc. y equipamiento
Subcontratista	Construye	Modelo de componentes	Implement, integración

13. Las dimensiones de la arquitectura de Zachman.

Las dimensiones de la arquitectura de Zachman son una forma abstracta de entender las necesidades de cada perspectiva.

Se busca dar respuesta a las siguientes preguntas:

Las dimensiones se caracterizan por:

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

14. Beneficios de la arquitectura de referencia

La arquitectura de referencia facilita las siguientes tareas:

- Evaluación de las inversiones actuales
- Análisis de los costos y beneficios
- Análisis y administración de riesgos
- Evaluación de distribuidores
- Evaluación de productos y herramientas

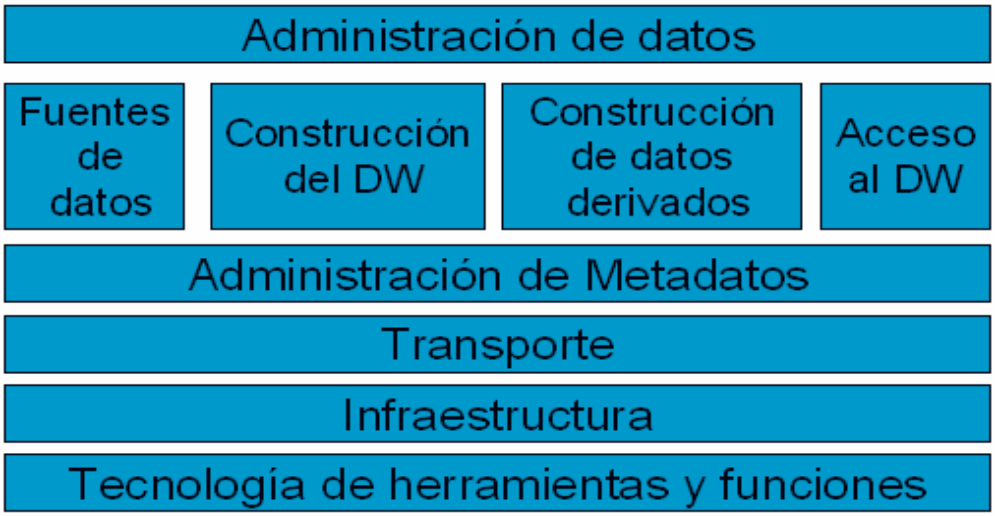
- Mantenimiento y mejoramiento
- Planeación y administración de proyectos
- Evaluar la tecnología
- Simulación de proyectos
- Arquitectura y diseño

15. Los bloques de construcción de la arquitectura de referencia

La arquitectura de referencia divide el Data Warehouse en bloques de construcción y capas.

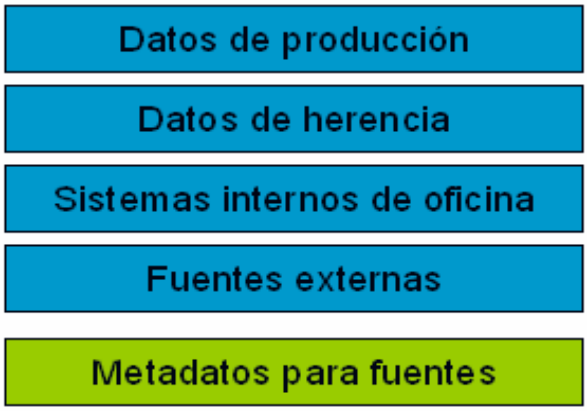
- Bloques: se relacionan con la funcionalidad específica del Data Warehouse.
- Capas: representan el ambiente necesario para la implementación de los bloques.

Una visión de alto nivel de la arquitectura de referencia sería la siguiente:

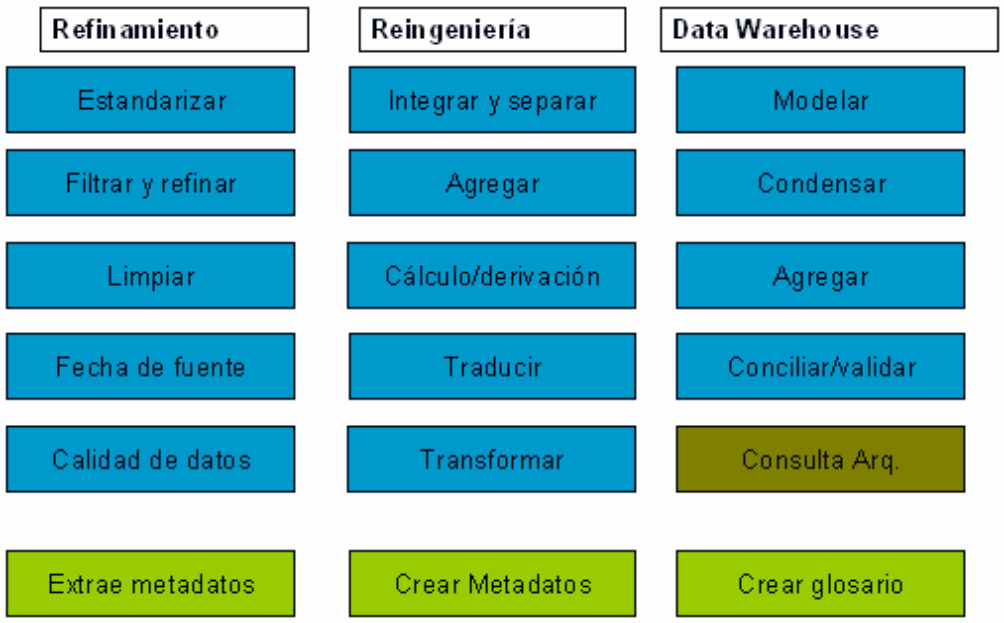


Los bloques del diagrama de Zachman:

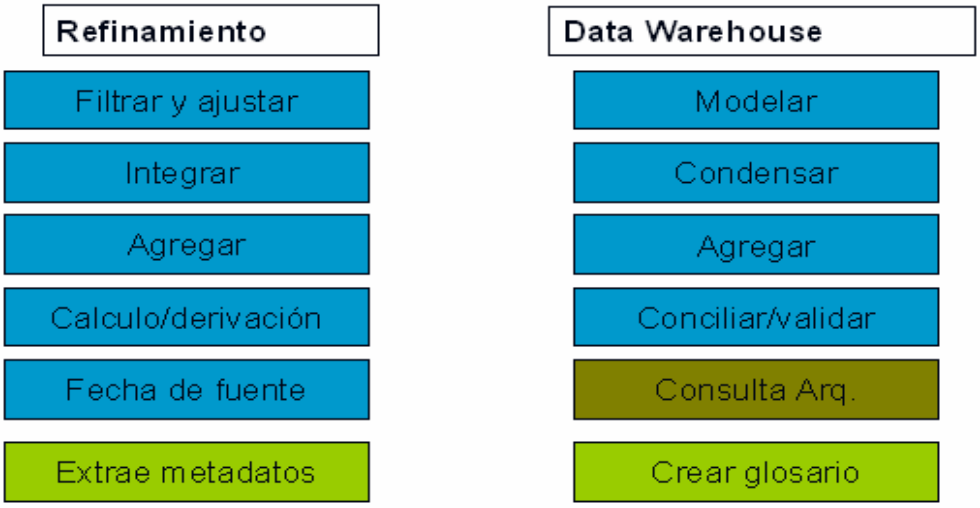
El bloque de fuentes de datos en detalle:



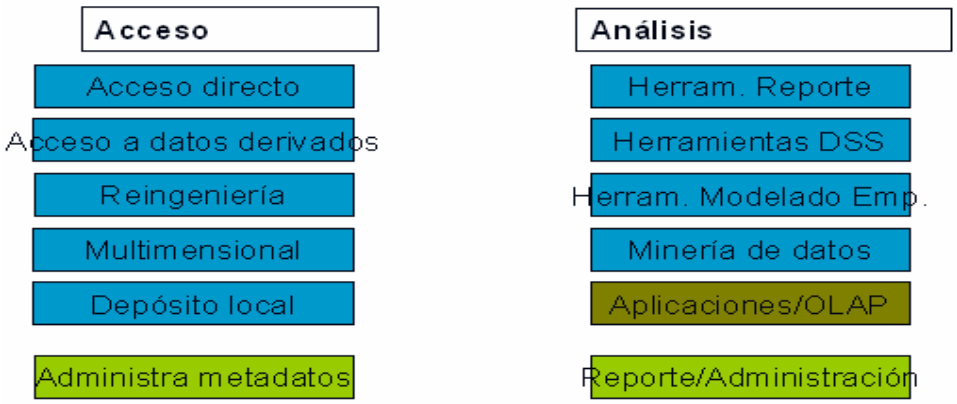
Bloque de construcción del Data Warehouse



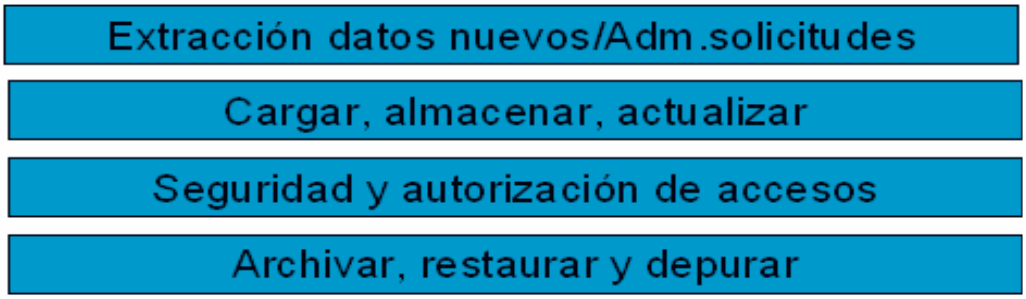
Bloque de construcción de datos derivados:



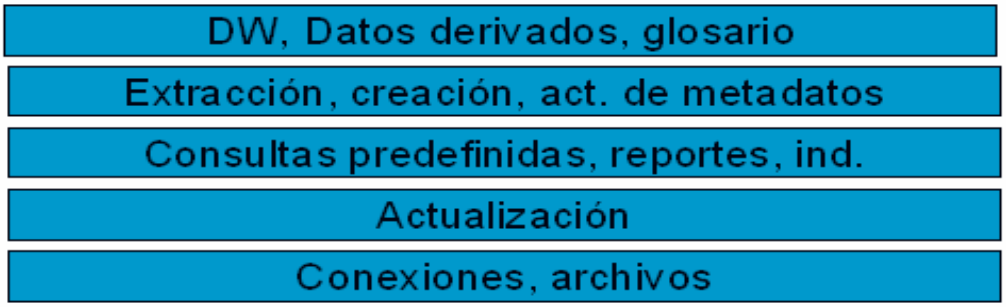
Bloque de acceso y uso del Data Warehouse:



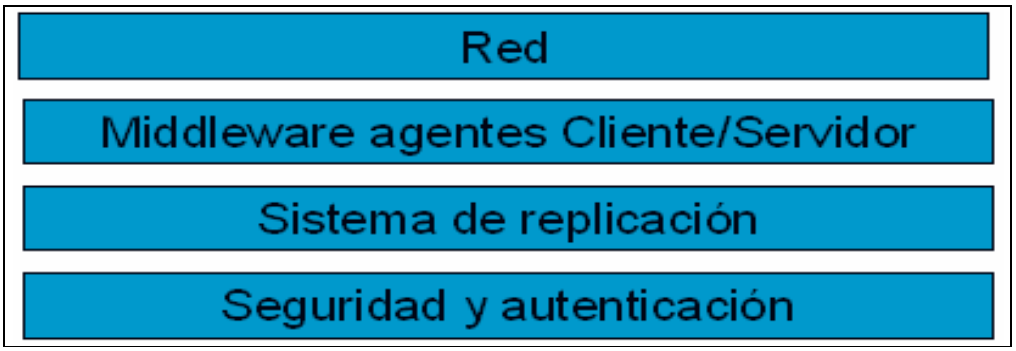
Las capas en el diagrama de Zachman:
La capa de administración de datos



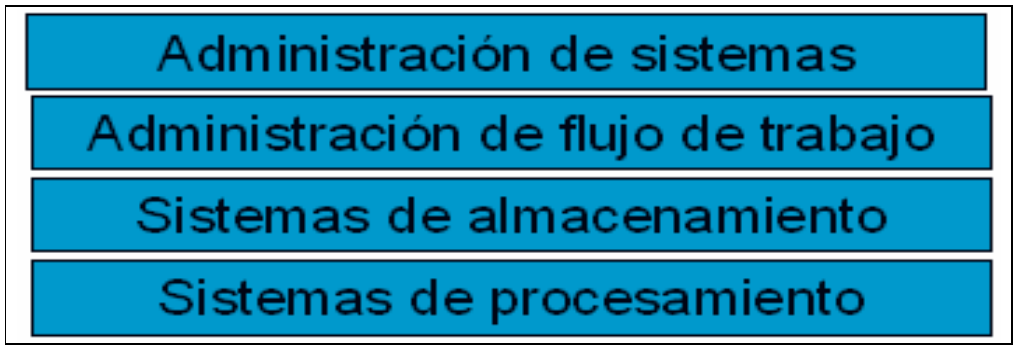
La capa de administración de metadatos



La capa de transporte:



La capa de infraestructura:



La arquitectura de referencia de Zachman para Data Warehouse nos ayuda a analizar y definir los componentes que deben ser implementados en el Data Warehouse y la forma en que se debe hacer esta implementación. Es útil también para determinar cuáles son los incrementos en la construcción del Data Warehouse.<

Autoevaluación

1. Explique el concepto de arquitectura, desde el punto de vista de los sistemas de información.

2. Explique la evolución de los sistemas de información en las cuatro generaciones.

3. ¿De qué manera ayudaría una arquitectura si se quisiera construir un automóvil?

4. ¿Qué es la arquitectura telaraña y cuáles son los problemas que presenta con respecto a la obtención de información para el análisis?

5. Dibuje la arquitectura de un Data Warehouse.

6. Explique el concepto de Datamart, proponga un ejemplo.

7. Explique el concepto de ODS (Operational Data Storage), proponga una aplicación, de un ejemplo.

8. ¿Cuál es la utilidad de un diagrama en Data Warehouse?

9. Explique las diferentes técnicas involucradas en la construcción de un Data Warehouse.

10. Explique el concepto de perspectivas de la arquitectura de Zachman, proponga un ejemplo.

11. Explique el concepto de dimensiones de la arquitectura de Zachman, proponga un ejemplo.

12. Explique los bloques y las capas de la arquitectura de referencia de Data Warehouse de Zachman.

13. ¿Cuál es la utilidad de la arquitectura de referencia de Zachman para Data Warehouse?

Para recordar

1. Históricamente, la tercera generación de sistemas de información está caracterizada por la dispersión de sistemas y bases de datos. La cuarta generación está caracterizada por la unificación desde el punto de vista del negocio.

2. Una arquitectura es necesaria porque nos dice el orden en el que se deben hacer las cosas y porque da a todos una visión de lo que será un sistema de información, gracias a que muestra un patrón reconocido universalmente.

3. La arquitectura telaraña es básicamente inestable y presenta problemas cuando de allí se quiere obtener información para el análisis.

4. El Data Warehouse ayuda a superar los problemas que presenta la arquitectura transaccional.

5. Un diagrama es útil porque permite a todos los participantes del proyecto tener una visión común.

6. En la construcción de un proyecto cualquiera, cada persona tiene una perspectiva distinta definida por el rol que desempeña.

7. Las dimensiones de la arquitectura de Zachman, nos ayudan a entender las necesidades de cada perspectiva.

8. La arquitectura de regencia de Zachman, específica para Data Warehouse, está compuesta de bloques y capas. Los bloques están relacionados al proceso de DataWarehousing y las capas son el soporte a este proceso.

9. La arquitectura de referencia de Zachman es un instrumento fundamental en el análisis, pues permite identificar en detalle todos los componentes del Data Warehouse.

SEMANA

4

La estrategia de Data Warehouse - Construcción de un Data Warehouse: Una metodología I

OBJETIVOS ESPECÍFICOS

- Comprender la importancia de una estrategia de Data Warehouse
- Presentar una metodología de construcción de un Data Warehouse

CONTENIDO

- Objetivos de una estrategia de Data Warehouse
- Aspectos generales de la estrategia
- Los dominios de la estrategia
- Las 10 reglas para tener un Data Warehouse exitoso
- Construcción del Data Warehouse
- La metodología de Barquin paso a paso

ACTIVIDADES

- Reconocer las estrategias y su importancia en el proceso de construcción de un datawarehouse

1. Objetivos de una estrategia de Data Warehouse

La estrategia de Data Warehouse, que se estudia en este capítulo ha sido planteada por Ramón Barquin, quien es consultor experto en Data Warehouse, expresidente y fundador de "The Data Warehouse Institute" (www.tdwi.org), institución líder en el mundo en materia de Business Intelligence y Data Warehouse.

Los objetivos de la estrategia de Data Warehouse planteada por Ramón Barquin son los siguientes:

- Definir la visión y una dirección de largo plazo
- Establecer un marco de trabajo para el desarrollo futuro
- Obtener consenso
- Identificar los requerimientos de infraestructura
- Establecer un cronograma inicial
- Formar un comité de administración
- Definir la visión y una dirección de largo plazo
- Establecer un marco de trabajo para el desarrollo futuro

2. Aspectos generales de la estrategia de Data Warehouse

El aspecto principal a tratar, en una estrategia de Data Warehouse, es el de la forma de construirlo. En este caso, hay tres alternativas:

- Desarrollar un Data Warehouse empresarial
- Construir DataMarts y luego crecer hacia un Data Warehouse empresarial
- Tener un enfoque mixto

Otros aspectos que se deben analizar son los siguientes:

- Adquirir nuevo hardware o utilizar el que se tiene
- Herramientas de software que se deben adquirir
- Destrezas necesarias
- ¿La empresa esta preparada?
- Áreas temáticas de mayor prioridad
- ¿Cómo puede ayudar el Data Warehouse a los usuarios?
- Identificar los factores críticos de éxito
- ¿La calidad de datos es aceptable? ¿Qué hacer si no la fuera?

16. Los dominios de la estrategia

La estrategia de Data Warehouse tiene 5 dominios. Cada uno abarca un aspecto que es relevante:

Estos dominios son los siguientes:

- Dominio del negocio
- Dominio de los datos
- Domino de los sistemas de información
- Dominio del Soporte a Decisiones
- Dominio de las personas

El dominio del negocio:

Se debe considerar los siguientes aspectos en el análisis de este dominio:

- ¿Qué es importante para el negocio?
- ¿Cómo es la estructura organizacional?
- ¿Cuáles son las metas y objetivos de los que toman las decisiones?
- ¿Qué reportes reciben estas personas?
- ¿Qué reportes utilizan?
- ¿Cuáles son las preguntas que necesitan responder para tomar decisiones?
- ¿Cuáles son sus factores críticos de éxito?

El dominio de los datos:

En el análisis de este dominio, se debe considerar los siguientes aspectos:

- ¿Que datos se recolectan?
- ¿De cuánta data hablamos?
- Donde y cómo esta almacenada?
- ¿Quién es el dueño?
- ¿Cuál es la calidad?
- ¿Se puede obtener con facilidad?
- ¿Cuáles son las estructuras?
- ¿De donde viene la data?
- ¿Qué bases de datos formales existen?
- ¿Qué manejadores de bases de datos existen?

El dominio de los sistemas de información:

En el análisis de este dominio, se debe considerar los siguientes aspectos:

- ¿Cómo es la arquitectura de Sistemas de información?
  - ¿Qué plataformas?
  - Lenguajes de programación, ¿RDBMS?
  - Red (LANs, WANs, Internet/intranet?
  - La seguridad
  - ¿Existen estándares?
- ¿Existe una arquitectura Cliente/Servidor?
- ¿Se tiene soporte multimedia?

El dominio de los sistemas de soporte a decisiones:

En el que, se debe analizar los siguientes aspectos:

- ¿Existe algún sistema de Soporte a Decisiones?
- ¿Los usuarios entienden los conceptos básicos de los Sistemas de Soporte a Decisiones?
- ¿Qué herramientas de soporte a decisiones existen?
- ¿Los usuarios están capacitados para utilizar las herramientas?

El dominio de las personas:

En este caso, se debe analizar los siguientes aspectos:

- ¿Cómo es la comunidad de usuarios?
    - Hackers
    - Lectores de reportes
- ¿Ellos saben de computación?
- ¿Cuántos usuarios se espera tener?
- ¿Cuál es su nivel de entrenamiento?
- ¿Son resistentes al cambio?
- ¿Donde están los usuarios?
- ¿Cuál es la actitud de la administración?

17. Las 10 reglas para un Data Warehouse exitoso

Tener en cuenta las siguientes 10 reglas de oro en la implementación de un Data Warehouse exitoso:

1. Tener el “sponsor” adecuado
2. El Data Warehouse es de los usuarios
3. Construir un prototipo ¡pronto!
4. Hacer que el Data Warehouse sea crítico para el negocio
5. Mostrar ejemplos concretos
6. Liberar el poder Visual del Data Warehouse
7. Educar a los usuarios de los nuevos beneficios de la solución
8. Estar cerca de la gente de la parte operativa.
9. Reconocer los cambios de prioridades en la administración
10. Tener una estrategia antes de construir el Data Warehouse

18. Introducción a la construcción del Data Warehouse (Metodología I).

La construcción de un data warehouse implica las siguientes actividades generales:

- Desarrollar una estrategia de Data Warehousing para la empresa
- Diseñar una arquitectura de alto nivel
- Escoger la tecnología, herramientas y soporte para la estructura
- Construir el Data Warehouse de manera incremental

19. La metodología de Barquin paso a paso

La metodología de Barquin, para la construcción de un Data Warehouse consta de una serie de actividades que se deben hacer en cada incremento. Cada actividad tiene determinados objetivos y determinados entregables, los que se detallan a continuación, estas actividades no se deben hacer necesariamente en el orden planteado.

6.1 Desarrollar el plan

Antes de construir se debe haber terminado con la estrategia y la arquitectura del Data Warehouse. En la estrategia, se han definido los objetivos centrales del negocio en función a la visión y misión. Luego, el plan se hace con la finalidad de hacer un listado de actividades detallado que nos permita alcanzar cada uno de los objetivos definidos en la estrategia.

Cada incremento estará orientado al logro de uno de los objetivos definidos en la estrategia. El primer paso en la construcción de un incremento del Data Warehouse es desarrollar un plan. Para el desarrollo del plan, debemos tener en cuenta los siguientes aspectos:

- Definir y establecer los objetivos específicos a ser cumplidos
- Listar los pasos a ser seguidos
- Determinar que recursos se necesitarán
- Personas y habilidades
- Tecnología
- Materiales
- Establecer los costos del proyecto
- Establecer los cronogramas del proyecto
- Establecer riesgos y establecer un plan de contingencias

6.2 Relevar los requerimientos de los usuarios

Consiste en identificar las necesidades de información específicas de cada área. Las actividades a realizar para un adecuado relevamiento son las siguientes:

- Identificar los usuarios
- Las entrevistas a los usuarios deben ser enfocadas al objetivo, breves y deben abordar el tema directamente.
- Entender los procesos de Negocio. En esto, pueden ayudar los modelos de datos existentes.
- Listar los requerimientos
- Entender los requerimientos
- Conducir el descubrimiento de la información, (inducir en las reuniones con los usuarios.)
- Crear el comité directivo de Data Warehousing
- Crear el comité consultivo de usuarios de Data warehousing. (Este comité estará encargado de validar los modelos que se vaya a diseñar.)
- Validar los requerimientos
- Alinear con la visión del negocio y la estrategia del Data Warehouse. (No se debe perder de vista los objetivos principales del negocio. Los que han sido definidos en la estrategia.)
- Alinear con la arquitectura corporativa de IT

6.3 Identificar los sistemas fuente

Es una actividad complementaria al relevamiento, y está orientada a ver si la información que se necesita para implementar el requerimiento está disponible o no. Para lograrlo, se debe:

- Estudiar y entender la arquitectura IT
- Realizar inventario de los sistemas transaccionales existentes
- Realizar inventario de los sistemas de análisis existentes
- Investigar fuentes potenciales del Data Warehouse
- Explorar e investigar fuentes externas a la empresa
- Explorar los temas de calidad de datos
- Entender la administración de cambios de los sistemas fuentes

6.4 Modelar los datos

Los modelos de datos se hacen utilizando las técnicas tradicionales para el caso del modelo del Data Warehouse y las técnicas dimensionales para el caso de Data Marts.

- Determinar si existen modelos de datos y procesos del negocio
- Revisar y validar los procesos de negocio
- Determinar si existe un repositorio de datos corporativo, modelos o herramientas

6.5 Diseñar la Base de Datos del data Warehouse

Una de las actividades críticas, en la construcción de un Data Warehouse, es el diseño de la Base de datos. Por ello se recomienda realizar las siguientes tareas:

- Alinear con los requerimientos del negocio
- Planear un nivel de staging
- Estimar volúmenes
- Considerar paralelismo y estrategias de segmentación
- Escoger un DBMS
- Identificar las necesidades de los datos derivados
- Generar scripts
- Entender los requerimientos de metadata

Así mismo se debe considerar la existencia de las siguientes tecnologías de almacenamiento:

Bases de datos relacionales: que son las utilizadas en el mundo Operacional, y que tienen buen desempeño con bases de datos grandes y buenos procesos de backup y restore.

Bases de datos multidimensionales: que son de acceso rápido, proporcionan múltiples vistas de la información pero tienen problemas cuando la Base de datos es muy grande.

6.6 Mapeo los datos

Es una tarea muy importante pues constituye la base de los procesos de ETL (extracción, transformación y carga), y permite el manejo del cambio. Consta de las siguientes actividades:

- Establecer mapeo de los requerimientos del negocio
- Determinar el rol del staging área
- Mapeo requerimientos a las necesidades de datos
- Crear el mapeo destino
- Mapeo los datos

Autoevaluación

1. ¿Por qué se debe tener una estrategia de Data Warehousing en una empresa?

2. ¿Cuál es la decisión más importante que se debe tomar como resultado de una estrategia de Data Warehousing?

3. ¿Permitirá una estrategia determinar los recursos necesarios para la construcción de un Data Warehouse?

4. ¿Cómo debe ser la estrategia frente a los sistemas de soporte a decisiones existentes?

5. ¿El Data Warehouse ofrecerá la misma herramienta de acceso a la información a todos sus usuarios?

6. Si se tiene definido primer tema a construir y se sabe que la calidad de los datos del sistema fuente es muy mala, y hay datos que no se capturan, ¿cuál sería el siguiente paso?

7. ¿Cuáles son las diferencias entre Estrategia y Plan?

8. ¿Cuáles son las diferencias entre una BD relacional y una BD multidimensional?

9. ¿Cuál es el objetivo del mapeo de los datos?

Para recordar

1. La estrategia de Data Warehouse, definirá la visión y misión de lo que se quiere lograr y estará alineada con la visión y misión del área de sistemas y de la empresa.

2. Uno de los objetivos principales de la estrategia es definir la forma en que se va a construir el Data Warehouse.

3. Los cinco dominios de la estrategia son temas que se deben considerar en el inicio de todo proyecto de Data Warehouse.

4. Tener el patrocinador correcto, hacer que los usuarios sean dueños de cada proyecto, mostrar resultados pronto, entre otras, son las reglas que no se deben perder de vista en un Data Warehouse.

5. El plan de Data Warehouse está orientado al cumplimiento de objetivos específicos previamente definidos por la estrategia.

6. Las actividades de la metodología no se tienen que hacer necesariamente en el orden planteado.

SEMANA

5

Construcción de un Data Warehouse: Una metodología II - Análisis de los requerimientos empresariales

OBJETIVOS ESPECÍFICOS

- Presentar una metodología de construcción de un Data Warehouse.
- Comprender los criterios básicos de análisis de los requerimientos empresariales.

CONTENIDO

- La metodología de Barquin paso a paso.
- Introducción al análisis de los requerimientos
- Análisis de los requerimientos empresariales
- Análisis de las fuentes de datos

ACTIVIDADES

- Analizar una consulta empresarial aplicando el método del análisis de la consulta empresarial.

La metodología de Barquin paso a paso(continuación)

La metodología de Barquin, para la construcción de un Data Warehouse, consta de una serie de actividades que se deben hacer en cada incremento. Cada actividad tiene determinados objetivos y determinados entregables, los que se detallan a continuación, estas actividades no se deben hacer necesariamente en el orden planteado.

6.7 Extraer los datos

Se deben realizar las siguientes actividades:

- Conceptuar los procesos de extracción
- Alinear los procesos de extracción al mapeo de datos
- Determinar el rol del staging área
- Considerar actividades de transformación y limpieza
- Escoger la data a extraer y el software de transformación
- Extraer los datos requeridos y colocarlos en el staging área (o direccionarlos en el data warehouse destino)
- Validar y probar los procesos de extracción de datos

6.8 Limpiar los datos

Es una tarea ardua que implica procesos de gestión de datos y de cambio en los sistemas de captura, para lograrlo:

- Conceptualizar los procesos de limpieza de datos
- Considerar necesidades de limpieza, sincronización y estandarización
- Establecer métricas de calidad mínima
- Determinar rol de la metadata
- Escoger el software de limpieza de datos
- Diseñar los procesos generales de limpieza
- Limpiar la data
- Validar y probar los procesos de limpieza

6.9 Transformar los datos

Depende del modelo de datos que se haya definido para el Data Warehouse y consiste en:

- Revisar la visión de los procesos de transformación de datos
- Detallar y describir las derivaciones necesarias, sumarizaciones y/o otras operaciones
- Determinar el rol del staging layer
- Determinar los metadatos
- Escoger el software de transformación de Datos
- Transformar la Data
- Validar y probar los procesos de transformación y los datos

6.10 Cargar el Data Warehouse

Es un proceso que tiene ciertas complicaciones. Consta de las siguientes actividades:

- Conceptualizar los procesos de carga
- Desarrollar el plan de carga
  - Calcular el tiempo
  - Establecer ventanas
  - Preparar la infraestructura técnica
  - Preparar el software y los datos
  - Desarrollar el plan de contingencia
- Considerar el rol del staging área
- Cargar los datos
  - Desarrollar y probar la carga inicial
  - Cargar en producción el Data Warehouse
- Validar la data cargada

6.11 Implementar la Metadata

En esta actividad, se deben crear los datos acerca de los datos, esto implica la creación de un repositorio que proporcione información que puede ser de tres tipos:

- Metadatos del negocio, que contienen las reglas del negocio que han definido para el data Warehouse, entidades y atributos.
- Metadatos técnicos, que contiene los modelos de datos a nivel técnico, así como los modelos de los procesos de carga.
- Metadatos operacionales, que son acerca de los procesos del data Warehouse, frecuencia de ejecución, prioridad entre otros. Estos metadatos permitirán administrar el Data Warehouse.

En general, un repositorio de Metadatos debe contener lo siguiente:

- Nombres de campos y definiciones
- Mapeo de los datos
- Tablas
- Índices
- Cronogramas de extracción, carga, etc.
- Criterios de selección
- Cálculos de los datos derivados
- Transformación de los datos

6.12 Establecer los procesos de administración

Para administrar el Data Warehouse se debe desarrollar las siguientes actividades:

- Desarrollar un plan de operación y mantenimiento del Data Warehouse
- Establecer un plan de administración de las operaciones de back-end
- Establecer un plan de administración de las operaciones de metadata
- Establecer un plan de administración de las operaciones de acceso de los usuarios
- Establecer un plan de administración del cambio

Por otra parte, se debe documentar los modelos y procesos de Data Warehouse. Así como establecer procedimientos de monitoreo del tamaño y uso de los datos, y procedimientos de administración de la seguridad por perfiles de acceso o por resguardo de información reservada.

6.13 Crear las aplicaciones del Data Warehouse

Es una de las actividades que tiene especial importancia debido a que permite al usuario el acceso y la exploración de la información que está en el Warehouse. Se debe tener en cuenta los siguientes criterios:

- Alinear con la visión del negocio y los requerimientos del usuario
- Desarrollar dentro de área de negocio y añadir prioridades
- Listar y documentar consultas orientadas a los requerimientos de usuarios
- Desarrollar pantallas de prototipos y revisar con el usuario
- Considerar los tipos de aplicaciones como alertas, herramientas OLAP, y minería de datos, herramientas de consultas y reportes
- Validar y probar los procesos de administración

6.14 Probar y validar el Data Warehouse

Considerar:

- Desarrollar un plan de prueba y validación
- Comprometer a los usuarios finales
- Establecer parámetros y métricas de prueba
- Validar la data
- Reconciliar los principales sistemas de soporte

6.15 Entrenar al staff y a los usuarios finales

Esta tarea es muy importante, pues Data Warehousing es un proceso que implica mucho aprendizaje. Los pasos a seguir son:

- Determinar los requerimientos de entrenamiento necesario
- Desarrollar el plan y calendario del entrenamiento
- Diseñar el contenido del entrenamiento

6.16 Implementar y hacer el siguiente incremento.

Se debe hacer un plan de implantación que permita la aceptación del producto por los usuarios.

Finalmente, se debe saltar al siguiente incremento.

Resumen de la Metodología

El plan es el conjunto detallado de tareas que deben hacerse para concretar las recomendaciones de la estrategia.

El relevamiento de requerimientos se debe complementar con un análisis de la situación de los datos.

Se debe llegar a un consenso cuando se trata de definir conceptos de negocio. Es muy frecuente que dos términos de negocio tengan el mismo nombre, pero diferentes significados dependiendo del área. Este problema no debe persistir en la implementación del Warehouse.

Extraer, limpiar, transformar y cargar son las actividades más costosas en la construcción de un Data Warehouse y pueden consumir el 80% del tiempo y los recursos.

Las aplicaciones de acceso al Warehouse son especialmente importantes, pues constituyen la herramienta del usuario para el uso del Data Warehouse.

7. Introducción al Análisis de los requerimientos

Las necesidades del negocio son la razón de ser del Data Warehouse. En este sentido, el análisis debe concentrarse en las necesidades empresariales y los métodos a aplicar deben representarlas de manera adecuada dentro del Data Warehouse.

8. Análisis de los requerimientos empresariales.

Se tienen dos métodos de análisis de los requerimientos empresariales. Estos son complementarios, pues el primero está orientado a definir el tema de negocio a abordar y el segundo está orientado a detallar el requerimiento dentro de este tema de negocio. Los métodos son los siguientes:

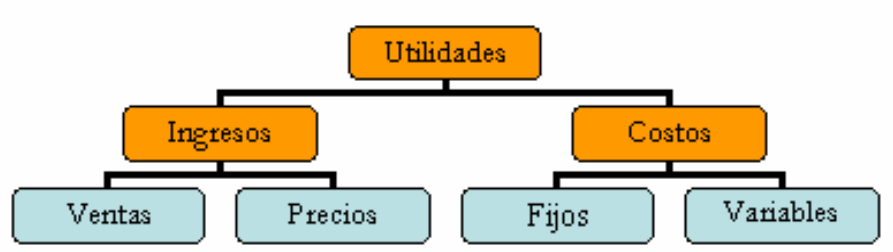
- Análisis de arriba hacia abajo
- Análisis de la consulta empresarial

8.1 Análisis de arriba hacia abajo

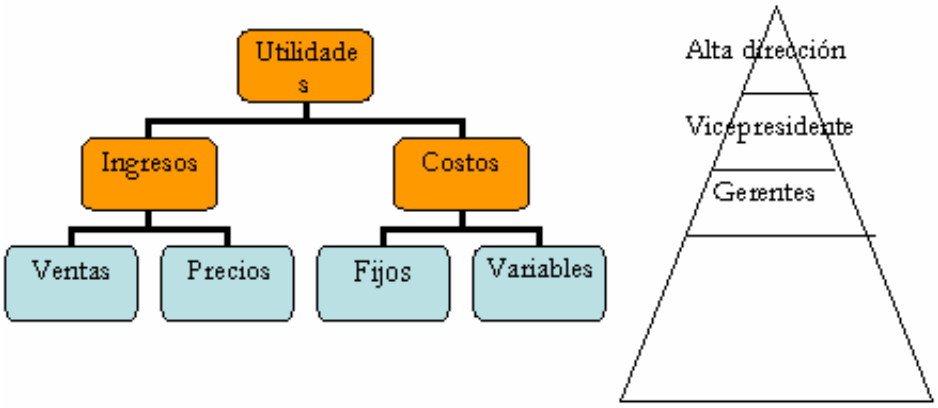
Permite la selección de la información correcta para el Data Warehouse. Los objetivos empresariales, desde la visión de arriba hacia abajo, enfocan el área en la que puede hacer un incremento del Data Warehouse.

El análisis de arriba hacia abajo tiene dos pasos:

Primero centrarse en los objetivos



Segundo, relacionar los objetivos organizacionales con las funciones de la organización.



Según el gráfico, se puede definir como áreas temáticas la administración de ingresos o la administración de costos. Cada tema requiere un conjunto diferente de información que debe manejar el data Warehouse, y un conjunto diferente de técnicas de análisis que deben emplear los usuarios finales.

La administración de ingresos tiene mucha relación con el pronóstico de ventas a futuro con base en las ventas pasadas. Los patrones ambientales y las tendencias de compras, también se puede requerir de fuentes de información externas. Por otro lado la administración de costos tiene que ver con el control operacional y la vigilancia de varias medidas de costos empresariales. La evaluación comparativa es una herramienta para la administración de costos.

8.2 Análisis de la consulta empresarial

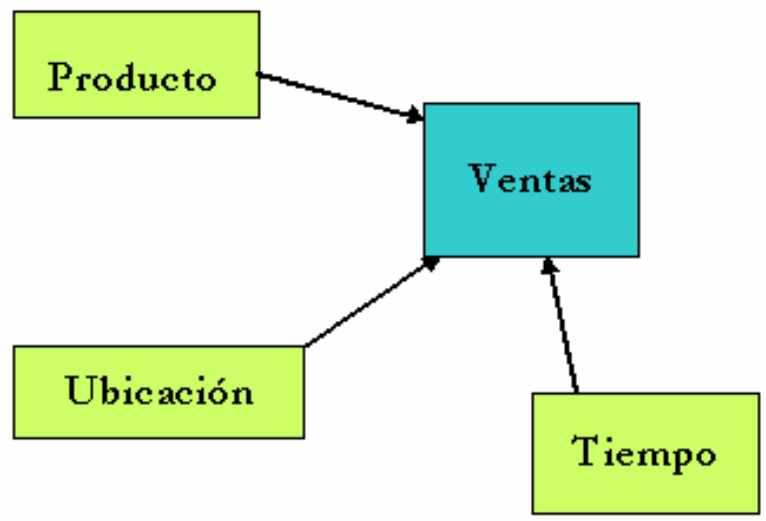
La visión de la consulta empresarial es la perspectiva de los datos del Data warehouse desde el punto de vista del usuario. Una de las razones de la popularidad del modelo de datos multidimensional o estrella, es que es un reflejo cercano de la forma en que un analista empresarial visualiza una consulta. De hecho, una tabla multidimensional es una representación exacta de una consulta multidimensional. Por ello, para el analista, proponer la consulta es lo mismo que consultar directamente una tabla multidimensional. De ahí que la consulta empresarial sea una solicitud de hechos, a veces llamados mediciones o medidas de varias dimensiones.

Las consultas empresariales contienen además sub-consultas o puntos de corte. Por ejemplo, por mes, ciudad o país. Estos puntos de corte deben también incorporarse como dimensiones y vincularse a las tablas de hechos.

Una consulta típica es la siguiente:

"Se necesita analizar las ventas de productos por tipo en todas nuestras tiendas en los últimos 12 meses"

Esta consulta se puede representar de la siguiente manera:



De esta manera, se define un modelo en el que las entidades tiempo, ubicación, producto son las dimensiones que permiten ver las ventas (medidas) desde cada una o desde la combinación de mas de una dimensión.

9. Análisis de las fuentes de datos.

Se analizan las fuentes de información con la finalidad de ver si es factible atender los requerimientos definidos en el análisis de los requerimientos empresariales. Los aspectos a considerar en el análisis son los siguientes:

- Tecnologías de almacenamiento
- Definiciones múltiples
- Campos nulos
- Formatos diferentes
- Codificación diferente
- Duplicidad

9.1 Tecnologías de almacenamiento

Se refiere al análisis de los tipos de datos, y de las plataformas que los soportan. Entre los tipos principales están las Bases de datos relacionales, archivos, datos comprados, datos no estructurados.

9.2 Definiciones múltiples.

Se pueden dar dos situaciones:

- Dos elementos de datos que tienen el mismo contenido pero diferente nombre
- Dos elementos de datos que tienen el mismo nombre pero diferente contenido

Estas deben ser evaluadas cuidadosamente.

9.3 Campos nulos

Se debe analizar las fuentes estructuradas con la finalidad de detectar campos nulos, como en la figura 1.

9.4 Formatos diferentes

Cuando campos que tiene el mismo significado se almacenan en formatos diferentes en bases de datos distintas. Dos tablas que tienen la fecha de nacimiento de los empleados, en una de ellas esta en alfanumérico y en la restante en fecha-hora.

Nro documento	Nombre	Saldo
10234567	Jorge Basadre	230,000.00
23451233		20,000.00
11558800	Pedro Picapiedra	
32456722	Vilma Marmol	
	Tony Mestas	10,000.00
	Toño	5,000.00

Figura 1

9.5 Codificación diferente

Cuando los dos campos de distintos sistemas transaccionales contienen la misma información pero codificada de distinta manera.

9.6 Duplicidad

Se da en los casos en que hay registros duplicados en tablas donde deberían ser únicos.

En los campos de tipo texto cuando se tiene diferentes longitudes y el dato se corta en el de menor longitud originando duplicados.

Autoevaluación

1. ¿Cuál es el orden en el que deben hacer las actividades de la metodología de Barquin?

2. ¿Cuáles son las actividades más costosas en la construcción de un Data Warehouse?

3. ¿En qué actividad se deben considerar los aspectos del crecimiento de la Base de datos y el tema de seguridad?

4. ¿Cuántas veces se debe repetir el ciclo de la metodología?

5. ¿Cómo se deben aplicar los métodos de análisis de requerimientos?

6. ¿Cuál es la característica principal del método de Arriba hacia abajo?

7. ¿Cuáles son las diferencias entre el método de arriba hacia abajo y el del análisis de la consulta empresarial?

8. ¿Cómo se debe hacer un análisis de las fuentes de datos en la etapa de Análisis de requerimientos?

CASO:

En una cadena de videos que cuenta con 70 tiendas a nivel nacional, se quiere hacer un análisis de los videos rentados por tipo de película, por ciudad, por distrito, por director, por año, mes y semana del mes de acuerdo con cada tipo de cliente.

Para recordar

1. Las tareas de extracción, limpieza, transformación y carga de los datos significan el 80% del tiempo y de los recursos de construcción del Data Warehouse.

2. Existen dos tecnologías de almacenamiento de Datos aplicables a Data Warehouse, estas son las BD relacionales y las BD multidimensionales

3. El método de análisis de Arriba hacia abajo y el método del análisis de la consulta empresarial son complementarios.

4. Una consecuencia del método del análisis de la consulta empresarial, es el modelamiento dimensional.

5. Todo análisis de requerimientos en Data Warehouse debe tener una parte de análisis de las fuentes de datos.

SEMANA

6

Planificación de un DataWarehouse – Identificación de requerimientos de negocio empresariales

OBJETIVOS ESPECÍFICOS

- Comprender la importancia de un proyecto Datawarehouse

CONTENIDO

- Proyecto de planeamiento y administración.
- Definición de requerimientos

ACTIVIDADES

- Determinar la importancia de planificar un proyecto

1. Proyecto de planeamiento y administración

Dentro de los diferentes tipos de proyectos de Data Warehouse, se pueden determinar diferentes factores que los caracterizan.

El factor más crítico es el de contar con un fuerte patrocinador de negocios. Ellos deberían tener una visión del potencial impacto del data warehouse en la organización. Debe tener la capacidad de "convencer" a sus pares para que apoyen el proyecto.

Un riesgo es cuando existe un solo patrocinador, ya que podría estancarse el proyecto si éste decide dejar la empresa o atender otros asuntos.

Si al empezar un proyecto de Data Warehouse no se encontrara a un patrocinador, no es razón para parar el proyecto. Lo que ocurrirá es que el proyecto se desarrollará con lentitud.

Otro factor importante es tener una fuerte compenetración y motivación en la construcción del data warehouse.

Un tercer factor es la viabilidad, existen puntos importantes a considerar como la parte tecnológica, o de recursos, pero la más crítica es la viabilidad de los datos. Es decir, si tenemos los datos en los sistemas operacionales para poder hacer el análisis esperado.

El siguiente factor no es decisivo para la continuidad del proyecto pero si influye en el éxito de éste. Este factor tiene que ver con la relación entre los negocios y la organización de TI. Aunque dicha relación no esté en armonía, el proyecto puede ser una excelente oportunidad para que ambos frentes avancen al mismo compás.

El último factor se relaciona con la actual cultura analítica en la compañía; es decir, si los analistas toman decisiones basadas en hechos y figuras o son basadas en su intuición o hechos anecdóticos.

a) **Alcance del proyecto**. - Debe estar alineado con la administración y organización. Al inicio puede centralizarse en un solo proceso de negocio.
b) **Justificación.** –No olvidar que requiere una justificación entre el costo y beneficio.
c) **Staff.** – El proyecto requiere la integración de una cantidad fundamental de recursos tanto de negocios como de TI. Los nombres de los cargos podrían variar pero proponemos algunos:
    a. Promotor de negocio
    b. Dirigente de negocio
    c. Líder de negocio
    d. Usuarios de negocio
    e. Analista de sistemas de negocio
    f. Experto en el tema del negocio: nivel de análisis muy agudo y participativo en el modelamiento.
    g. Desarrollador de aplicaciones analíticas
    h. Educador del data warehouse

ADAPTADO DE :

KIMBALL Ralph, ROSS Margy
2002. The Data Warehouse Toolkit – The complete guide to dimensional modeling. Editorial :John Wiley & Sons Inc. Ciudad : New York. Pag: 331-340

Definición de requerimientos

El entendimiento de los requerimientos es esencial para el éxito del data warehouse.

El levantamiento de información se debe hacer de manera adecuada y en reuniones no muy técnicas (en términos) con el usuario. El objetivo es hablar con ellos referente a ¿Qué es lo que ellos hacen? , ¿Por qué lo hacen? , ¿Cómo es que lo hacen? , ¿Cómo esperan hacer decisiones en el futuro?.

Podemos plantear dos formas de recabar información :

- Entrevistas: alta participación del usuario.
- Sesiones de facilitación

Si decidimos realizar una entrevista, es necesario, definir al entrevistador. Es muy útil tener a otra persona en la reunión que escriba lo que se dice en la reunión. Además, no pretender conocer todo o saber todo.

En el momento de las reuniones con el usuario, se debe:

a) Inicio

1. Establecer prioridades
2. Plantear los objetivos
3. Focalizarse en el proyecto y objetivos de la entrevista, evitando mezclar temas de software y hardware.

b) Flujo de entrevista

1. Consiga que el usuario opine.
2. Pregunte sobre sus responsabilidades y como encajan en la empresa.
3. Pregunte sobre sus KPI (Indicadores de performance).
4. Determine como estos KPI, se traducirán al modelo dimensional.
5. En caso el entrevistado tenga mas experiencia en los datos, podría tratar de bosquejar la dimensionalidad del negocio.
6. Si el entrevistado es aún mas analítico, podría preguntarle referente a los tipos de análisis que realiza, con el fin de tener información referente al acceso de datos y las herramientas a utilizar.
7. Si se reúne con ejecutivos de negocio, no podemos ir al detalle como en los casos anteriores, en lugar debería preguntar sobre su visión para un mejor uso de la información en la organización, es decir, nuestro entregable debe satisfacer las demandas y expectativas del negocio.

c) Resumen

1. Pregunte al usuario sobre cómo sería un proyecto exitoso (reportes fáciles o reportes útiles, etc.).
2. Exprese que no todo lo que se dijo se deberá implementar en una primera fase.
3. Agradezca la participación de los usuarios por sus valiosos aportes.

d) Entrevistas referente a los datos.
   1. Entérese de la data disponible
   2. Entérese de la calidad de la data
   3. Analice si los requerimientos demandados son satisfechos con la data disponible.

e) Documentación Post-Levantamiento de información.

   1. Revise sus apuntes.
   2. Revise los reportes proporcionados por los usuarios, con la finalidad de enriquecer la dimensionalidad.
   3. La documentación es necesaria, aunque no muy del agrado del personal.
   4. La documentación sirve para la validación del usuario y como referencia para el equipo del proyecto sobre los temas tratados en las reuniones.

ADAPTADO DE :

KIMBALL Ralph, ROSS Margy
2002. The Data Warehouse Toolkit – The complete guide to dimensional modeling. Editorial :John Wiley & Sons Inc. Ciudad : New York. Pag: 340-347

Autoevaluación

1. ¿Por qué documentar el proyecto?
2. ¿Por qué planificar un proyecto?
3. ¿Qué factores críticos puede usted mencionar para un proyecto data warehouse?

Para recordar

1. Un proyecto debe tener un patrocinador
2. Es importante las entrevistas con los usuarios ya que son ellos justamente quienes nos dirán que necesitan y de ésta forma dar una solución adecuada.
3. Nuestra solución debe ir de la mano con la visión de la empresa con respecto al proyecto.
4. Es importante tener la documentación de las reuniones y de las fases del proyecto, ya que en ellas tendremos información valiosa de los usuarios las cuales serán tomadas como referencia en la solución analítica.

SEMANA

8

Modelamiento de datos en Data warehouse – Conceptos de Modelamiento dimensional

OBJETIVOS ESPECÍFICOS

- Identificar las técnicas de modelamiento de datos utilizadas en Data warehouse.
- Comprender la técnica de modelamiento dimensional.

CONTENIDO

- Conceptos preliminares
- Modelamiento de datos en la arquitectura de Zachman
- Modelamiento del tiempo en Data warehouse
- El proceso de negocio
- El modelo dimensional o esquema estrella
- Ventajas del esquema estrella

ACTIVIDADES

- Determinar la importancia de modelo dimensional

o Conceptos preliminares

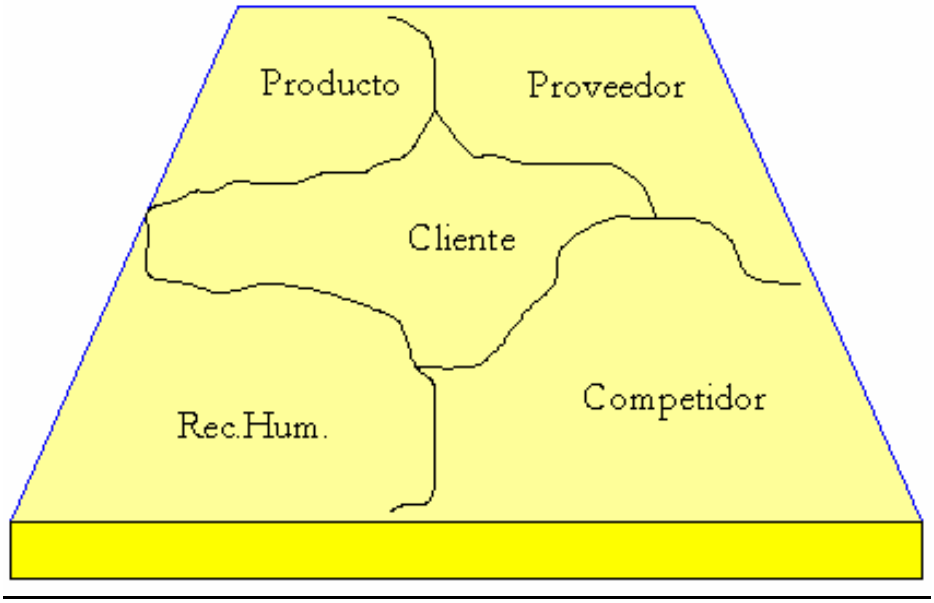
Los componentes fundamentales de la arquitectura de datos del data warehouse son los siguientes:

- Área temática
- Modelo conceptual
- Modelo lógico
- Modelo físico

Área temática

Un área temática es una entidad primaria que es importante para la organización. Un área temática típicamente es un sustantivo, por ejemplo Cliente, producto, recurso humano entre otros.

Todo negocio está compuesto de un conjunto de áreas temáticas:

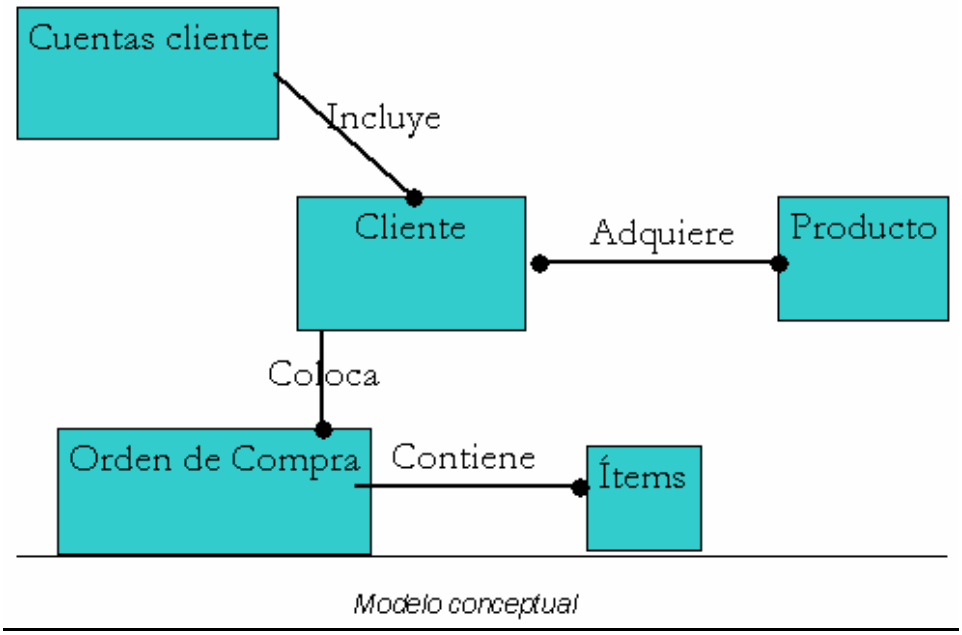


Modelo conceptual

Representación gráfica y textual del análisis que identifica los datos que necesita una organización para lograr su misión, sus metas, sus objetivos, funciones y estrategias. Un modelo de datos identifica entidades y sus relaciones entre ellas, proporcionando una visión conceptual del negocio.

Modelo lógico

Es el modelo que representa las entidades y su estructura inherente. Además de las relaciones entre ellas, es independiente de las aplicaciones individuales. Contiene la implementación de los atributos de las entidades y las reglas de negocio (Diagrama entidad-relación).



Modelo conceptual

Modelo físico

Es la instancia física del modelo lógico. Está conformado por los estándares de codificación, tipos de datos, longitudes, constraints, índices, particiones.

o Modelamiento de datos en la arquitectura de Zachman

Si se construye la matriz de Zachman para el caso de modelamiento de datos en Data Warehouse, se obtiene lo siguiente:

	Dato	Función	Red	Persona	Tiempo	Motiv.
Alcance	A.T.	Función. Principal.	Ubic. Emp	Princ. depart	Evento Emp	Metas
Modelo Empres	Mod.Con.	Proceso Negocio	Ubic Red	Depart.	Evento Neg	Objetiv
Modelo sistema	Mod.Log.	Aplicación	Func Nod	Sección	Evento Sist	Regla Neg
Modelo tecnolog.	Mod.Fisic.	Arq. De Aplic.	Inter Nod	Usuario	Ejecución	Caract Operac
Compon	BD	Program.	Red	Aut.	Interrup.	Car Mo

Leyenda:
Mod.Con. : Modelo conceptual
Mod.Log. : Modelo lógico
Mod. Fis. : Modelo físico
Bd : Base de datos
A.T. : Área temática

o Modelamiento del tiempo en un Data Warehouse

Entre las diferencias principales de una base de datos de Data warehouse y una operacional están el carácter histórico y la no-volatilidad de la primera.

Es necesario que las entidades del Data warehouse tengan los atributos que le permitan almacenar la historia de los datos y un registro de todos los cambios. Hay diversas maneras de lograr esto. En este capítulo, se presentan las que serían aplicables al caso del Warehouse. El caso de los Data Marts se tratará en los capítulos de modelamiento dimensional. Las formas son las siguientes:

- **El tiempo en una tabla agregada**

Dependiendo de la frecuencia de actualización de la tabla, bastará con colocarle un campo que ayude a identificar el periodo de proceso de la data. Por ejemplo si se tiene una tabla conteniendo el cliente y la compra mensual realizada por este cliente se tendría:

Compra mensual x cliente
Mes
DNI
Nombre
Número de artículos
Monto de compra del mes
Teléfono

Si en este se agrega el campo “Mes” se puede tener las compras mensuales e históricas del cliente por cada mes.

- **El tiempo en una tabla detallada de actualización diaria**

Para el caso de una tabla de actualización diaria, que pretende mantener los cambios por cada uno de los registros independientemente del día en que se haya producido, se debe colocar dos campos fecha. El primero indicará la fecha de inicio de vigencia del registro y el segundo indicará la fecha de fin del registro. Por defecto, estos campos tienen un valor muy antiguo para el caso de inicio de vigencia del primer registro y una fecha a un futuro lejano para el caso del último registro vigente.

Ejemplo: Se desea modelar la tabla de clientes, manteniendo la historia de los cambios de cada registro

Para conseguirlo, se ponen dos columnas adicionales una fecha de inicio de vigencia y otra de fin de vigencia del registro. La columna de fecha de fin de vigencia pasa a ser parte de la llave primaria para evitar duplicidad de registros con la misma vigencia.:

Cliente
Identificador único de cliente PK
Fecha de fin de vigencia PK
Fecha de inicio de vigencia
DNI
Nombre
Dirección

4. El proceso de negocio

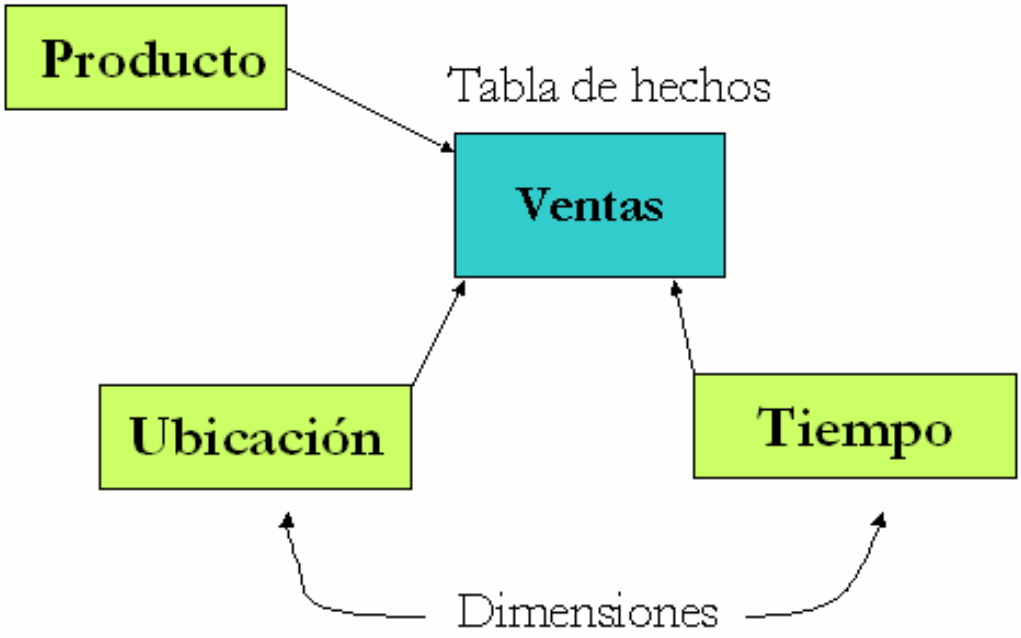
En el análisis de la consulta empresarial se debe identificar lo que los usuarios hacen con la información, de dónde viene la información y como esta debe ser transformada para cobrar significado.

Como consecuencia del análisis de la consulta empresarial se origina el modelo dimensional o esquema estrella.

5. El modelo dimensional o esquema estrella

El modelo dimensional le permite al usuario ver la data mediante múltiples dimensiones, por ejemplo ver las ventas por producto, por tienda, por mes por año. Un modelo dimensional es un modelo simple que muestra medidas, dimensiones y sus relaciones y que puede ser presentado al usuario para verificación. La información deberá ser presentada utilizando etiquetas de negocio que le sean familiares al usuario final. Este modelo puede ser utilizado para crear un esquema físico.

Un modelo dimensional se crea para dar respuesta a requerimientos de análisis como el siguiente: *"¿Cuáles fueron los 10 productos más vendidos fabricados por la compañía XYZ basados en las ventas totales por sector para cada trimestre de los dos últimos años?".*



Medidas:

Las medidas dicen lo que está ocurriendo en el negocio, son datos cuantitativos acerca de un área temática. Responden a la pregunta ¿Cuánto? o ¿Cuántos?, y generalmente son numéricos.

Ejemplos:
-¿Que sectores producen las ***utilidades*** más altas en el año?
-¿Cuál fue la ***ganancia*** por vendedor?
-¿Cuántas ***unidades*** fueron vendidas por cada producto?

Una medida puede basarse en una columna de una tabla del sistema operacional o puede ser calculada, y se almacena en la "Fact table" o tabla de hechos en el Warehouse.

Dimensiones:

Las dimensiones son los calificadores que dan sentido a las medidas, organizan los datos en base a los componentes de una pregunta, por ejemplo ¿qué?, ¿dónde?, ¿cuando?

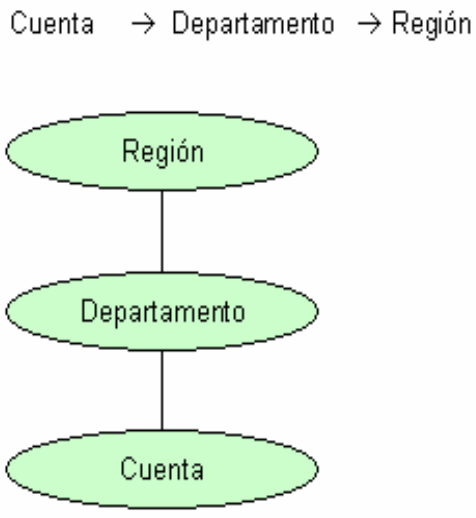
Las dimensiones se almacenan en tablas denominadas tablas de dimensiones.

Elementos de una dimensión

Cada dimensión está compuesta por ítems relacionados o elementos. En general, las dimensiones son jerararquías de Ítems relacionados. Cada elemento representa un nivel diferente se agregación.

Las jerarquías en una dimensión permiten hacer "Drill Down" o "Drill Up".

La dimensión geografía:

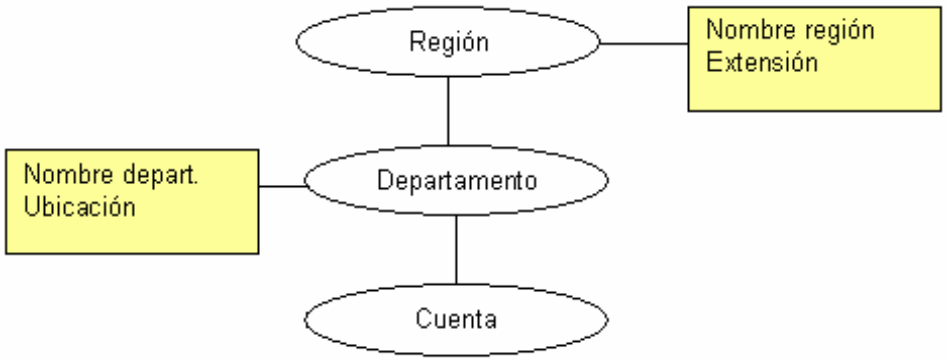


Atributos de una dimensión

Los atributos contienen descripciones y otra información asociada con los elementos de la dimensión. Por ejemplo, el atributo nombre de la cuenta contiene

la descripción del elemento cuenta. El elemento departamento, puede tener como atributos el tamaño del departamento, la cantidad de habitantes, entre otros.

Los atributos facilitan al usuario final la construcción de las consultas haciendo uso de términos de negocio con los cuales ellos estén familiarizados.



6. Ventajas del esquema estrella

i. Sencillo, porque es fácil visualizar la consulta empresarial en un modelo, y es de fácil entendimiento por el usuario.
ii. De fácil acceso, pues los hechos o medidas se pueden visualizar a través de algunas o todas las múltiples dimensiones del modelo.
iii. Buen tiempo de respuesta, por la forma en que se ha diseñado este modelo proporciona tiempos de respuesta muy buenos cuando se hacen consultas.
iv. Es un modelo des-normalizado y orientado al análisis.

Autoevaluación

1. Mencione los componentes de una arquitectura de datos de Data warehouse.

2. Mencione tres áreas temáticas para Cibertec.

3. Describa las razones por las que se construye la matriz de Zachman del acápite 2 de este capítulo

4. Mencione las principales diferencias entre un modelo lógico y un modelo físico.

5. Los métodos de modelamiento del tiempo explicados son aplicables al Data warehouse, ¿serán estos campos suficientes si además se sabe que en estas estructuras se integrarán diversas fuentes de información?
6. ¿Cuál es la consecuencia del análisis de la consulta empresarial?

7. ¿Por qué se dice que el modelo dimensional es de fácil entendimiento por parte del usuario?

8. ¿Las jerarquías de una dimensión son conceptos que se plasman en el modelo físico?

9. ¿Se pueden tener fechas como medidas?

Para recordar

1. La base de datos del Data warehouse debe tener estructuras que puedan garantizar la historia de la información y la no volatilidad así como la integración.

2. El área temática es la característica que diferencia a una base de datos "Orientada a un tema" de una base de datos orientada a la aplicación.
3. El modelo dimensional permite ver la data a través de múltiples dimensiones de análisis.

4. Las medidas cuantifican el negocio o los hechos del negocio.

5. El modelo dimensional es de fácil entendimiento.

6. El modelo dimensional es flexible y tiene buen tiempo de respuesta.

SEMANA

9

Modelamiento dimensional: Caso práctico y Conceptos Avanzados

OBJETIVOS ESPECÍFICOS

- Aplicar la técnica del modelamiento dimensional.
- Medidas y dimensiones especiales

CONTENIDO

- Caso de modelamiento dimensional
- Casos a tener en cuenta

ACTIVIDADES

- Desarrollar los casos de modelamiento dimensional planteados.

1. Caso de modelamiento dimensional

Las siguientes entrevistas se hicieron en una empresa:

Jorge Acosta, Vicepresidente de la empresa:

"No estoy satisfecho por la forma en que utilizamos la información que tenemos sobre nuestros clientes para obtener una ventaja competitiva. Necesitamos ser capaces de analizar las tendencias de nuestros compradores por producto y por región. Esto nos ayudará a desarrollar campañas especiales de marketing que produzcan el máximo impacto."

Pedro Salazar, Analista de mercados:

"Actualmente, me toma mucho tiempo obtener la información que necesito para el análisis que hago, y por eso quisiera poder tener un sistema que me ahorre tiempo en la obtención de la información. Necesito comparar la rentabilidad y costo para todos nuestros productos y proveedores. Necesito conocer a nuestros mejores clientes y por supuesto, saber cuales son sus patrones de compras en los últimos 2 años.

Hasta este año nuestra base de clientes se ha dividido en distritos de venta que corresponden a la ciudad donde están localizados. Pero acabamos de crear dos regiones: la región 1 para Lima, y la región 2 para todas las otras ciudades del interior.

Los reportes que mas necesito son:
Rentabilidad mensual, costo, ganancia neta por línea de producto y por proveedor.
Rentabilidad y unidades vendidas por producto, por región y por mes.
Rentabilidad mensual por cliente
Rentabilidad trimestral por proveedor.

La mayoría de nuestros análisis están basados en resultados mensuales. Es posible que necesitemos bajar a un detalle semanal o para un periodo contable

Diana Zúñiga: Administradora de datos:

La data del sistema de ingreso de órdenes es almacenada en la Base de datos "'Ordenes" y, actualmente, proporciona toda la información necesaria para el área de marketing. El código de producto utilizado por los analistas se almacena en la tabla "Catalogo" como Número de catálogo. El código de línea de producto se almacena en la tabla "Stock" como el Número de stock. Y el nombre de la línea de producto se almacena como descripción. En cuanto a las líneas de productos, cada una tiene muchos productos y cada fabricante tiene también muchos productos.

El sistema de compras es completamente distinto y tiene toda la data de costos por cada producto y proveedor. La información es almacenada en un archivo plano llamados "costos.unl". Cada fila en este archivo contiene el número de catálogo, código del proveedor y costo unitario.

La data del cliente es almacenada en la Base de datos "'Ordenes", y aún no se ha añadido la información de Región es nuestra Base de datos.

Solución:
Área temática: Ventas
Identificar las dimensiones:

Del análisis del problema se determinan las siguientes dimensiones:
-Producto
-Cliente

-Proveedor
-Geografía
-Tiempo

Los elementos de las dimensiones y las jerarquías serían las siguientes:

Producto: Línea de producto → Producto

Cliente: Cliente

Proveedor: Proveedor

Geografía: Ciudad → Distrito

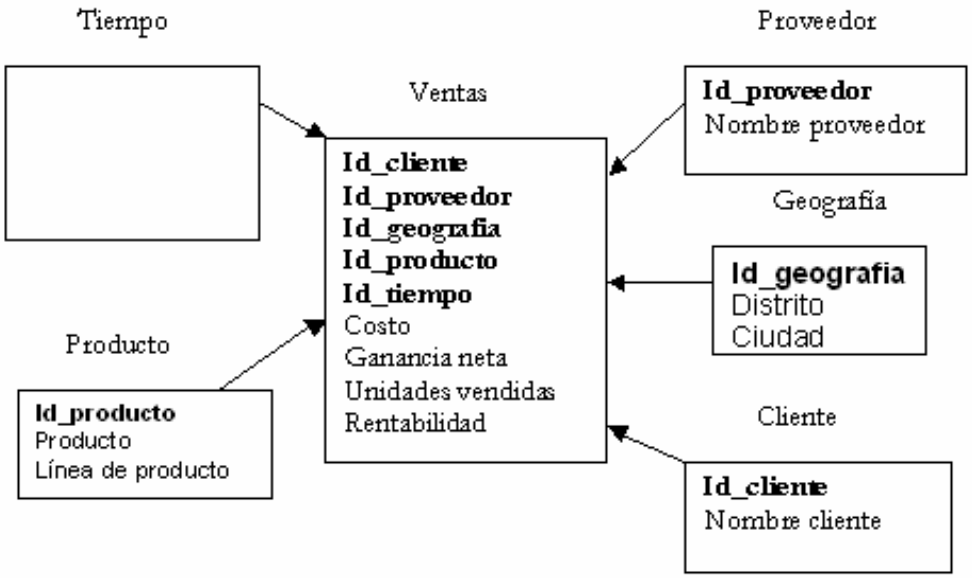
Tiempo: Año → Trimestre → Mes → Semana

Identificar las medidas:

Las medidas son las siguientes:

-Costo
-Rentabilidad
-Unidades vendidas
-Ganancia neta

El diagrama estrella



La granularidad de la tabla de hechos

La granularidad de la tabla de hechos está determinada por el producto de las granularidades de las dimensiones. Esto es:

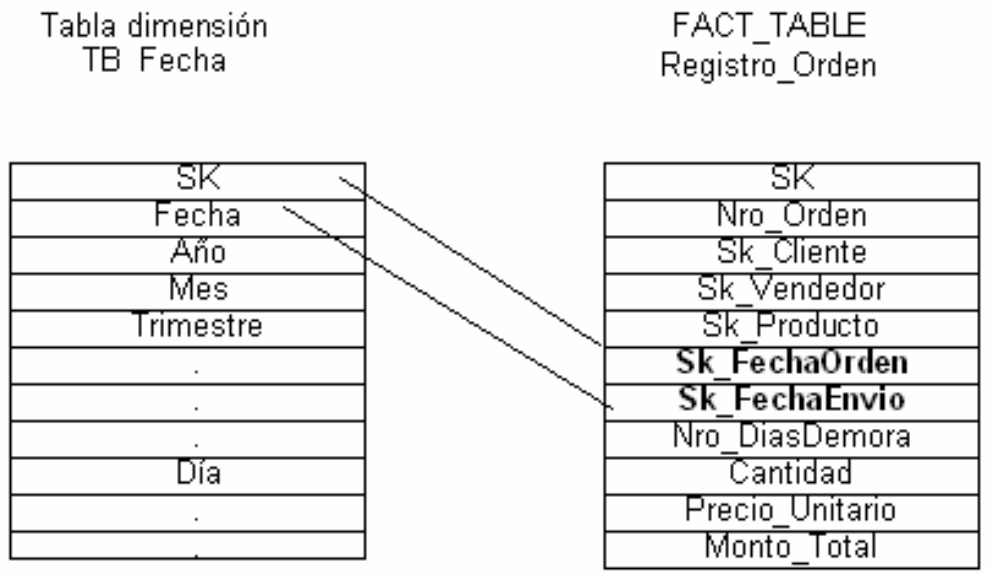
Producto x Cliente x Proveedor x Distrito x Semana

2. Casos a tener en cuenta

En esta parte de la sesión verá algunos casos importes en el modelo dimensional.

a. **Más de una tabla dimensión de fecha.-** Existen algunos modelos dimensionales donde se hace referencia a más de una fecha en el hecho.

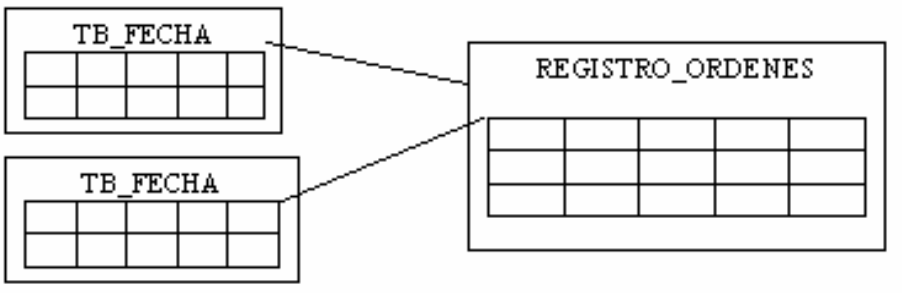
Ejemplo: "….. Una orden es solicitada en una fecha determinada y enviada días después…."



El modelo dimensional, aparentemente está correcto, pero al momento de implementarse tendríamos un problema al tener **solo una tabla Fecha** que se relaciona con **DOS o más Surrogate Key en la Fact Table.**

El surrogate Key es la llave principal en un tabla dimension.

En estos casos, se debe tener tantas tablas dimensión como Surrogate Key en la fact table, pero como es muy conveniente repetir todo el contenido de una tabla, usaremos una salida adecuada**.**



Debe crear la ilusión de 2 tablas independientes empleando vistas con SQL. Es importante crearlas con los nombres de columnas diferentes.

Create View vw_FechaOrden (**Sk_FechaOrden, FechaOrden, AñoOrden, Mesorden,……., DiaOrden**)

As

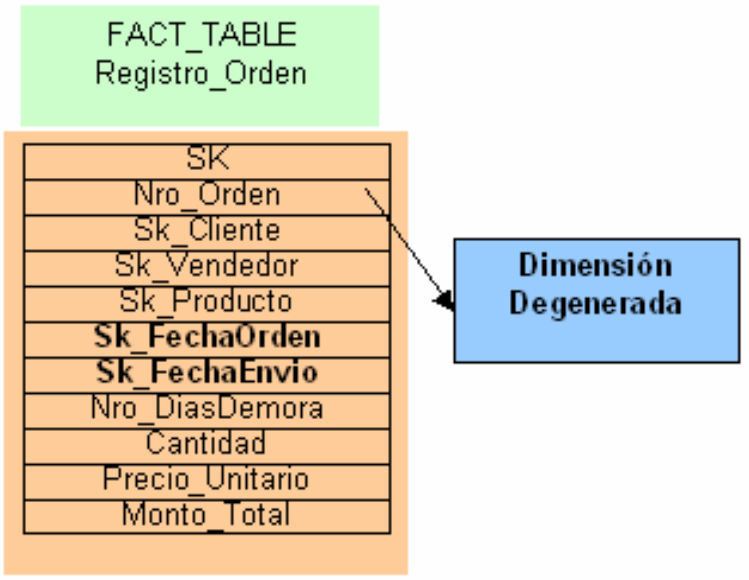
Select Sk, Fecha, Año, Mes,.. , Dia From Tb_Fecha

Create View vw_FechaOrden (**Sk_FechaEnvio, FechaEnvio, AñoEnvio, MesEnvio,……., DiaEnvio**)

As

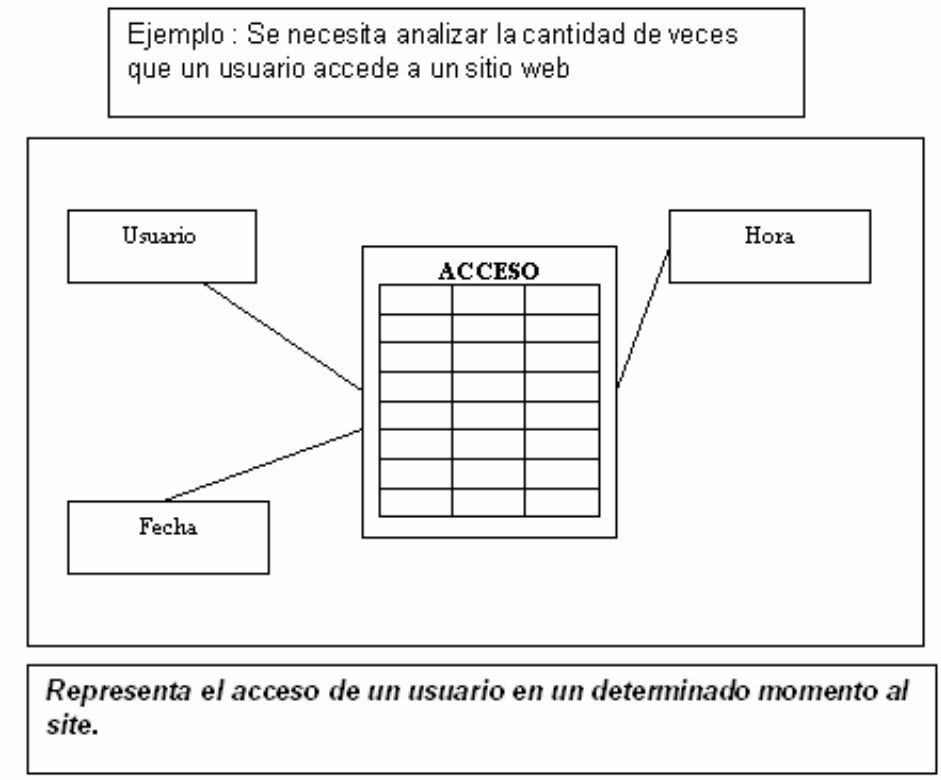
Select Sk, Fecha, Año, Mes,.. , Dia From Tb_Fecha

b. **Dimensiones Degeneradas.-** Son identificadores transaccionales. Son valores no numéricos que residen en la fact table. Tienenpor característica que si se convierten en tablas dimensión sólo tendrían dos columnas (Surrogate Key y el número de la Orden)



Es factible dejar en la fact table el valor completo del número de la orden , ya que si crea una tabla dimensión Numero_Orden, se tendría tantos registros como Ordenes se tuviera y solo 2 campos.

c. **Medidas en Fact Table.-** Mucha veces es difícil definir una medida para la fact table, en los casos donde el hecho de por sí solo no genera ninguna cantidad o valor para poder medir.



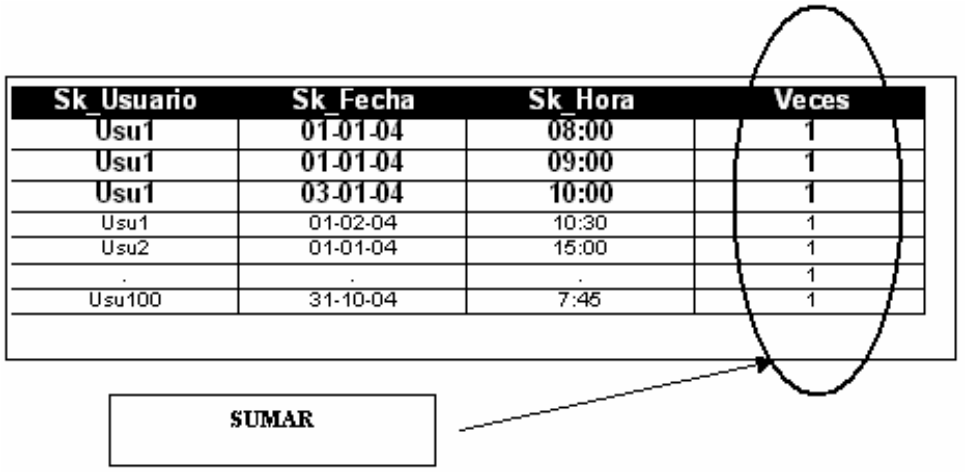
La fact table queda: (para evitar los surrogate keys hemos colocado los valores descriptivos)

Sk_Usuario	Sk_Fecha	Sk_Hora	Veces
Usu1	01-01-04	08:00	1
Usu1	01-01-04	09:00	1
Usu1	03-01-04	10:00	1
Usu1	01-02-04	10:30	1
Usu2	01-01-04	15:00	1
.	.	.	1
Usu100	31-10-04	7:45	1

La Fact table presenta 3 (Sk's) y una medida (Veces). La medida "veces", la implementara con la finalidad de poder operar sobre ella.

El valor de dicho campo siempre será "1"

Por ejemplo, si quisiéramos saber cuántas veces el Usu1 ha ingresado al site en enero de 2004.



d. **Hechos no realizados.-** Muchas veces se desea analizar qué hechos no sucedieron, aunque parezca contradictorio se puede preguntar los siguiente:

¿Qué productos no se han vendido?

Es factible tener una tabla de hechos, de hechos no realizados.

Autoevaluación

1. Una fábrica de alimentos balanceados, tiene 5 plantas de producción a nivel nacional, en cada planta hay cinco departamentos, y cada departamento tiene entre 2 y 4 líneas de producción que son administradas por un supervisor. Se ha entrevistado al gerente de personal y a un analista del área de desarrollo de personal, y se ha obtenido lo siguiente:

   Gerente de personal: "*Es prioridad de nuestra empresa analizar la productividad de nuestros empleados con la finalidad de optimizar nuestros programas de entrenamiento y la distribución de la carga de trabajo. Esto nos permitirá tener empleados motivados y mejorar el clima organizacional de la empresa.*"

   Analista del área de desarrollo de personal: "*Es necesario contar con información actualizada y que se pueda analizar con facilidad. En la actualidad es complicado obtener la información de nuestros empleados dispersa en cada una de nuestras plantas*. *Lo que necesitamos es comparar las horas de capacitación recibidas con las horas de trabajo en planta para nuestros trabajadores, por cada planta y por categoría de trabajador (cargo) Necesitamos también analizar la evolución del sueldo de nuestros trabajadores en los últimos 2 años, así como los ascensos que se hubieran producido*"

   Además se requiere los siguientes reportes:
   -Horas en Capacitación/Horas trabajas por planta, por área, por cargo, por empleado, por semana.
   -Número de tardanzas por planta, por área, por cargo, por mes.

2. Se desea diseñar un DataMart para un banco. El banco ofrece un portafolio de servicios que incluye, cuentas corrientes, ahorros, préstamos personales, hipotecarios, tarjetas de crédito y otros. El objetivo principal del banco es hacer campañas efectivas de marketing ofreciendo nuevos servicios a sus clientes que son personas y empresas.

   Se ha recopilado los siguientes requerimientos, de las entrevistas a los usuarios:
   - Se requiere ver 5 años de datos históricos para cada cuenta. Para todos los meses será suficiente ver el estado al final del mes y para el mes actual el estado al día de ayer, no es necesario tener otros días en el mes actual.
   - Cada tipo de cuenta tiene un saldo. Es necesario agrupar diferentes tipos de cuentas y comparar sus saldos. También se requiere contabilizar el número de transacciones.
   - Se necesita hacer análisis demográficos en base a los datos de los clientes.

Para recordar

1. Los pasos a seguir en el modelamiento dimensional, después de identificar el área temática son:
   - Entrevistar a todos los usuarios determinando sus requerimientos de información.
   - Verificar que toda la información requerida este sistematizada y se pueda obtener.
   - Definir las dimensiones y medidas.
   - Definir la granularidad, la frecuencia de actualización y graficar
   - Validar el modelo con el usuario.

SEMANA

10

Taller: Modelamiento Dimensional

OBJETIVOS ESPECÍFICOS

- Afianzar los conocimientos de modelamiento dimensional.

CONTENIDO

- Descripción de los pasos para el modelamiento dimensional – Casos propuestos.

1. Pasos de modelamiento dimensional.

En los casos que se plantearán en clase, es importante poder desarrollar el modelo dimensional siguiendo 4 pasos importantes.

•**Seleccionar el Proceso del Negocio**.- En este paso debemos definir claramente el proceso o procesos que estaremos modelando. Es importante poder comprender correctamente el proceso ya que es a partir de éste que todo nuestro modelo será construido.

Ejemplo: Pago de planillas.

•**Nivel de Granularidad.**- Otro punto importante por definir es el nivel de detalle el cual se desea analizar, es decir, el nivel de detalle disponible de los datos al usuario. Es importante poder definir el nivel mas atómico, es decir, lo mas detallado posible ya que nos permitirá conocer mas lo que estamos analizando

Ejemplo: El pago a un empleado en particular.

•**Dimensiones.-** Luego de haber definido el hecho en primera instancia, las dimensiones son más fáciles de poder determinar, ya que ellas describirán los actuantes en el hecho. Es importante el diseño de las tablas dimensión ya que es a partir de ellas que vamos a poder establecer diversos niveles de agrupamiento.

Ejemplo: Empleados, Oficinas, Tiempo, etc.

•**Facts Table (Hechos)**.- Habiendo definido el hecho y las tablas dimensión es importante definir las medidas a analizar que forman parte del hecho.

Ejemplo: Monto, bruto, monto, neto, etc.

ADAPTADO DE :
FUENTE: KIMBALL Ralph, ROSS Margy
2002. The Data Warehouse Toolkit – The complete guide to dimensional modeling. Editorial: John Wiley & Sons Inc. Ciudad: New York. Pag: 32-38

2. Casos propuestos en clase

Se expondrán diversos casos para realizar el modelamiento dimensional en clase, los cuales serán desarrollados por los alumnos y luego desarrollados por el profesor del aula, identificando los puntos señalados anteriormente.

SEMANA 11

Diseño de la base de datos de Data Warehouse

OBJETIVOS ESPECÍFICOS

- Identificar los criterios de diseño de la base de datos de Data warehouse.

CONTENIDO

- Las necesidades de recursos.
- Los recursos de hardware.
- El manejador de base de datos.
- Particionamiento o fragmentación de tablas
- Consideraciones adicionales

ACTIVIDADES

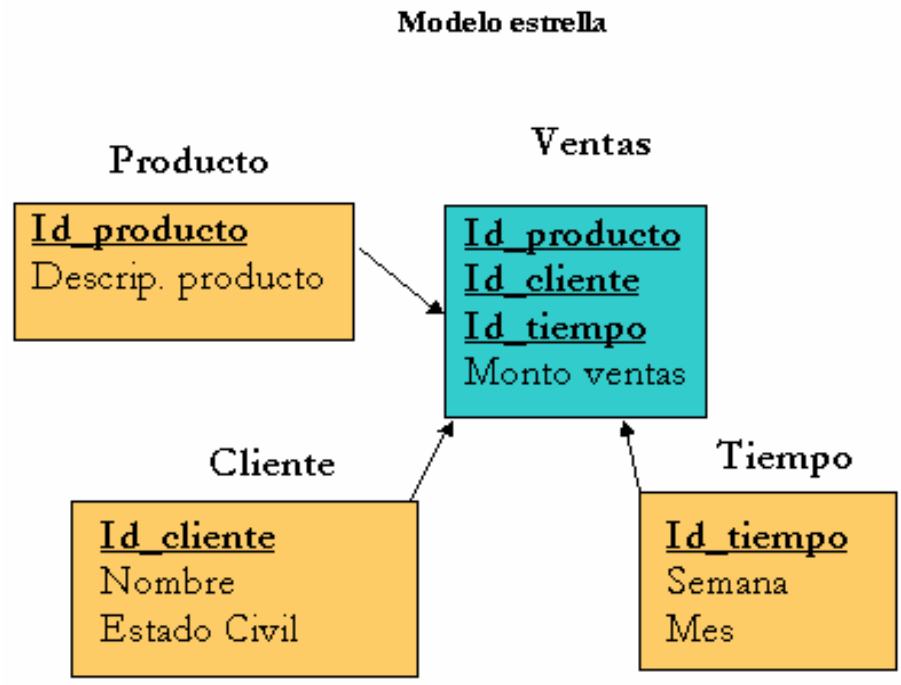
- Contestar las preguntas del cuestionario de auto evaluación.
- Evaluación Continua

1. Las necesidades de recursos

Para determinar las necesidades de recurso de hardware que tiene un Data warehouse se debe analizar el tamaño de la base de datos que va a construir.

En particular, el crecimiento de un modelo dimensional se puede complicar debido a las necesidades de almacenamiento.

Con bases de datos grandes se necesita mantener una performance aceptable, a continuación se analiza las necesidades de recursos del modelo estrella.



En la mayoría de los casos, las dimensiones son tablas que contienen un número relativamente pequeño de registros, hay dimensiones relativamente grandes como la dimensión cliente en el caso que se esta revisando.

Id_cliente	Nombre	Estado Civil
1	Juan Pérez	S
2	Sonia Salazar	C
3	Sofía Alvarez	D
..	..	...

La Fact-table o tabla de hechos tiene muchos registros, en el peor caso tendría tanto registros como combinaciones posibles de dimensiones haya, es decir en el

caso de la figura tendría cantidad de productos x cantidad de clientes x el número de registros de la tabla tiempo. Esta característica hace que la Fact-table sea una tabla que puede crecer mucho por lo que hay que contemplar el espacio en disco suficiente para almacenarla.

Id_prod	Id_tiempo	Id_cliente	Venta
1	2	7	100.00
4	89	32	200.00
8	3	3245	234.50
7	67	784	657.00
..	..	..	..

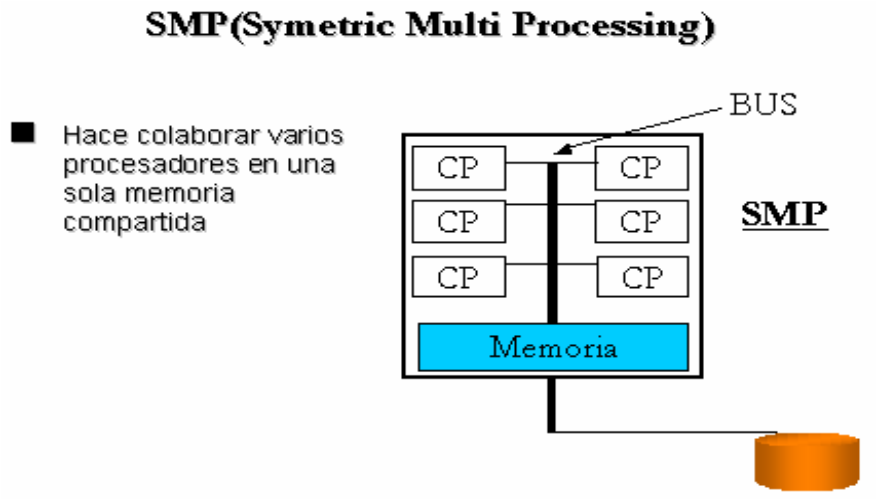
2. Los recursos de hardware

Con la finalidad de aprovechar mejor los recursos de hardware, es necesario revisar las diferentes arquitecturas de los servidores que se pueden utilizar en Data warehouse.

Arquitecturas de hardware

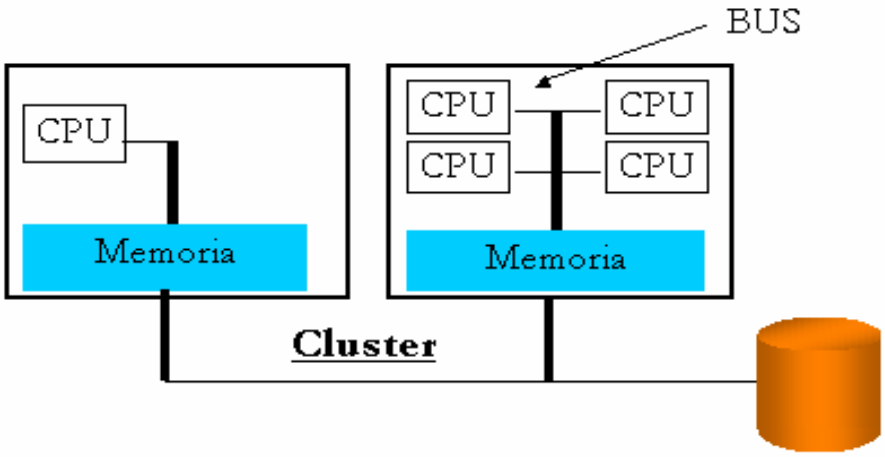
- La característica fundamental a explotar en un servidor es el paralelismo.
- Al ejecutar "n" tratamientos sobre "n" procesadores, los tiempos de respuesta serán idénticos.
- Se puede dividir un tratamiento haciendo colaborar varios procesadores, disminuyendo el tiempo de respuesta.

Las arquitecturas que se encuentran en el mercado son las siguientes:



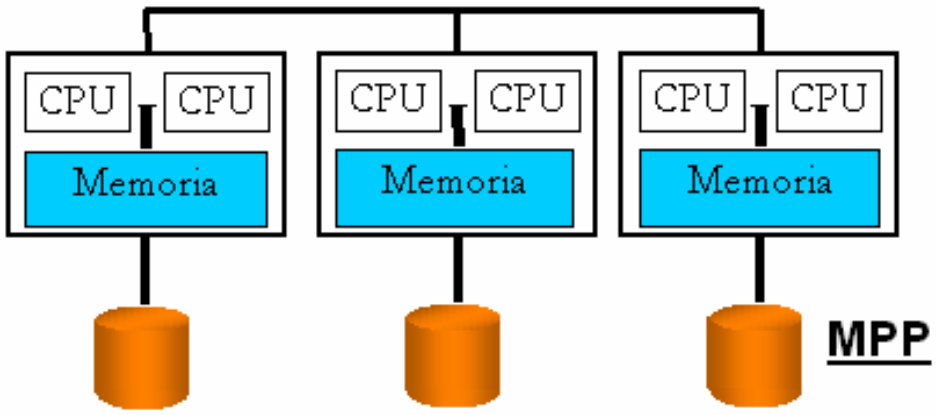
Cluster

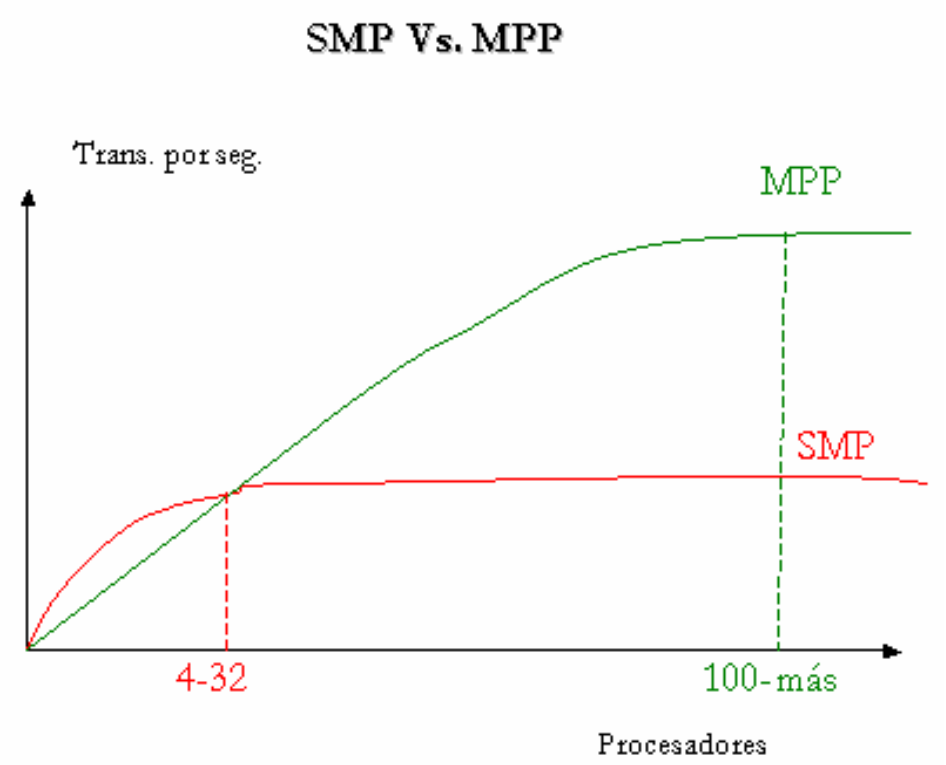
- Arquitectura que permite la colaboración de varias máquinas compartiendo discos



MPP(Massively Parallel Processing)

- Hace colaborar varios procesadores, cada uno con su propia memoria.



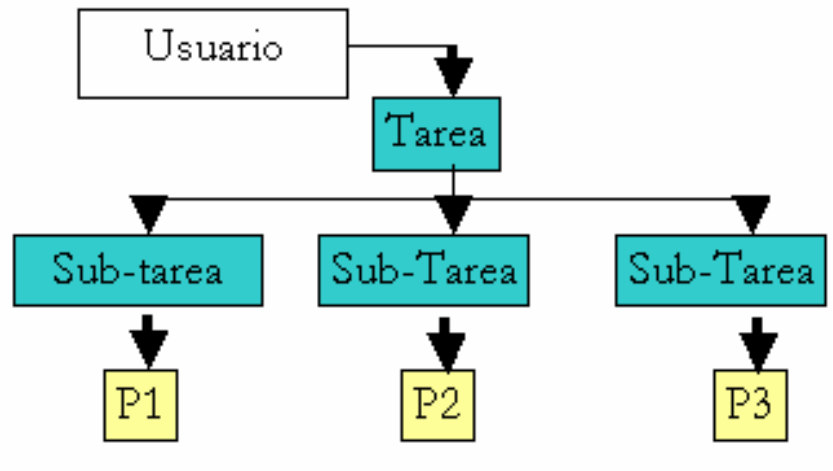


El siguiente gráfico muestra una comparación entre la arquitectura MPP y la SMP, como se puede ver en el gráfico un computador de menos de 32 procesadores se desempeña mejor cuando es de una arquitectura SMP y cuando se trata de tener más procesadores la arquitectura MPP tiene un mejor desempeño.

3. El manejador de base de datos

- Los Manejadores de Bases de Datos deben utilizar los recursos de hardware al máximo.
- Deben hacer dos tipos de operaciones:
  - Consultas complejas (Volúmenes grandes).
  - Cálculos complejos (batch).
  - Carga (lectura y actualización)

¡Error!

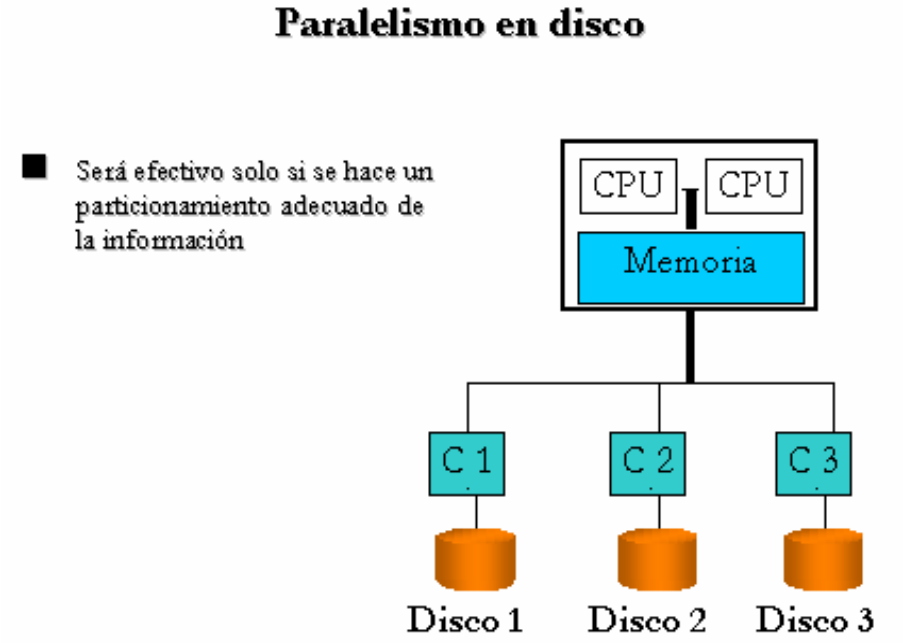


El motor de base de datos se debe afinar con la finalidad de obtener el mejor rendimiento tanto en los procesos de carga del Data warehouse así como en los procesos de explotación de la información.

Paralelismo en disco

- Al acceso al disco es el principal problema de un Manejador de Base de Datos.
- El paralelismo en disco, se puede implementar incluso en máquinas que tengan un solo procesador, incrementando el número de controladores de disco.

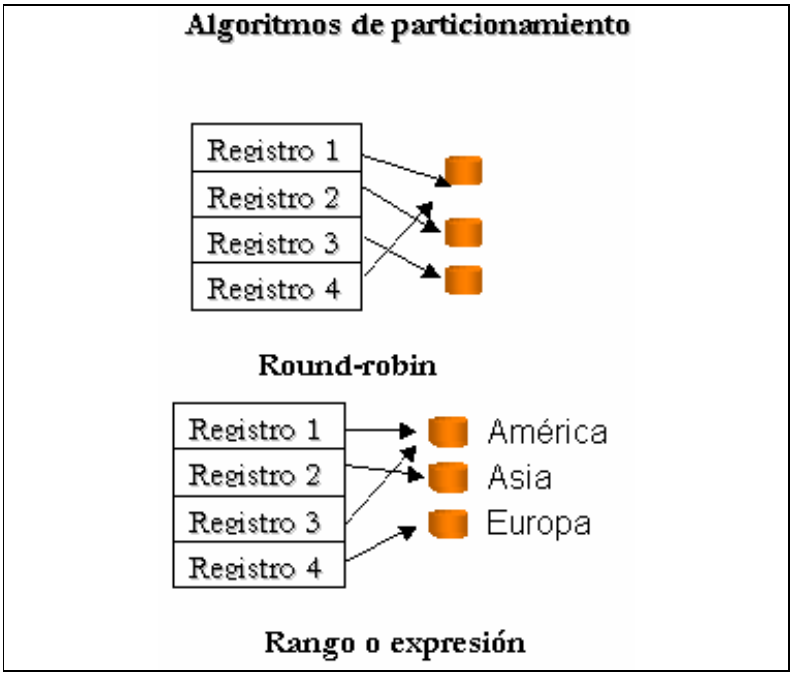
¡Error!



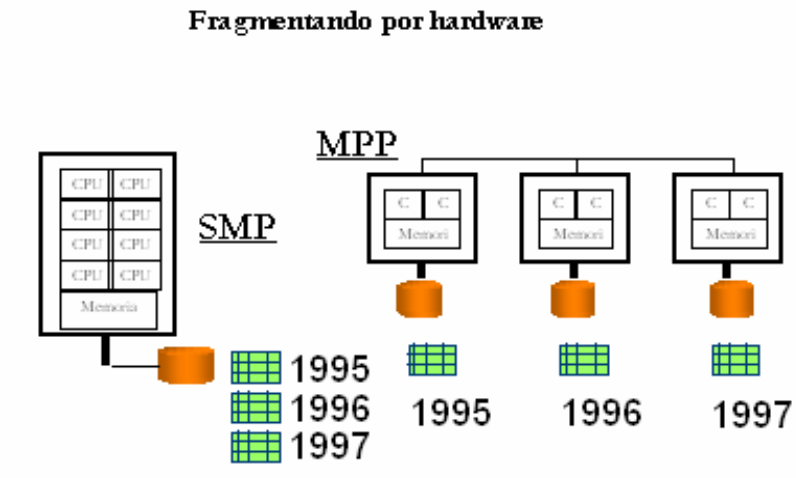
4. Particionamiento o fragmentación de tablas.

- Es el proceso de dividir una tabla en unidades más pequeñas.
- Ventajas.
  - Mejora en el tiempo de respuesta de los querys.
  - El proceso de backup y de recuperación es incremental y se acelera.
  - Disminuye el tiempo requerido para la carga en tablas indexadas.
- El particionamiento no es gratuito.
  - Se requiere querys más “inteligentes” para determinar en que partición esta la data consultada.
  - Se requiere metadata adicional para saber en que partición esta la data.

¡Error!

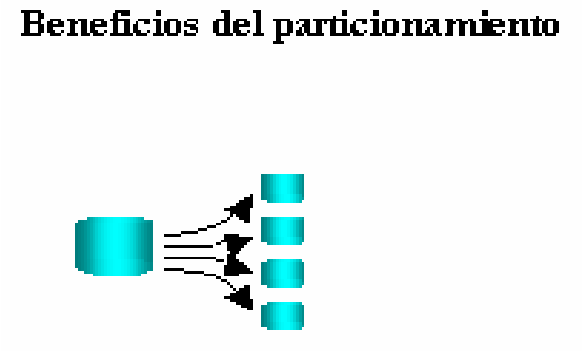


¡Error!



La estrategia de fragmentación depende fundamentalmente de los querys a realizar y la arquitectura del hardware.

¡Error!

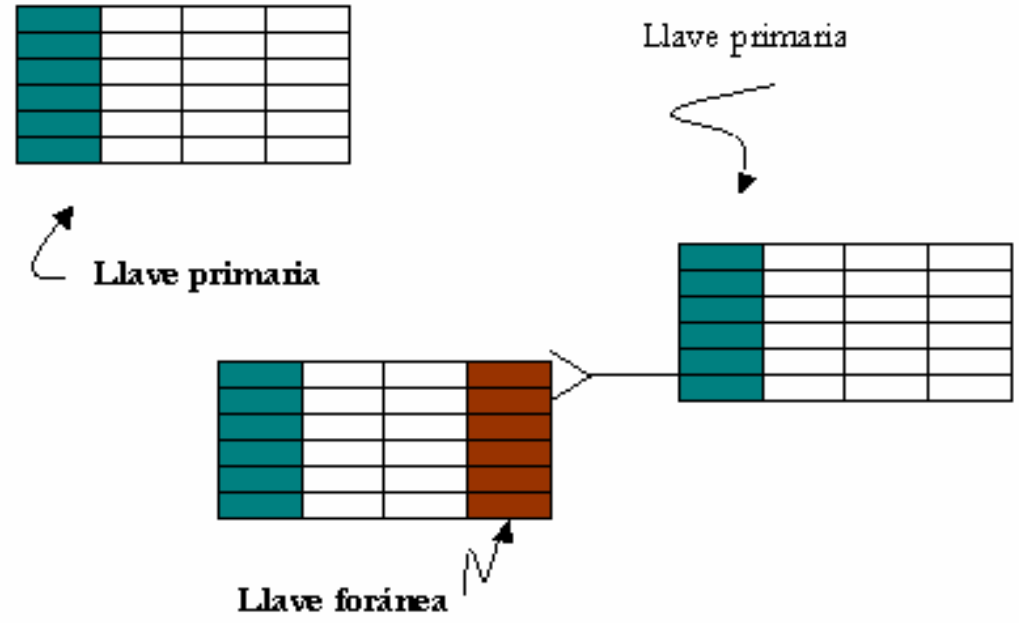


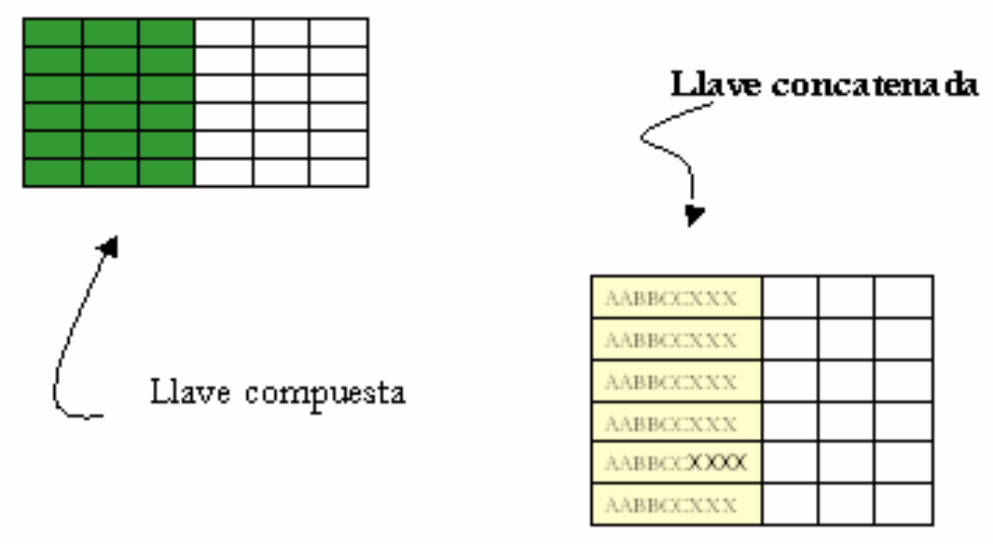
- Incrementa el paralelismo.
- Reduce tiempos de Backup.
- Incrementa la disponibilidad.
- Mejora la administración.
- Reduce los conjuntos de datos para los querys.
- Facilita eliminar los datos antiguos.

5. Consideraciones adicionales

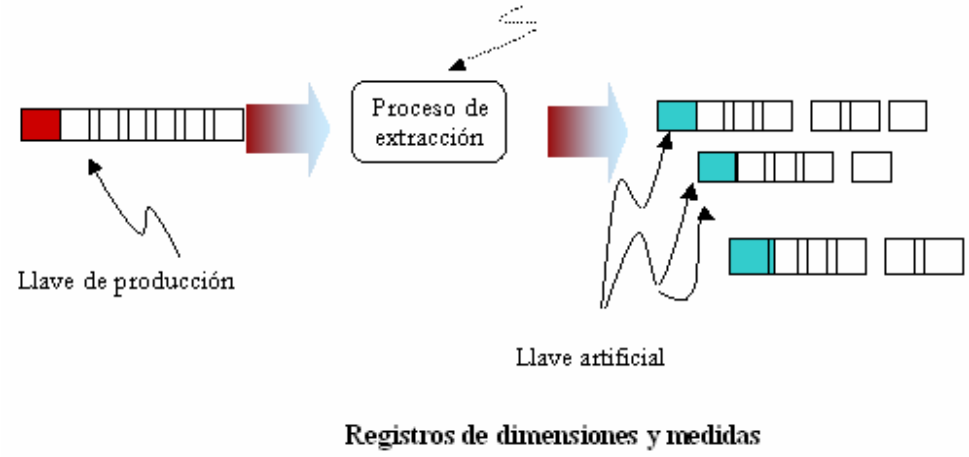
Las principales tipos de llaves en una base de datos operacional son las siguientes:

- Llave primaria, identificador único de un registro
- Llave foránea, garantiza la integridad referencial
- Llave compuesta, constituida por varias columnas.
- Llave concatenada, una sola columna con valores concatenados.





Cuando se construye un DataMart, las llaves del sistema operacional deben ser transformadas a llaves artificiales. El proceso de extracción reemplaza la llave del sistema operacional por una llave artificial.



Características de la llave artificial

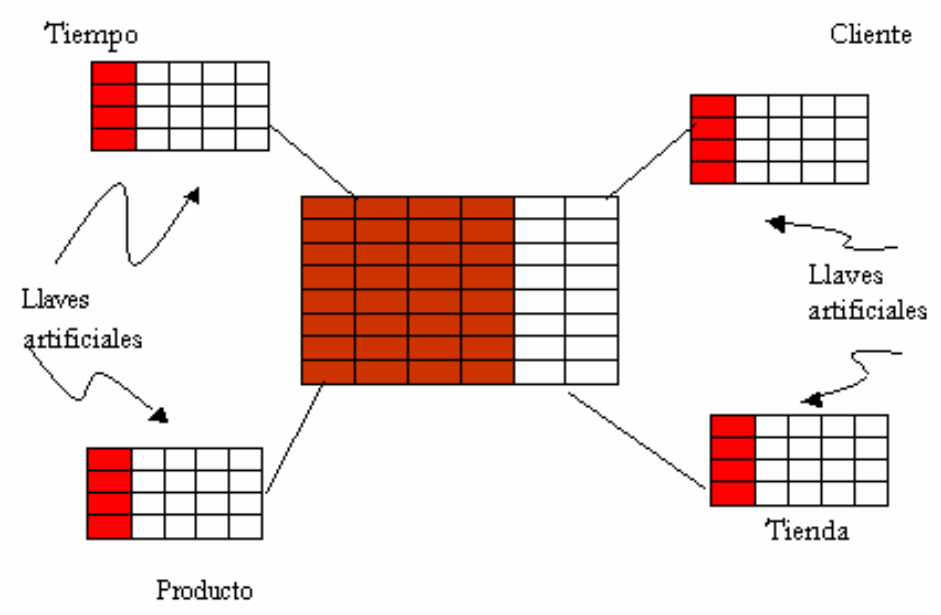
- Sustituye a la llave del sistema operacional.
- Es de tipo “integer” y es un secuencial que se inicia en 1.

Razones para el uso de llaves artificiales

- Las llaves del sistema de producción pueden ser reutilizadas.
- Permite manejar los cambios que se puedan producir en el sistema operacional, en los campos descriptivos.
- Permite el manejo de los cambios en la misma llave.

El esquema estrella y las llaves artificiales

En el modelo estrella se deben utilizar llaves artificiales, pues mejorar la performance de los querys, facilitan el manejo del cambio y permiten reutilizar las llaves de los sistemas operacionales. Asi mismo permiten reducir el espacio en disco utilizado por la tabla de hechos, y es obligatorio utilizarlas cuando se va a integrar fuentes distintas



Constraints

- Check constraint: Define el dominio de un campo determinado.
- Not-Null constraint: Excluye a los nulos del dominio de un campo.
- Primary Key constraint: Asegura que la tabla no tendrá duplicados.
- Foreing Key constraint: Asegura la integridad entre tablas.

Índices

- Son estructuras de datos utilizadas por el DBMS para acelerar el acceso a los datos.
- Razones para usarlos:
  •Mejoran la performance.
  •Garantizan la unicidad de campos, por ejemplo en la llave primaria.

Autoevaluación

1. ¿Que criterios se deben tener en cuenta en el diseño de la base de datos de Data Warehouse?

2. Describa las arquitecturas de servidores más conocidas

3. ¿Cómo se calcula el número máximo de registros que puede tener una Fact-table y el espacio requerido en disco?

4. ¿Cuáles son las estrategias de fragmentación de tablas, describa cada una de ellas?

5. ¿Cuáles son los criterios que definen la forma de fragmentación de las tablas?

6. Analice la curva de rendimiento comparativo entre la arquitecturas SMP y MPP, ¿que puede concluir?

7. ¿Cuales son las razones para el uso de la llave artificial?

Para recordar

1. Un paso fundamental en la implementación de un data warehouse es el dimensionamiento de la base de datos.

2. Es necesario conocer la arquitectura del hardware para poder definir una buena estrategia de fragmentación de tablas.

3. Las estrategias de fragmentación son función de los querys y de la arquitectura del hardware.

4. Se debe utilizar el paralelismo para optimizar los procesos de lectura y escritura de la Base de datos.

5. Todas las bases de datos se deben afinar con la finalidad de aprovechar al máximo los recursos de hardware.

SEMANA

12

Poblando el data warehouse: extracción, transformación y carga - estandarización y limpieza de datos

OBJETIVOS ESPECÍFICOS

- Identificar las características del proceso de poblamiento del data warehouse.
- Comprender los problemas y las soluciones de estandarización y limpieza de datos

CONTENIDO

- El “Staging área”.
- Poblamiento del Data Warehouse.
- Estandarización y limpieza de datos.
- La importancia de los metadatos.

ACTIVIDADES

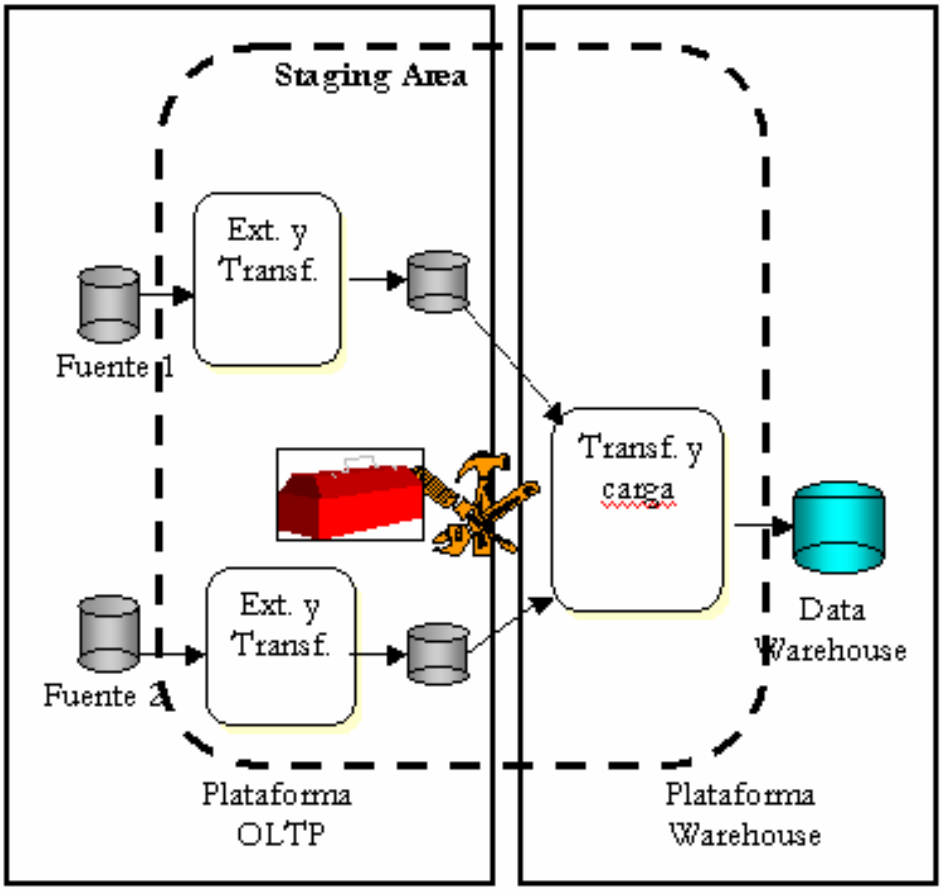
- Contestar las preguntas del cuestionario de autoevaluación.

6. El “Staging área”

El “Staging área” es el lugar de tránsito de los datos en su camino de la fuente al Data Warehouse. La mayor parte del esfuerzo en la construcción de un Data Warehouse se despliega en el “Staging Área”, donde se construyen y se implementan los procesos de extracción, limpieza, transporte, transformación y carga de los datos.

Normalmente el Data Warehouse y los sistemas transaccionales residen en plataformas de bases de datos distintas debido a que las configuraciones que tienen ambos ambientes son muy diferentes, y con la finalidad que los procesos de Soporte a decisiones, que normalmente son pesados, no afecten a los sistemas operacionales.

La herramienta que se utiliza para la construcción de los procesos del “Staging área” es la herramienta ETL, que es una herramienta especializada en el tratamiento de los datos, sobre todo en el manejo de volúmenes grandes.



7. Poblamiento del Data Warehouse

El proceso de poblar un Data Warehouse se puede dividir en 5 tipos de subprocesos:

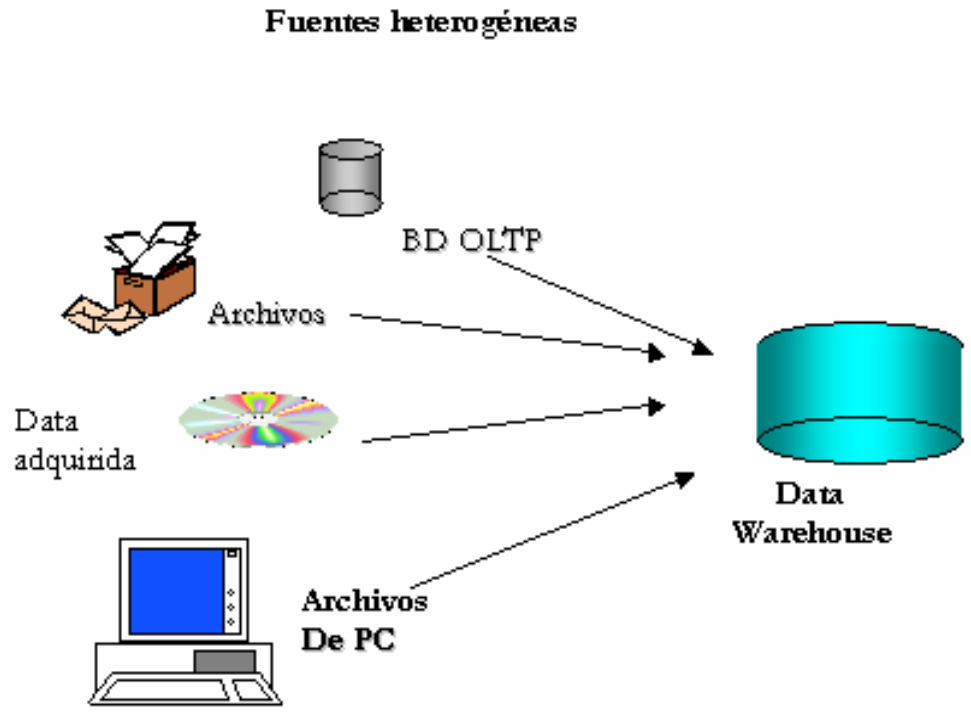
- Descubrir
- Extraer
- Transformar
- Transportar
- Cargar

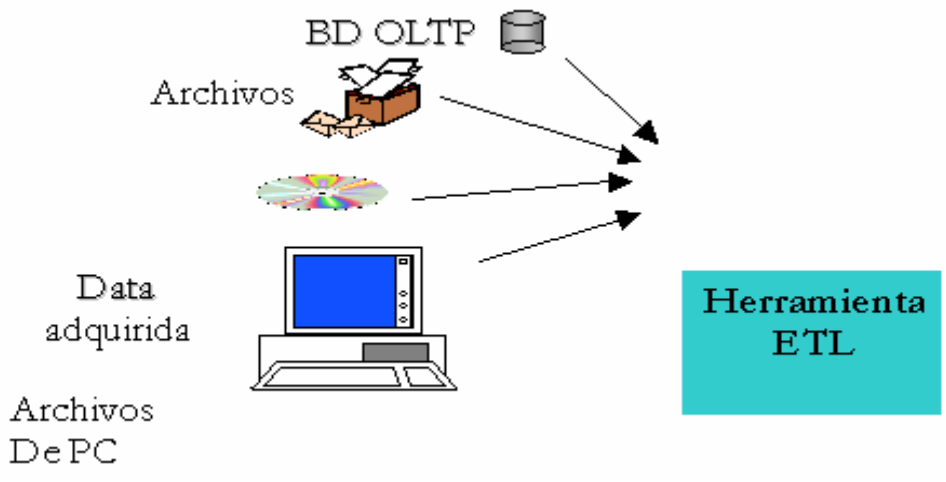
Descubrir

En esta etapa se analiza la fuente de información, seleccionando los datos a extraer, los niveles de calidad de estos y la disponibilidad de los mismos.

Extraer

El proceso de extracción se realizara sobre fuentes heterogéneas, es por ello que se debe contar con una herramienta ETL abierta a todas las fuentes y a todas las plataformas.

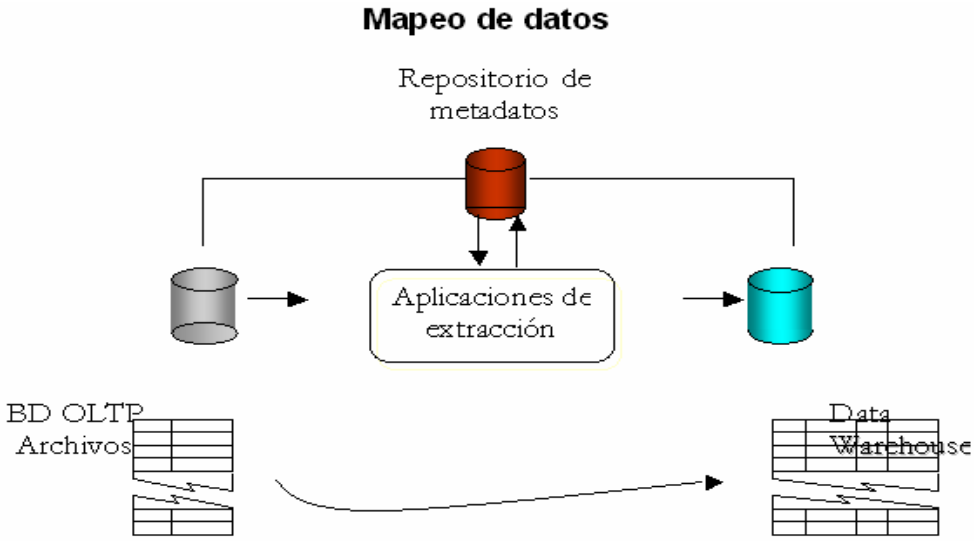




Transformar

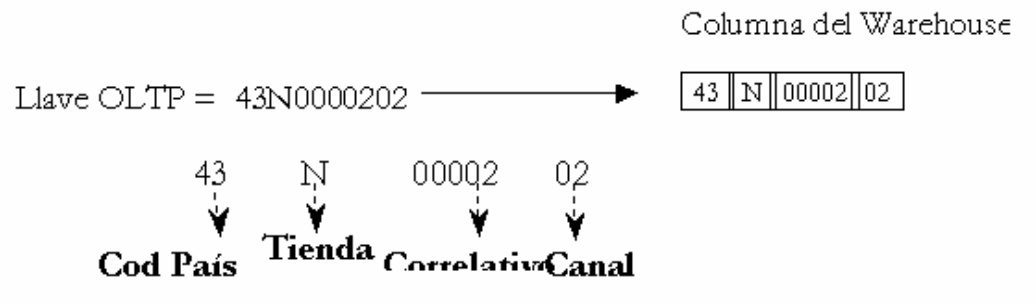
El proceso de transformación se encarga de cambiar los formatos de datos del sistema fuente al sistema destino, así como de realizar la integración de las fuentes y la estandarización de los datos

El componente mas importante de los procesos de transformación es el mapeo de los datos, que es la base de las definiciones de las reglas de transformación, constituye la fuente más importante de metadatos y es la base sobre la cual se manejan los cambios.



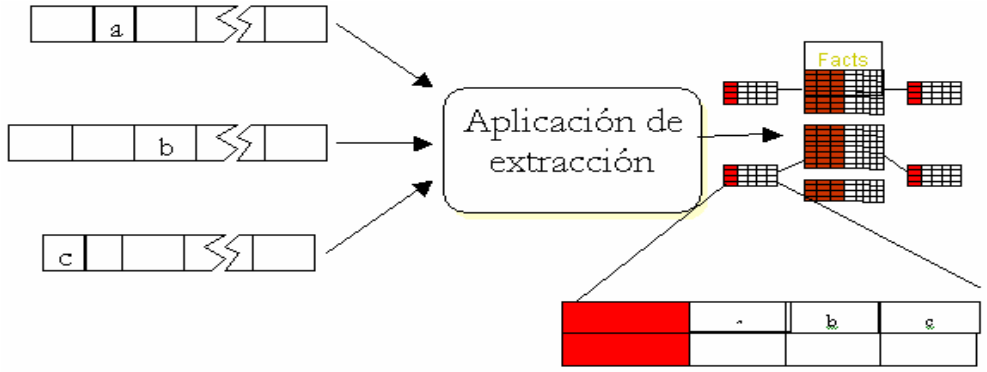
Los procesos de transformación serán muy variados y dependerán de las reglas del negocio, entre los tipos más importantes se pueden distinguir los siguientes:

- Conversión de llaves concatenadas



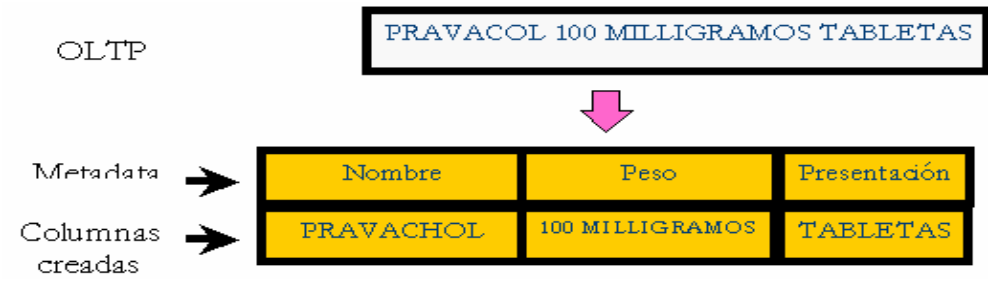
-Consolidación de datos

¡Error!

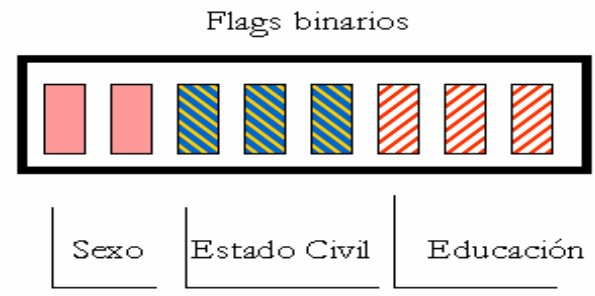


- Separación de campos “Free-Form”

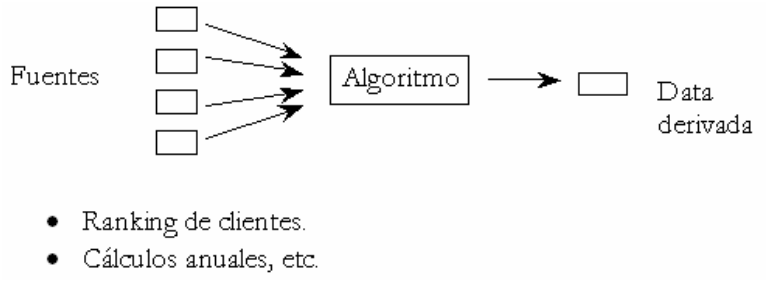
¡Error!



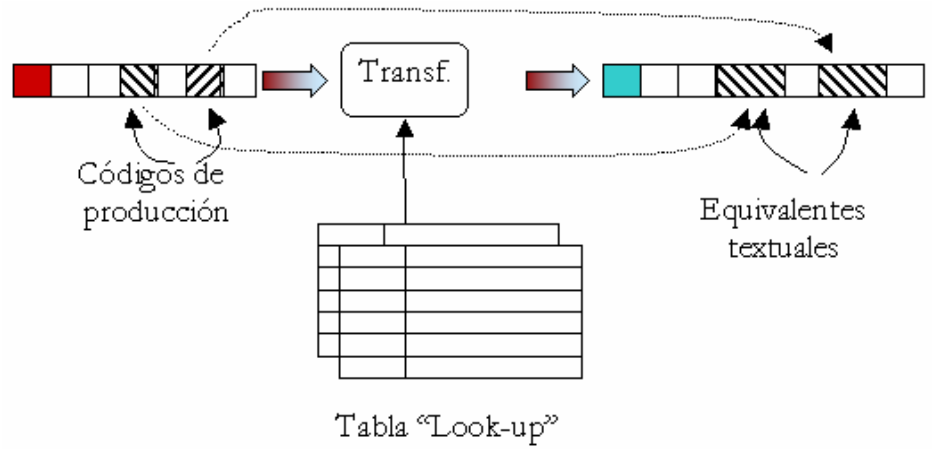
- Separación de datos que tienen codificación binaria



- **Derivar datos a partir de las fuentes**



- **Transformando códigos de producción**



- **Asignación de llaves artificiales**

¡Error!

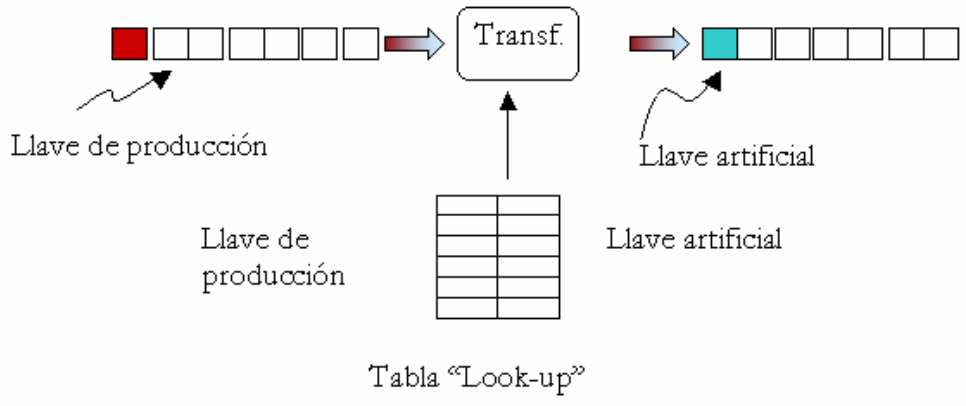


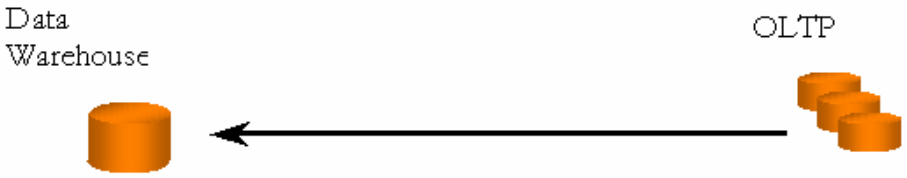
Tabla "Look up"

Es una tabla de referencia que básicamente tiene dos columnas que contienen las equivalencias entre los códigos de las fuentes y los códigos de Data Warehouse

Especificaciones del proceso ETL

El proceso ETL se especifica en una tabla similar a la de la figura siguiente que incluye el mapeo de la fuente al destino y en la que se incluye las reglas de transformación a implementar.

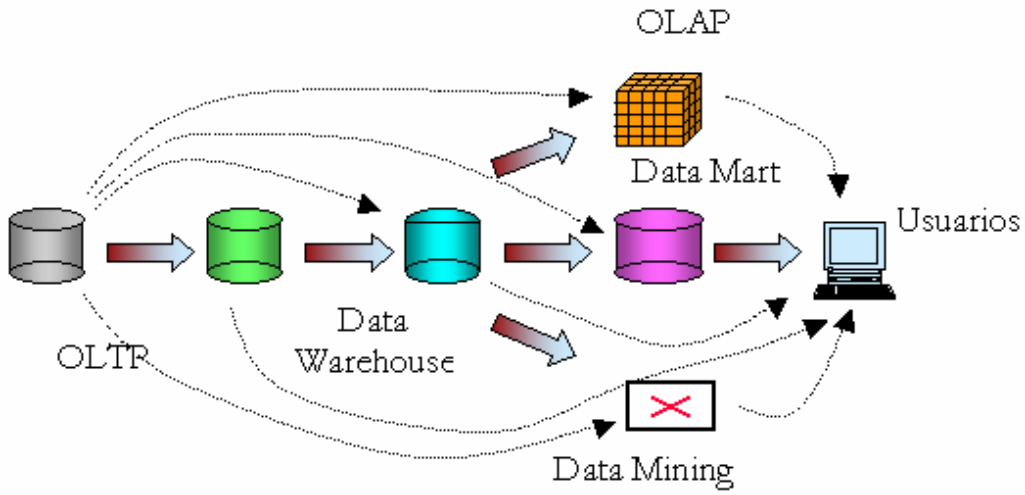
¡Error!



Atributo DWH	Transformación	DBMS-OLTP	
		DBMS/ Archivo	Atributo

Fuentes y objetivos

¡Error!



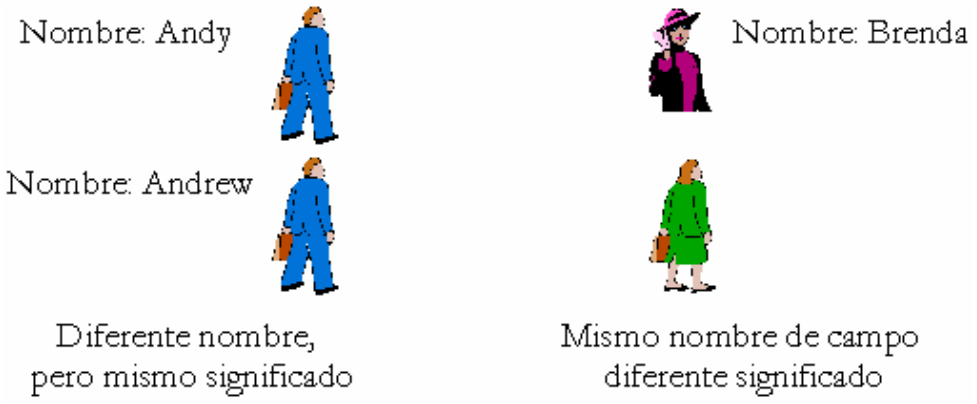
En general existirán procesos ETL entre la fuente y el warehouse o entre el warehouse y los datamarts o entre el ODS y los modelos de minería de datos, o todas las combinaciones posibles como se muestra en la figura anterior

8. Estandarización y limpieza de datos

Estandarización de datos

Es el proceso orientado a la uniformizar los datos en base a las definiciones y luego en base a la realidad. Por ejemplo un caso típico es el que se presenta es cuando existen campos que contienen diferente valor como por ejemplo “Andy” y “Andrew” y que se refieren a la misma persona real, o el caso en el que el campo contiene los mismos valores “Brenda” y “Brenda” pero que en la realidad corresponden a personas distintas.

Este problema se puede resolver con dos tipos de procesos que son complementarios. El primero un proceso automático que tenga rutinas que permitan identificar automáticamente estos registros y el segundo un proceso de gestión visual que tenga como finalidad complementar al primero.



Limpieza de datos.

El problema de la calidad de los datos se puede enfrentar en parte con rutinas de limpieza que permitan reducir el número de registros con error.

En el siguiente cuadro se muestra un caso típico de una tabla en la que se registran el número de documento y el nombre digitados y en la que se puede distinguir errores de digitación comunes.

Al igual que en el caso anterior el problema se puede enfrentar con rutinas que permitan identificar estos registros de manera automática complementadas por procesos de gestión visual de la información.

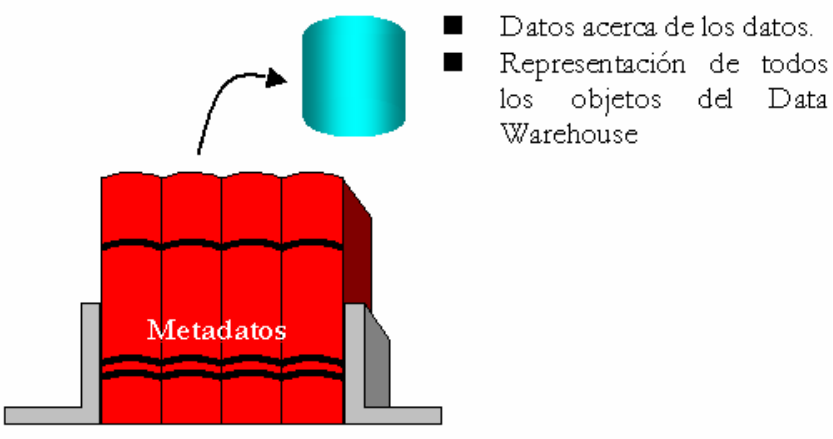
No. Doc.	Nombre
02336589	Juan Pérez Costa
2336589	Pérez Costa, Juan
02336689	Juan Pérez Costa

Posibles soluciones

- Construir rutinas de limpieza y transformación.
- Comprar herramientas especializadas en el tratamiento de nombres.
- Establecer procesos de gestión visual de la información.

9. La importancia de los metadatos

¿Qué son los Metadatos?
¡Error!



Importancia de los metadatos

- Los metadatos son como las fichas de catálogo de una biblioteca que ayudan a saber el contenido y la ubicación de un libro.

Importancia de los metadatos en el desarrollo del Data Warehouse

Cada etapa en la construcción del Data Warehouse genera un conjunto de metadatos propios, estos metadatos se deben unificar en un solo repositorio.

Los metadatos que se generan en cada etapa son:

- En la extracción de las fuentes:
  - Identificación de campos fuente.
  - Registro de cambios.
  - Resolución de inconsistencias.
  - Mapas
  - Transformaciones.

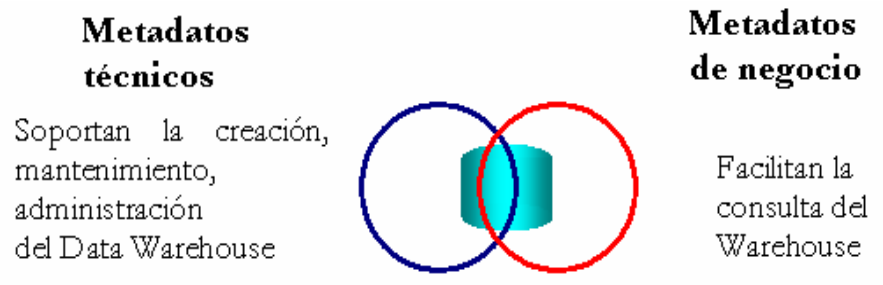
- En el Staging Área:
  - Integración y segmentación.
  - Resúmenes, adiciones.
  - Cálculos previos y derivaciones.
  - Transformaciones.

- En el bloque de Acceso y uso:
  - Proporciona un mapa de navegación para la exploración de la información.
  - Las herramientas de explotación generan metadatos propios.

Tipos de Metadatos por los usuarios que los utilizan:

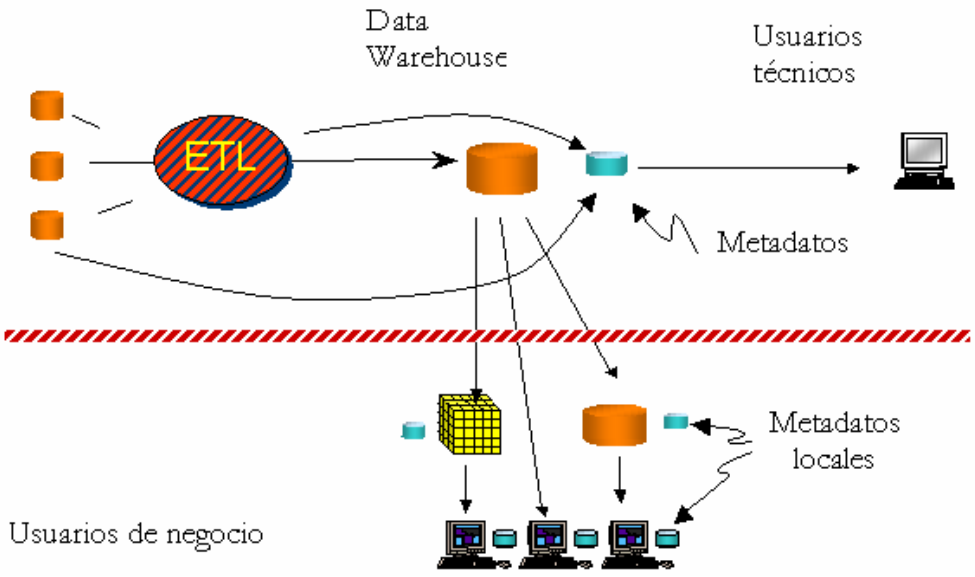
De acuerdo al tipo de usuario los metadatos pueden ser: Metadatos técnicos y metadatos del negocio

¡Error!



Arquitectura de metadatos

La implementación de un Data Warehouse requiere también la implementación de un repositorio unificado de Metadatos, este repositorio recibirá los metadatos que se generan en todas los bloques de la arquitectura del Data Warehouse.



¡Error!

Los usuarios de los metadatos tendrán necesidad de ver los siguientes metadatos:

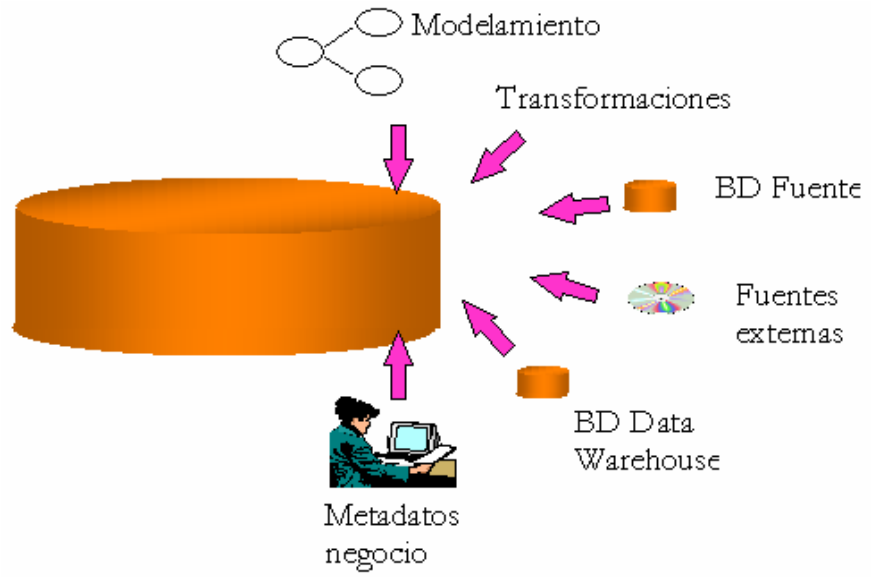
- Usuarios técnicos:
  - Datos sobre el proceso ETL
  - Datos sobre el DBMS.
  - Archivos, arquitectura.
  - Modelos físico, lógico.
  - Mapeos.
- Usuarios de negocio:
  - Áreas de negocio.

- Definiciones de reglas de negocio.
- Como utilizar las herramientas.
- Significado de la información
- Ubicación de la información

Fuentes de metadatos

Como se puede apreciar en la arquitectura de metadatos, las fuentes de estos últimos son diversas y al igual que las fuentes de datos de un datawarehouse requieren procedimientos ETL que lean los metadatos locales de cada herramienta y los centralicen en un solo repositorio unificado

¡Error!



Las fuentes de metadatos son:

- Lógica de programas.
- Comentarios en los programas
- Comentarios en archivos de datos.
- Secuencias de jobs y sus comentarios.
- Metadatos del repositorio de la herramienta CASE.
- Modelos de datos.
- Diccionarios de la base de datos.
- Documentos que contengan reglas de negocio

Autoevaluación

1. Defina “Staging área”

2. ¿Cuál es el proceso más costoso en Data Warehouse?

3. ¿Cuáles son las características principales de una herramienta ETL?

4. ¿Qué es una tabla “Look-Up”?
5. ¿Cómo sería un proceso típico de asignación de llaves artificiales?
6. Describa los procesos de agregación y consolidación de datos con ejemplos
7. Describa el problema de la estandarización de datos con ejemplos y proponga soluciones
8. Describa el problema de limpieza de datos con ejemplos y proponga soluciones
9. ¿Por qué es importante implementar el repositorio de metadatos?
10. ¿Cuáles son los tipos de metadatos?
11. ¿Quiénes son los usuarios de los metadatos?
12. Describa las fuentes de metadatos, ¿es posible tener acceso a todas?

Para recordar

1. El “Staging area” es el área más importante de un Data Warehouse, en ella se concentran la mayor parte de los recursos cuando se construye un Data Warehouse.
2. Los procesos ETL son los que permitirán construir el Data Warehouse
3. La herramienta ETL debe ser abierta a todos los tipos de datos y todas las plataformas como sea posible.
4. Las tablas “Look-up” son componentes esenciales de los procesos de transformación
5. El problema de estandarización de nombres es común a la mayoría de implementaciones de Data Warehouse.
6. La estandarización y la limpieza de datos esta relacionada directamente con la calidad de los datos y se apoya en procesos automáticos y de gestión visual.
7. Los usuarios de los metadatos son básicamente técnicos y de negocio.
8. Los metadatos del negocio deben ser presentados adecuadamente y son los que permiten al usuario explorar la información.

SEMANA

13

Poblando el data warehouse, primera carga y procesos de actualización - El acceso a los datos.

OBJETIVOS ESPECÍFICOS

- Comprender los procesos de primera carga y actualización.
- Presentar una metodología de construcción de un Data Warehouse.
- Identificar las formas de acceso a los datos.

CONTENIDO

- Los procesos de primera carga.
- Los procesos de actualización del warehouse.
- Atributos de la calidad de datos.
- Anomalías en la data operacional.
- Finalidad del Datawarehouse
- La fábrica de información.
- Técnicas de acceso a los datos

ACTIVIDADES

- Contestar las preguntas del cuestionario de autoevaluación.

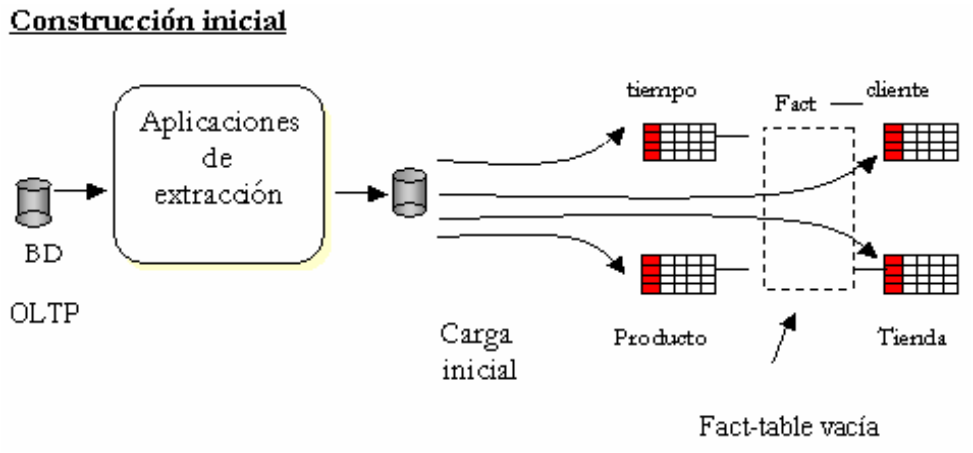
10. Los procesos de primera carga

Los procesos de carga al Data Warehouse tienen características muy particulares lo que hace necesario, en muchos casos, implementar procesos exclusivamente para la primera carga y procesos diferentes para los refrescos periódicos.

El aspecto más importante en un proceso de primera carga es el volumen de información que se va a llevar desde la fuente hasta el Data Warehosue, en la primera carga se suele llevar toda la información histórica lo que hace que los volúmenes de datos a extraer, procesar, transportar y cargar sean muy grandes. En algunos casos particulares cuando el volumen de información no sea muy grande se puede utilizar el mismo proceso para ambos fines

Muchas veces el proceso de carga de volúmenes considerables de datos puede tomar mucho tiempo por lo que se deberá utilizar herramientas especializadas en el tratamiento masivo de información como los “Bulk Loaders” para la descarga y carga de información en tablas y los algoritmos “hash” cuando se trate de Tablas “Look-up” en los procesos de transformación.

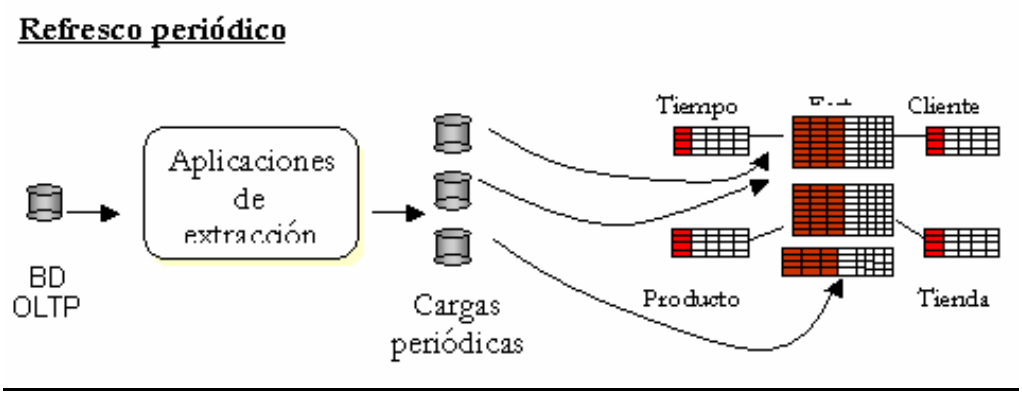
En el caso particular de un modelo estrella, en el proceso de primera carga se deben cargar las dimensiones, como se ve en la figura siguiente, y la parte histórica de la tabla de hechos, en ese orden necesariamente, debido a que después de cargadas las dimensiones se podrán generar las tablas “Look-up” que permitirán hacer las asignaciones de las llaves artificiales de la Fact-table.



11. Los procesos de actualización del warehouse

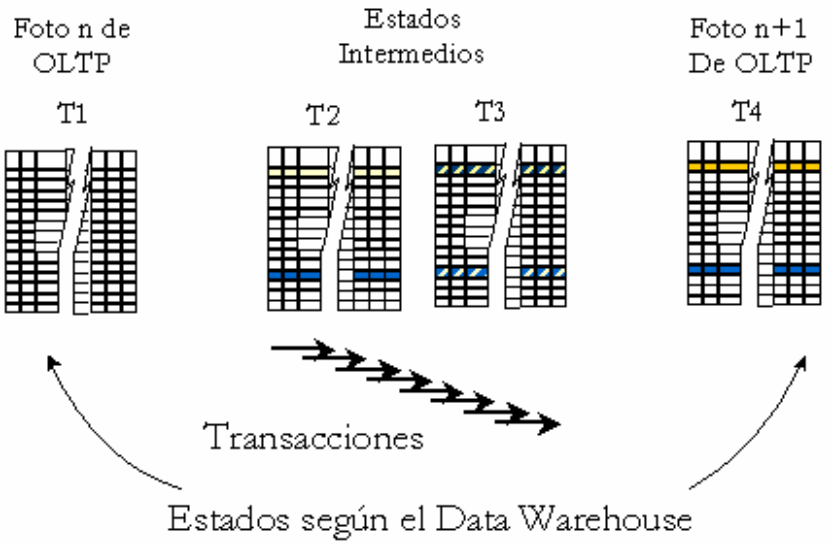
En el caso de los procesos de refresco periódico, es importante identificar la información que cambió en el último periodo de tiempo desde la ultima actualización, el volumen de información a tratar puede ser relativamente menor al caso de la primera carga.

La característica del proceso de carga deberá ser tal que permita el proceso más rápido posible, de manera que las múltiples fuentes de datos se puedan cargar sin interferir unas con otras ni tampoco con los procesos de lectura y análisis que se hacen en el warehouse.



Capturando los datos que cambian

Uno de los problemas a resolver en el diseño de un proceso de refresco del data warehouse es la técnica para identificar los registros que cambiaron en el sistema transaccional. Así también no se debe perder de vista que por la naturaleza periódica del proceso de carga siempre habrá transacciones (estados intermedios en la figura siguiente) que no se capturarán para el Data Warehouse, esta es una limitación intrínseca en todo proceso batch.



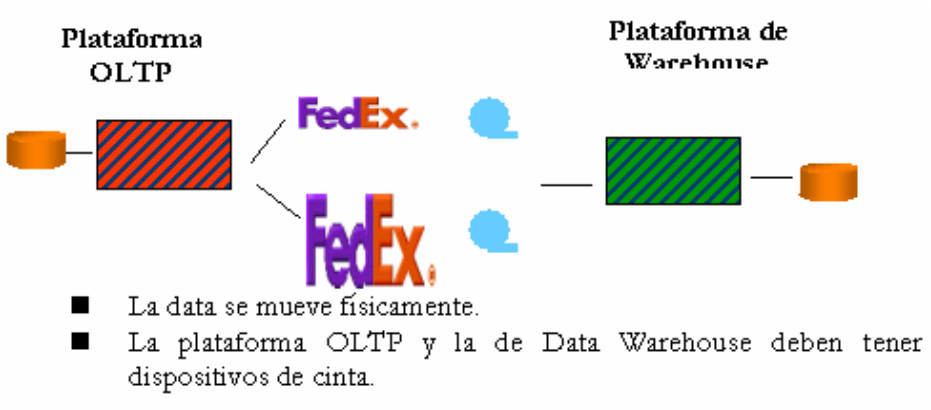
Algunas técnicas que se pueden utilizar para identificar los cambios son las siguientes:

- Comparar archivos.
- Sobre la base de fecha de cambio.
- Triggers.
- El log del DBMS.

Transferencia de datos

En función al tamaño de los archivos que se muevan de la plataforma OLTP a la plataforma de Data Warehouse se debe considerar una estrategia adecuada, para ello considerar los anchos de banda de las redes de comunicaciones que une las distintas plataformas.

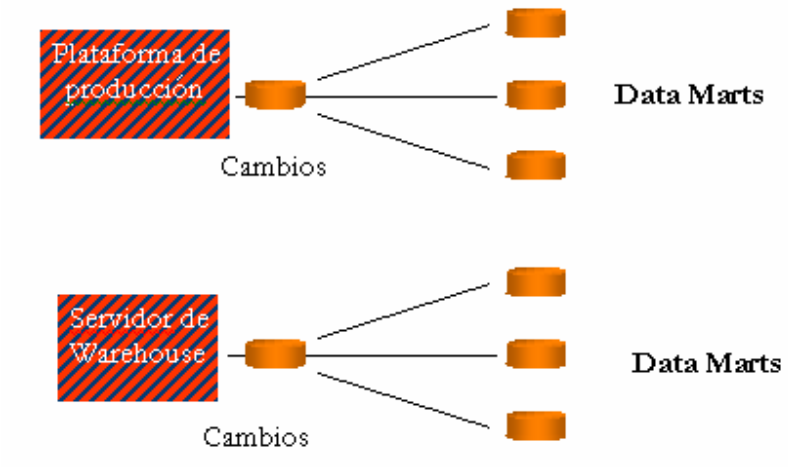
En los casos de la primera carga, dependiendo del volumen de datos a transportar, es posible que sea necesario transportar la información en medio fisico.



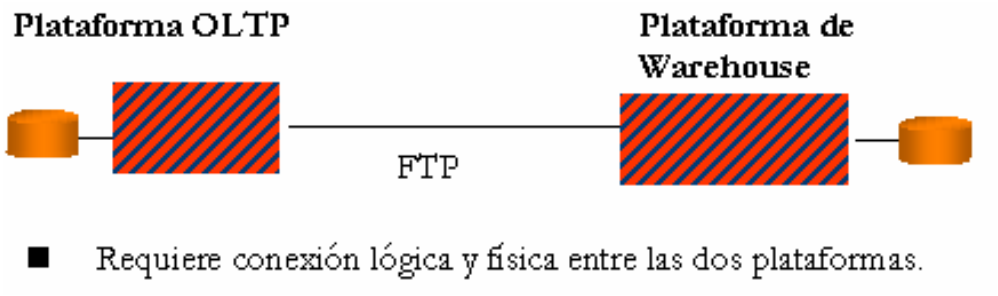
Alternativas para la transmisión de la información son las siguientes:

-Replicación de uno a muchos, cuando una BD replica a múltiples BD.
-Replicación de muchos a uno, cuando se replica de múltiples BD a una BD, también llamado centralización.
-Replicación uno a uno, cuando se replica de una BD a otra BD.

Cualquiera de estos esquemas puede aplicarse tanto de la fuente al warehouse como del warehouse a los DataMarts.



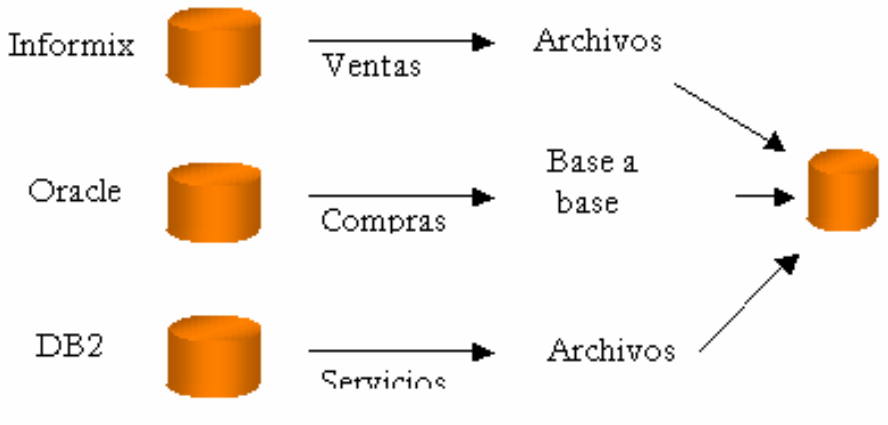
- Protocolo TCP/IP



¡Error!

- Otras formas de transferencia:

- Transferencia de Base a base (replicación).
- Transferencia de archivos.



12. Introducción a la calidad de datos

La calidad de los datos de las fuentes es un problema que debe ser enfrentado en el proceso de construcción de un Data warehouse. Los sistemas operacionales usualmente capturan muchos datos pero solamente validan aquellos que son de interés para el proceso en particular.

Es así que con la finalidad de mejorar la calidad de los datos de los sistemas operacionales, se originaran cambios en los sistemas transaccionales con la finalidad de mejorar la información a futuro, y procedimientos de gestión de datos que permitan mejorar la información histórica.

En general se definirá niveles mínimos de calidad también para el Data warehouse pues es probable que no toda la información de las fuentes pueda ser corregida o gestionada.

13. Dimensiones de la calidad de datos

El problema de la calidad de los datos se puede analizar de manera general desde tres grandes perspectivas, independiente de si se trate de la fuente o del Data warehouse.

Estas dimensiones son las siguientes:

- **Calidad de las definiciones de los datos**: Referida a si se cuenta con definiciones de reglas de negocio claras, completas y precisas.
- **Calidad del contenido**: Los valores de los datos deben ser concordantes con las reglas del negocio.
- **Calidad de la presentación**: Transformar la data en información accesible cuando sea necesario, es decir si la información esta disponible en la oportunidad que se requiera. Esta dimensión sería analizada después de construido el Data warehouse.

14. Atributos de la calidad de datos

Los atributos de la calidad de datos están referidos al contenido de los datos estructurados de las bases de datos a analizar, esta base de datos puede ser la del sistema transaccional o la base de datos de Data warehouse. Los atributos de la calidad de datos son los siguientes:

i. Exactitud

Es un atributo que no es identificable o cuantificable con mucha facilidad pues son valores de datos que están permitidos pero que no corresponden a información real o consistente. Se puede medir calculando el número de registros inexactos entre el número de registros totales.

Nombre	Sexo
Pablo Mármol	M
Sara Olivos	M
Pedro Picapiedra	M
Vilma Mármol	F

ii. Existencia

Es el número de registros en NULL entre el número de registros totales, nos dice el porcentaje con registros de campos nulos.

Nombre	Saldo
Jorge Basadre	230,000.00
	20,000.00
Pedro Picapiedra	
Vilma Marmol	5,000.00

iii. Validez

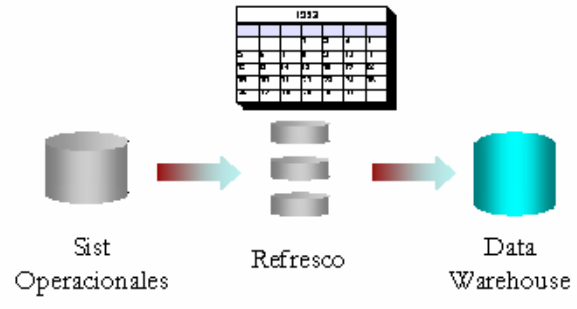
Es el atributo que mide el porcentaje de registros con valores fuera del rango o de los dominios definidos.

Nombre	Sexo
Pablo Mármol	M
Sara Olivos	X
Pedro Picapiedra	M
Vilma Mármol	F

Donde el dominio del campo sexo es: (M: Masculino F: Femenino)

iv. Temporalidad

Es atributo que mide la disponibilidad de los datos en un periodo de tiempo definido, por ejemplo el periodo de refresco del Data Warehouse.



v. Unicidad

Es el porcentaje de tablas con llave primaria y de llaves primarias que estén activas.

vi. Consistencia

Es el grado de integridad de los datos, referido a la existencia y validez de llaves primarias y llaves foráneas en la base de datos a analizar.

15. Anomalías de la data operacional

Las principales anomalías de los datos que se encuentran en las bases de datos operacionales son las siguientes:

- Errores en el ingreso de datos (digitación).
- Validaciones inadecuadas en los sistemas de ingreso de información.
- Prioridades de los sistemas operacionales.
- Datos que se ingresan en campos “Free-Form”.
- Anomalías en reglas de negocio (a través del tiempo).

Una tabla típica se vería como en la siguiente figura:

ID	Nombre	Dirección
10234567	Digitall Equipment	Av.Javier Prado 375
23451233	DEC	Avenida Javier P. 375
11558800	Digital Corp.	Av. J Prado 365 Lima
32456722	Dicital Consulting	Av J.P. 375-Lima 12
90876663	Digital Equip.	Av. Javier Prado Este 3

16. Soluciones

Existen en el mercado herramientas especializadas en limpieza de determinados tipos de datos, por ejemplo Trillium es una herramienta de limpieza de datos de nombres en ingles.

En general una solución de calidad de datos es un proceso en el que puede haber varias herramientas o algoritmos que ayuden a identificar los datos errados y a corregirlos. Este proceso debe ser complementado y apoyado por un trabajo manual arduo, en el que participaran las áreas usuarias de los sistemas operacionales y del mismo Data warehouse.

17. Finalidad del Datawarehouse

La finalidad de construir un Datawarehouse es poder analizar la información histórica comprendiendo el pasado y el presente para poder decidir el futuro.

El Datawarehouse es una base de datos que necesita una infraestructura de acceso a la información adecuada a cada tipo de usuario y a cada tema específico. Además se debe proporcionar información oportuna, integrada, en cualquier lugar.

18. El potencial del Datawarehouse

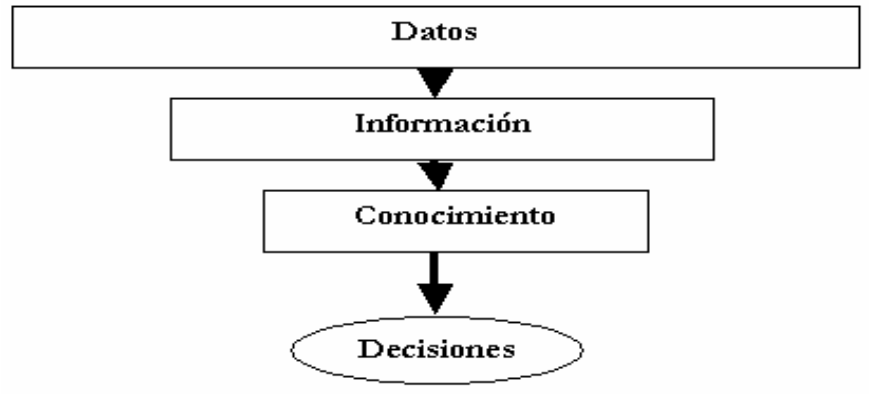
Derivar valor empresarial de un Datawarehouse es un esfuerzo complejo por ello es necesario proveer la infraestructura de acceso adecuada.

Las herramientas que debe tener el Datawarehouse deben ir desde las que permitan explorar la información en detalle hasta las que proporciones vistas agregadas de la información y que permitan tomar decisiones a diferentes niveles jerárquicos y funcionales.

Es así que un Datawarehouse proporcionará información para la toma de decisiones estratégicas a un nivel gerencial, y también para la toma de decisiones operativas como a nivel de atención al cliente en un Call center.

El potencial del Datawarehouse esta en extraer conocimiento a partir de los datos, y para lograrlo debe proporcionar las herramientas adecuadas.

¡Error!

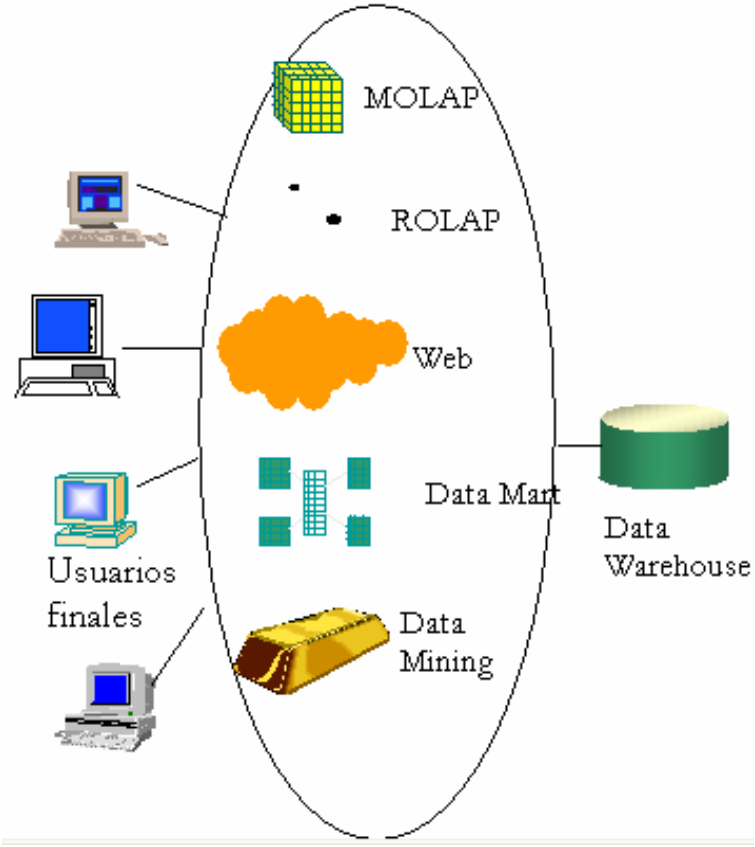


Las aplicaciones típicas de un datawarehouse son las siguientes:

- Análisis de rentabilidad y crecimiento.
- Administración estratégica.
- Conocimiento del cliente.
- Administración de relaciones con el cliente.
- Administración de los activos.
- Permite medir resultados.

19. La fábrica de información

El conjunto de aplicaciones de un Datawarehouse, que sirven para dar acceso a la información recibe el nombre fábrica de información. Esta fábrica de información tiene una diversidad de herramientas de consulta análisis y exploración de la información.



Al ser la fábrica de información el enlace entre el usuario y el Datawarehouse requiere de una arquitectura.

20. Los usuarios

Los usuarios son la razón de ser del Datawarehouse, pues son ellos quienes pueden extraer el conocimiento de la información con la ayuda de su experiencia.

Por ello los usuarios se caracterizan por o siguiente:

- Entienden la semántica de los datos del Warehouse.
- Aprenden a explorar el Warehouse.
- Tienen la experiencia.

Se debe proporcionar a cada usuario la herramienta más adecuada de acuerdo a su rol, función y de acuerdo a su experiencia con las herramientas de informáticas, para lograrlo se debe analizar y clasificar a los usuarios

Desde la perspectiva del Datawarehouse se tienen los siguientes tipos de usuarios:

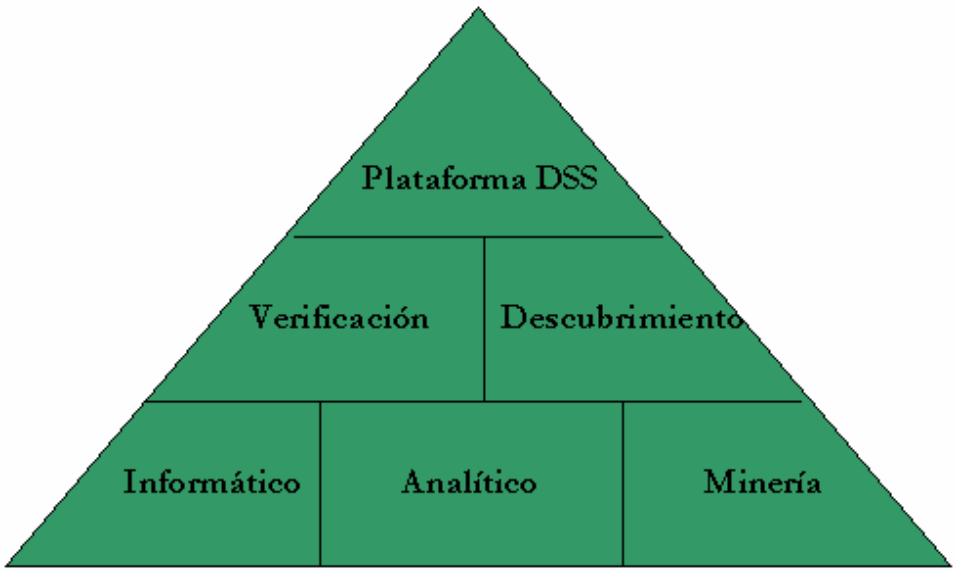
- Por funciones:
  - Contabilidad, marketing, producción.
- Por jerarquía:
  - Ejecutivos, analistas, apoyo.
- Por nivel de competencia:
  - Ocasionales, regulares y expertos.

21. Técnicas de acceso a los datos

Las técnicas de acceso al Datawarehouse son las formas en que se tiene el acceso a la información, estas se clasifican en:

- Procesamiento informático.
- Procesamiento analítico.
- Minería de datos.

El siguiente gráfico muestra la relación de las técnicas con las formas de análisis:



Autoevaluación

1. ¿Por qué se debe tener un proceso de primera carga distinto al proceso de refresco periódico?

2. ¿Cuáles son las características principales de un proceso de primera carga?

3. ¿Cuáles son las implicancias de un proceso de refresco que tiene cierta periodicidad, con respecto a las transacciones del sistema OLTP?

4. ¿Cuáles son las formas en las que se puede saber que registros cambiaron en el sistema OLTP?

5. ¿Qué formas de transmisión de datos se debe emplear en función al tamaño de la información?

6. ¿Una base de datos de mala calidad puede ser determinante en la factibilidad de la implementación de un área temática en el Warehouse? ¿Por qué?

7. ¿Las dimensiones de la calidad de los datos se pueden aplicar a cualquier base de datos o archivo?

8. Explique el atributo exactitud de la calidad de los datos.

9. ¿En el Data warehouse se tienen loas datos con todos los atributos al 100% de calidad?

10. Explique en que consistiría una solución para el problema de Calidad de datos.

11. ¿Cuál es la finalidad del Datawarehouse?

12. ¿Cómo se puede llegar de los datos al conocimiento y cual es el fin de este proceso?

13. ¿Cuáles son las aplicaciones típicas de un Datawarehouse?

14. ¿Cómo debe ser la estrategia frente a los sistemas de soporte a decisiones existentes?

15. ¿Qué es la fábrica de información?

16. ¿Por qué es importante analizar a los usuarios del datawarehouse?

17. Describa los componentes funcionales de la plataforma DSS

Para recordar

1. El volumen de la información a cargar será determinante en la forma de construir los procesos de primera carga y refresco del Data Warehouse.

2. El trabajo con grandes volúmenes de información requiere técnicas especializadas como el manejo de los utilitarios de las BD que permiten la carga y descarga masiva de datos.

3. La premisa cuando se construyen los procesos de carga es obtener el tiempo de respuesta más corto en todos los casos..

4. La naturaleza periódica del proceso de carga hace que no se puedan capturar todas las transacciones en el Data Warehouse.

5. La calidad de datos en un problema a enfrentar en el proceso de construcción de un Data Warehouse.

8. Se puede establecer métricas de calidad en el Data warehouse y en cualquier base de datos en general.

9. Una solución al problema es un proceso de negocio que debe estar fuertemente apoyado por herramientas y por un arduo trabajo manual., en el que se debe involucrar a los usuarios
10. La finalidad del Datawarehouse es extraer valor de los datos.

11. La fábrica de información facilita el acceso al Datawarehouse.

12. Es importante el análisis de los usuarios por roles, funciones, y experiencia, de manera que se les puede proporcionar la herramienta más adecuada.

SEMANA

14

Disponibilidad de soluciones en el mercado - Consultas y reportes como herramientas de acceso a los datos.

El Proceso KDD – MDX (I)

OBJETIVOS ESPECÍFICOS

- Identificar las soluciones y los proveedores que hay en el mercado
- Identificar las características de las consultas y reportes.
- Comprender el proceso de descubrimiento de datos
- Comprender los conceptos básicos del MDX

CONTENIDO

- Herramientas en el mercado
- Consultas DSS y OLTP.
- El procesamiento informático.
- Introducción al proceso KDD y MDX
- Conceptos Básicos de MDX

ACTIVIDADES

- Contestar las preguntas del cuestionario de autoevaluación.

22. Herramientas en el mercado

En el mercado existen muchas herramientas que poseen la capacidad de administrar completamente una solución datawarehouse, mientras otras tienen por objetivo la explotación de la data.

Mencionaremos algunos productos del mercado:

- *COGNOS (http://www.cognos.com)*
    - **Planeamiento y Consolidación**
        - ***Cognos Planning: Crea y contribuye a los planes y presupuestos***
        - ***Cognos Controller: Reportes Financieros y consolidados basados en web.***
        - ***Cognos Finance: Visualiza la información financiera de manera integrada.***
    - **Tablero de decisiones**
        - ***Cognos Metrics Manager: Crea y comparte tableros de decisión , asi como asignar valores***
    - **Inteligencia de Negocio**
        - ***Cognos ReportNet : Reportes de Producción y de negocios.***
        - ***Cognos PowerPlay :Implementa soluciones OLAP y de análisis.***
        - ***Cognos Visualizar : Permite visualizar datos.***
        - ***Cognos DecisionStream : Integración de datos y herramientas ETL.***
        - ***Cognos NoticeCast : Monitorea la actividad del negocio***
        - ***Cognos Performance Applications : Conjunto de métricas y reportes pre-definidos.***
- *PROCLARITY (Fuente: **http://www.proclarity.com/**)*
    - **Servidores Analíticos**
        - Dashboard Server: Es una capa de presentación dinámica para análisis, los cuales son proveídos por el Proclarity Analytics Server. El Dashboard Server, provee la facilidad de monitorear la performance del negocio asi como dar la posibilidad al usuario de poder comparar su desarrollo contra los objetivos de la empresa.
        - Business Logic Server: Incluye las herramientas del lado del cliente para poder crear, publicar y administrar la lógica de negocio que esta almacenada en el servidor de análisis. Nos permite el almacenamiento centralizado de mejores prácticas analíticas pre-definidas, incluyendo lógica de negocio y reglas; Indicadores clave de performancia; miembros y conjuntos.

- Analytics Server: es el fundamento y base para todo la familia de productos analíticos. Este servidor de 3 capas administra las conexiones y las consultas hacia los cubos, entregando los resultados al usuario. Además administra el procesamiento de las consultas, la seguridad centralizada, el escalamiento.

- **Familia de Clientes Analíticos**

  - Dashboard viewer. Permite ver la información de la siguiente manera:
    - Descomposición en árbol
    - Perspectiva
    - Mapa de performance
  - Web Standard. Permite :
    - Capacidad de analizar la información prácticamente desde cualquier lugar
    - Casi cero código para desarrollar facilidades
    - Cliente liviano
  - Web Professional. Permite :
    - Análisis de datos no estructurados, como causa efecto, exploración de datos y cálculos avanzados.
    - Cliente web liviano y versión windows.
  - Business Reports for Excel. Permite :
    - Exportar datos actualizados a Microsoft Excel.
    - Crear reportes ricos en formato.
    - Posibilidad de actualizar la información diariamente, mensualmente, semanalmente, etc.
  - ProClarity for Reporting Services
    - Integración con Microsoft Reporting Services.
    - Crear reportes OLAP en Reporting Services, sin usar complejos MDX.
    - Posibilidad de vincular reportes con vistas ProClarity.
  - ProClarity for SharePoint Portal Server
    - Presentación de KPI's en el portal.
    - Mínimo soporte y mantenimiento, es posible realizarlo a través del Proclarity Business Logic Server.
    - Facilidad de crear y agregar vistas analíticas al portal

- *INTELLIBROWSER (Fuente: http://www.downloaddatabase.com/databasesoftware/intellibrowser-powerful-olap-tool.htm)*

IntelliBrowser es una herramienta OLAP grafica para Microsoft OLAP and Analysis Services, la cual provee análisis de datos multidimensionales, visualización y reporteo. Con el cliente OLAP del IntelliBrowser, se pueden desarrollar rápidas y sofisticados análisis para extraerla inteligencia de negocio de la información corporativa.

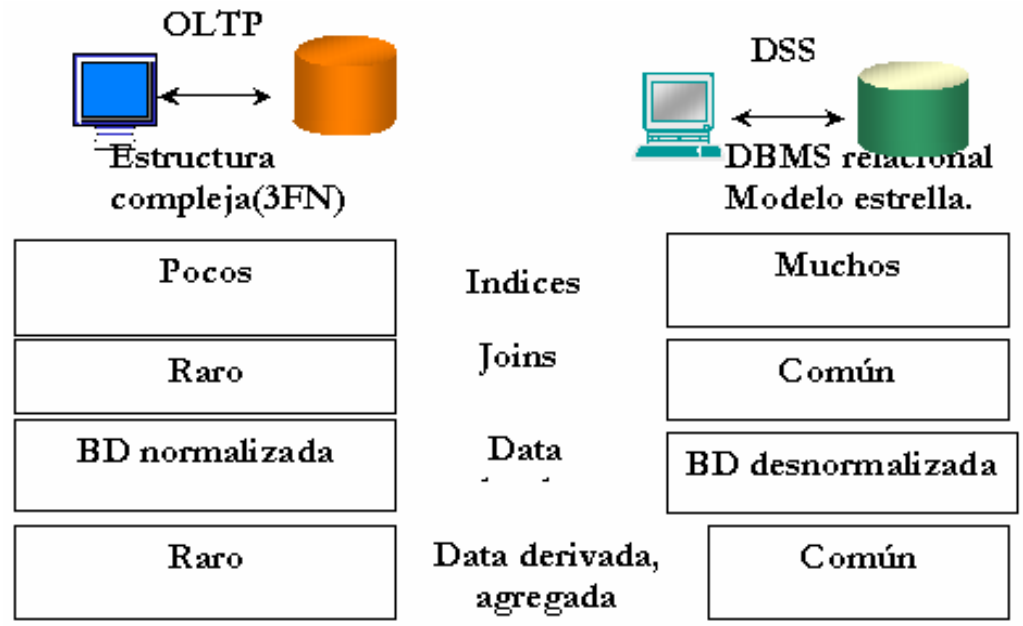
23. Consultas DSS y OLTP

Las consultas DSS (Decision Support Systems), son las orientadas a leer la información de una base de datos diseñada para el soporte a las decisiones, como es el caso del DataWarehouse.

La plataforma DSS tiene características muy particulares que la diferencian de la plataforma de un sistema transaccional convencional en línea OLTP(On-Line Transaction Process)

El siguiente cuadro muestra las principales diferencias entre un sistema DSS y un OLTP:

¡Error!



Características de las consultas OLTP:

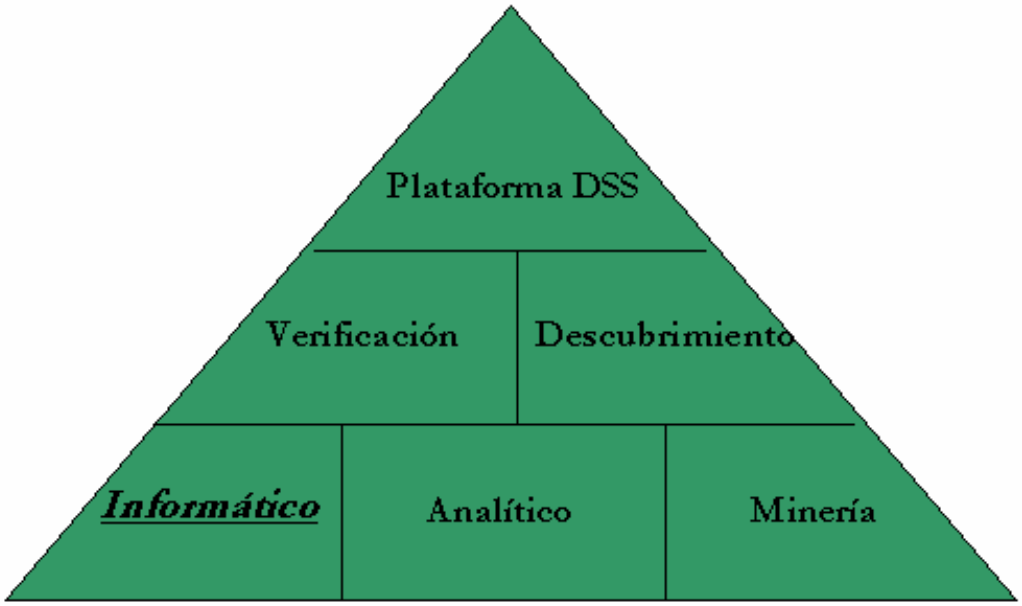
- Son pequeñas y predefinidas.
- Entrada pequeña y mensajes de salida
- Los usuarios no escriben las consultas

Características de las consultas DSS:

- Consultas indeterminadas.
- Consultas voluminosas y “lentas”.
- Los usuarios definen las consultas
- La data está para periodos de tiempo largos

24.El procesamiento informático

El procesamiento informático, es aquel orientado a la generación de consultas y reportes desde el Datawarehouse, este proceso en algunos puede ser en línea en otros en batch.
En la plataforma DSS se ubica en el tipo de procesamiento orientado a la verificación o consulta.
¡Error!



El procesamiento informático es conocido también como consultas de inteligencia de negocios.

Etapas del procesamiento informático:

1. Definir el aspecto empresarial
2. Definir la hipótesis.
3. Determinar necesidades de datos.
4. Acceder y recuperar los datos.
5. Analizar y presentar resultados.

25.Las necesidades de los usuarios

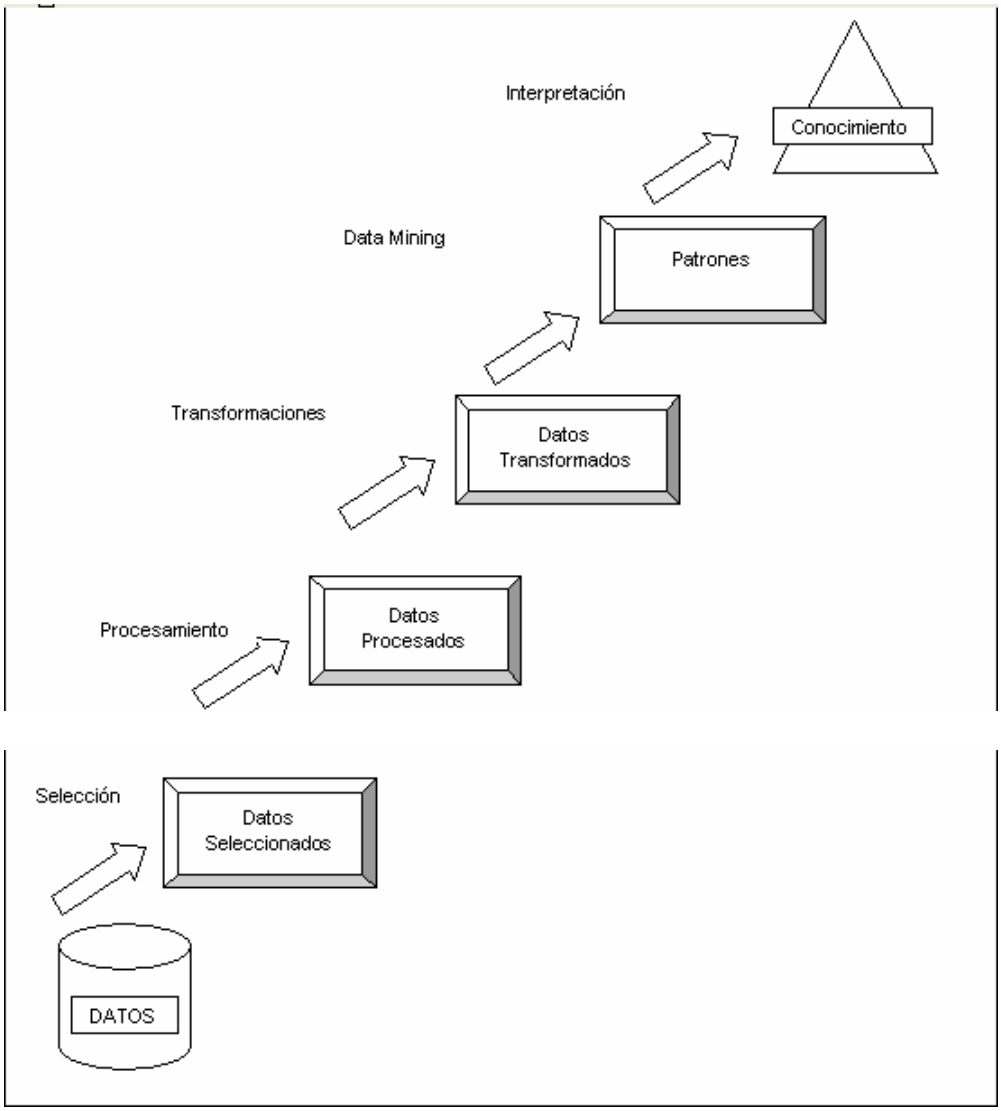
Los usuarios tendrán necesidades muy particulares en cada caso se tendrá que proporcionar la solución mas adecuada.

Entre las necesidades más comunes se tiene:

- Cuadros, gráficos, reportes estándar.
- Consultas predefinidas y análisis *ad-hoc.*
- En el Data Warehouse debe haber datos con diferentes niveles de granularidad.

26. Introducción al proceso KDD

KDD, es un proceso de identificar patrones que no son conocidos con anterioridad, dichos patrones son nuevos, útiles y validos dentro del cúmulo de datos.



PROCESO KDD

Proceso KDD :

- **Selección**
  - Identificar las fuentes de datos involucradas en el análisis
  - Tener presente los objetivos en el momento de decidir que datos serán usados
  - Conocer la infraestructura donde los datos residen

- **Procesamiento**
  - Depuración de los datos
  - Hacer que los datos sean de calidad

- Evite inconsistencias y errores en el dominio de campos
- Debe existir estándares en el proceso de depuración de datos.

- **Transformación**
  - Describir los datos totalmente
  - Poder enriquecer los datos empleando información de otras fuentes distintas

- **Datamining**
  - Descubrir tendencias y patrones
  - Uso de algoritmos para procesos KDD
  - Varias técnicas como: Clasificación, Segmentación, etc.

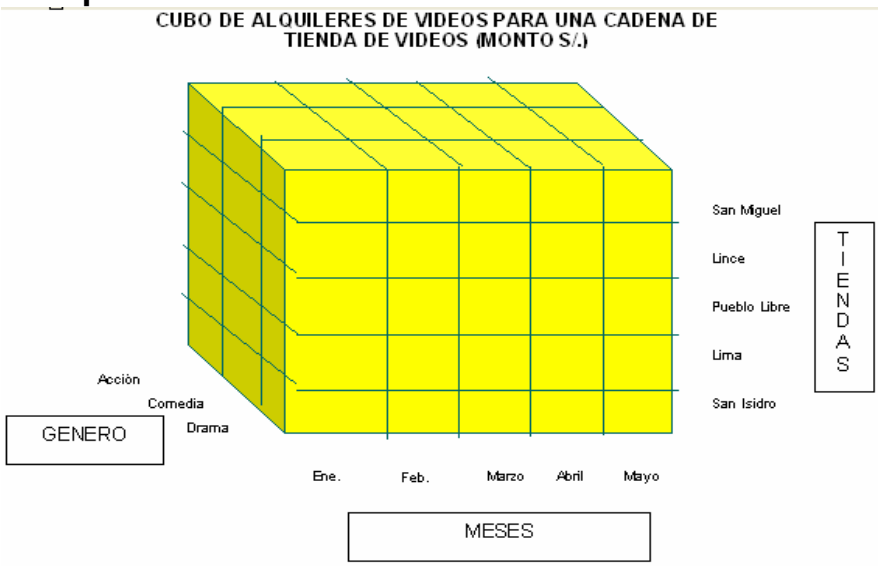
- **Interpretación**
  - Evalúe los patrones encontrados y su posible uso en el proceso de toma de decisiones.
  - Obtenga conocimiento

27. Introducción al MDX

MDX (Multidimensional expresión), es el lenguaje de consulta empleado en soluciones OLAP, es factible crear cálculos y consultas.

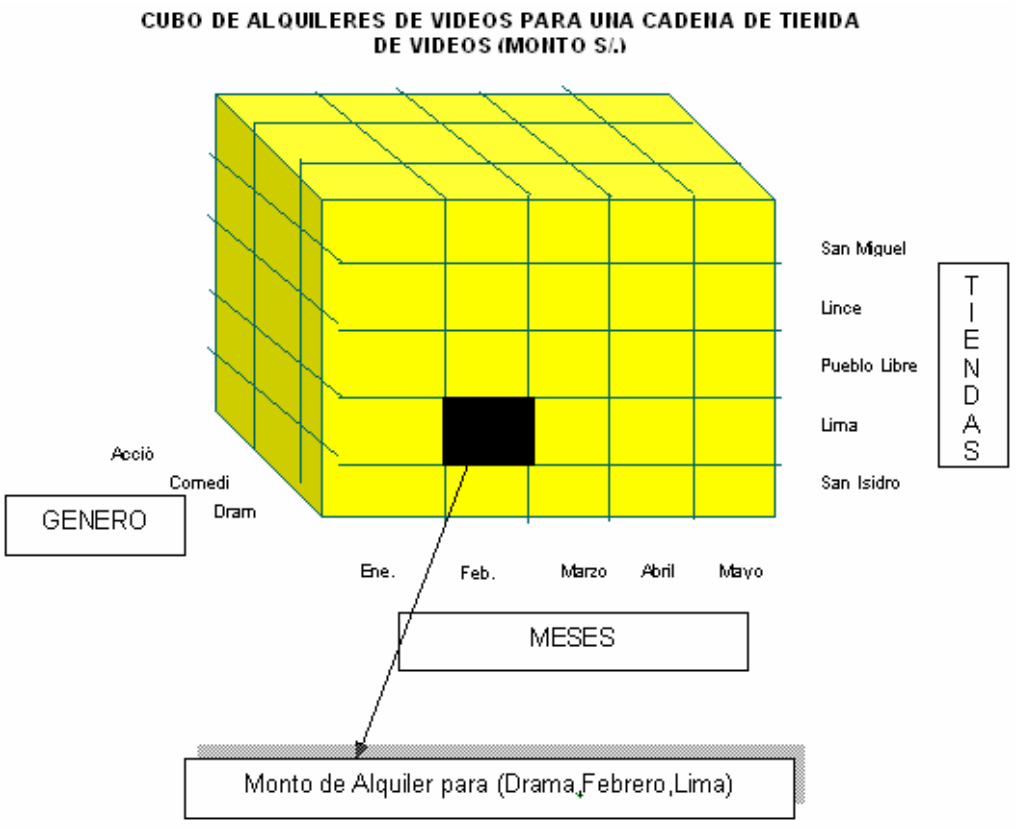
El MDX, es también empleado por las aplicaciones clientes con el fin de poder recuperar los datos provenientes de las bases de datos OLAP , con la facilidad de poder personalizar las consultas.

28. Conceptos Básicos de MDX

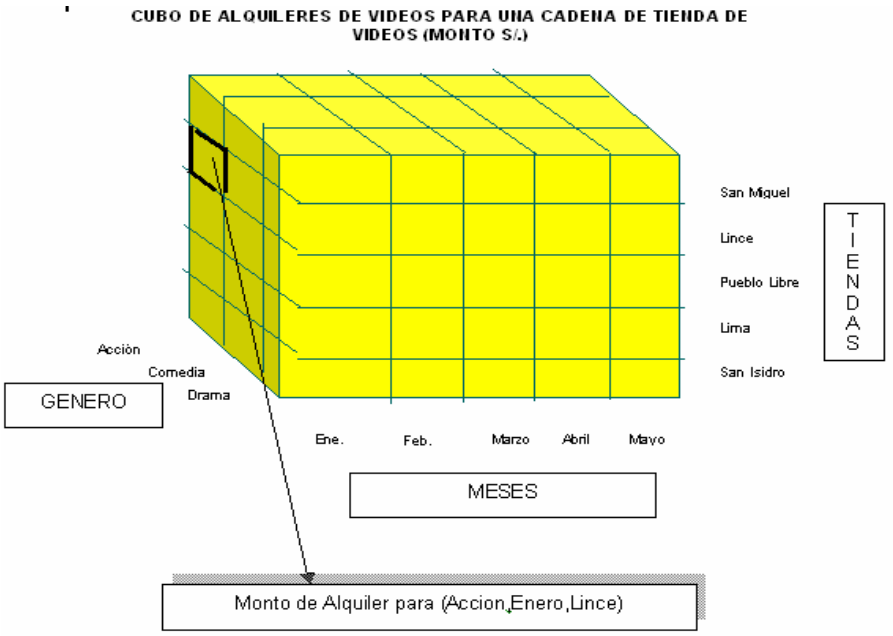


Vamos a trabajar sobre el cubo mostrado arriba, es importante antes de comenzar con algunas definiciones conocer como ubicarnos dentro del cubo.

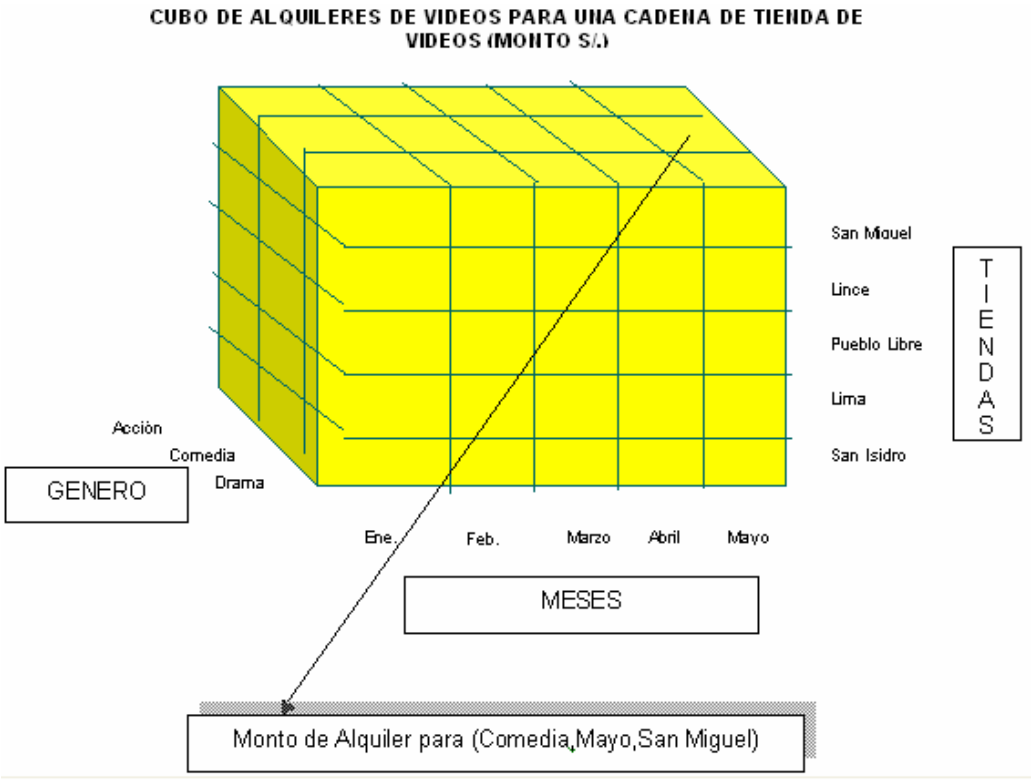
Ejemplo 1: Ubicar el monto de alquiler para el genero “Drama”, en el distrito de “Lima”, en el mes de “Febrero”.



Ejemplo 2: Ubicar el monto de alquiler para el genero “Acción”, en el distrito de “Lince”, en el mes de “Enero”.



Ejemplo 3: Ubicar el monto de alquiler para el genero “Comedia”, en el distrito de “San Miguel”, en el mes de “Mayo”.



Vamos a implementar las consultas anteriores usando MDX, para luego definir ciertos conceptos:

Ejemplo 1: Ubicar el monto de alquiler para el genero “Drama”, en el distrito de “Lima”, en el mes de “Febrero”.

```
SELECT
   {
      ([Measures].[MontoAlquiler])
   } on columns,
   {
      ([Genero].[Drama],[Tiendas].[Lima],[Meses].[Febrero])
   } on rows
FROM Alquiler
```

Ejemplo 2: Ubicar el monto de alquiler para el genero “Acción”, en el distrito de “Lince”, en el mes de “Enero”.

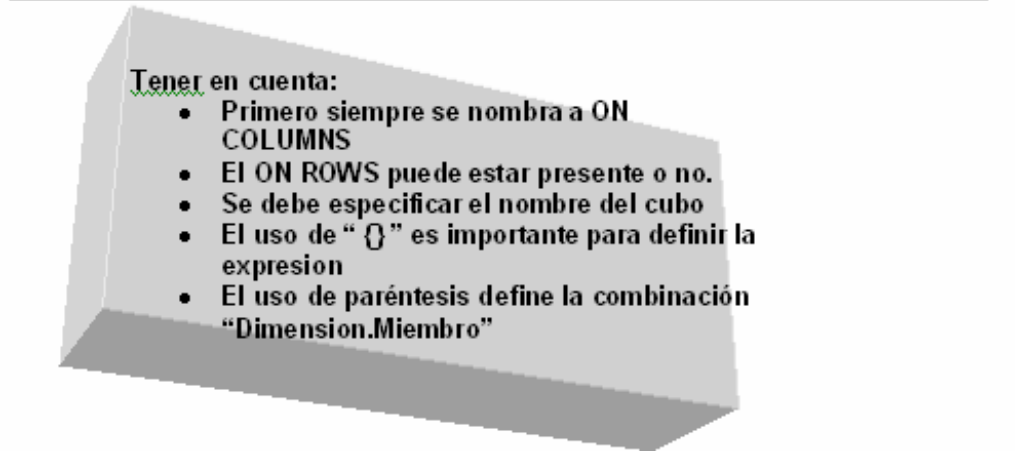
```
SELECT
   {
      ([Measures].[Monto Alquiler])
   } on columns,
   {
     ([Genero].[Accion],[Tiendas].[Lince],[Meses].[Enero])
   } on rows
FROM Alquiler
```

Ejemplo 3: Ubicar el monto de alquiler para el genero “Comedia”, en el distrito de “San Miguel”, en el mes de “Mayo”.

```
SELECT
   {
      ([Measures].[Monto Alquiler])
   } on columns,
   {
     ([Genero].[Comedia],[Tiendas].[San Miguel],[Meses].[Mayo])
   } on rows
FROM Alquiler
```

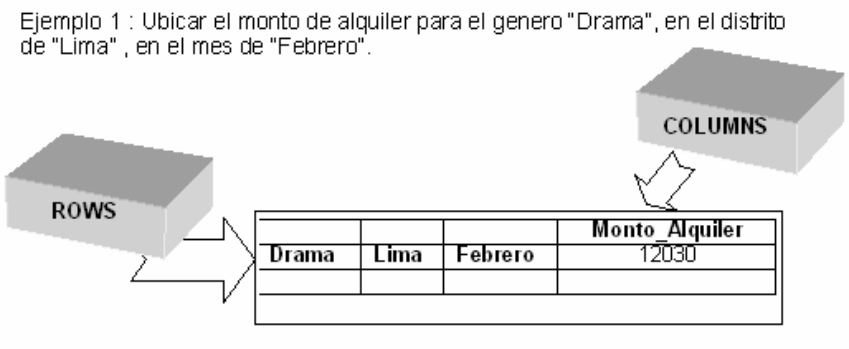
La estructura se explica de la siguiente manera:

```
SELECT
  { (Dimension.Miembro) } on columns,
  { (Dimension.Miembro) } on rows
FROM <Nombre de Cubo>
```



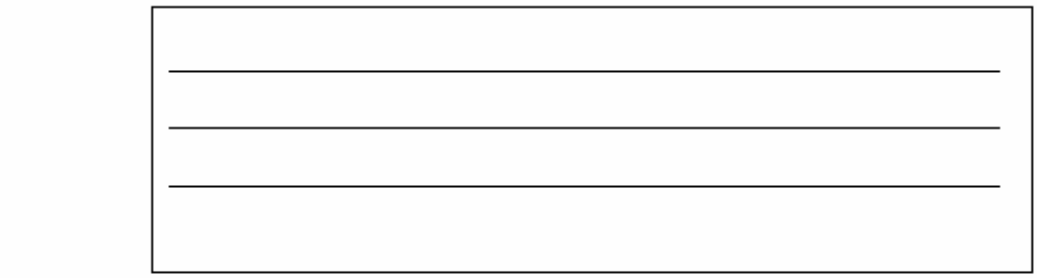
La salida de los SELECTS para los ejercicios seria de la siguiente manera:

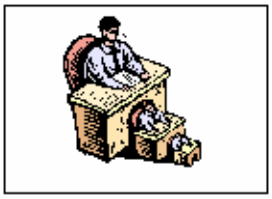
Ejemplo 1: Ubicar el monto de alquiler para el genero “Drama”, en el distrito de “Lima”, en el mes de “Febrero”.



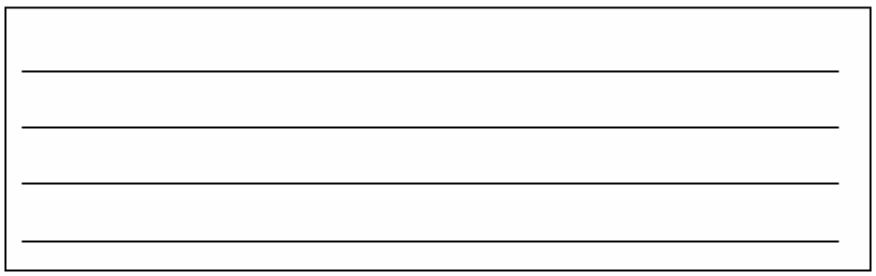
Hágalo usted:

Ejemplo 2 : Ubicar el monto de alquiler para el genero "Acción", en el distrito de "Lince" , en el mes de "Enero".

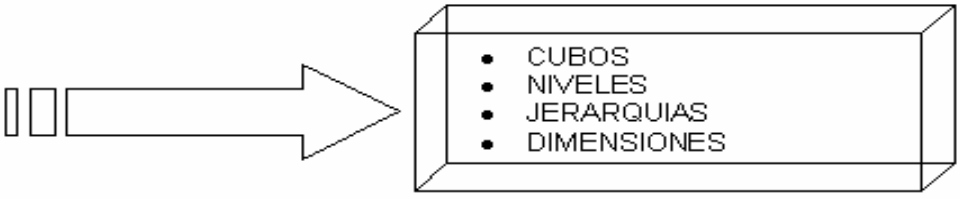


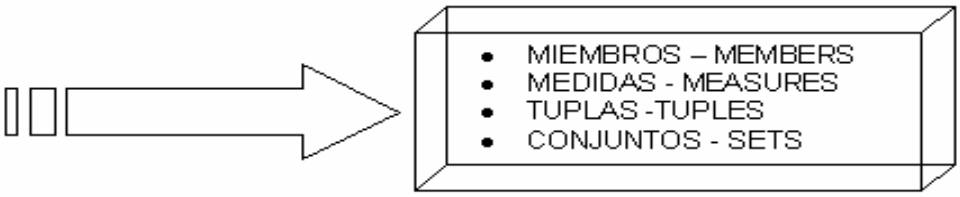


Ejemplo 3 : Ubicar el monto de alquiler para el genero "Comedia", en el distrito de "San Miguel" , en el mes de "Mayo".

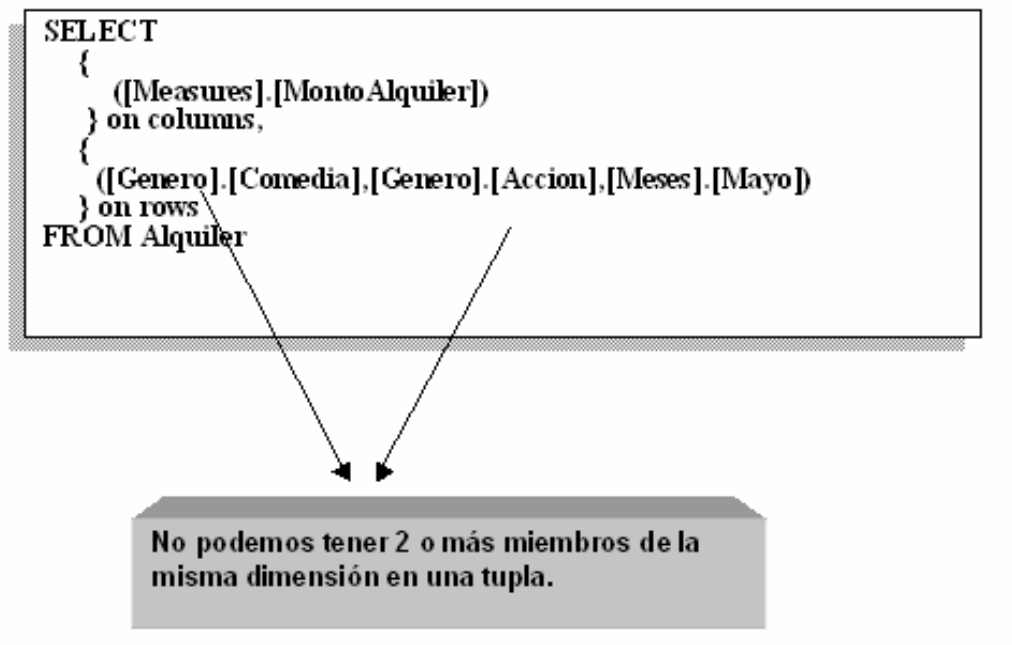


Para comprender las consultas MDX, se debe conocer los siguientes elementos.





Cuando es un error.



Autoevaluación

1. ¿Cuáles son las diferencias entre los sistemas DSS y OLTP?
2. ¿Cuáles son las características fundamentales de una consulta DSS?
3. ¿Cuáles son las características fundamentales de una consulta OLTP?
4. ¿Por qué se dice que el procesamiento informático es para el apoyo de una función de verificación?
5. Mencione las etapas del procesamiento informático
6. ¿Cuáles son las necesidades de los usuarios en lo que se refiere a procesamiento informático?
7. Investigue otras herramientas de soluciones OLAP.
8. ¿Por qué no usar SQL con un modelo dimensional?
9. ¿Cómo esta conformado un cubo?
10. ¿Qué puede hacer con MDX?

Para recordar

1. Los sistemas DSS tienen características opuestas a los OLTP.
2. El procesamiento informático se le conoce como consultas de Inteligencia de negocios.
3. El procesamiento informático es una actividad de varios pasos.
4. Las necesidades de los usuarios abarcan un amplio espectro.
5. MDX, es el lenguaje de consulta en cubos
6. Una tupla es la combinación de varios miembros de varias dimensiones, pero solo un miembro de cada dimensión participante
7. Un conjunto es la agrupación de tuplas, con el mismo orden y cardinalidad

SEMANA

15

MDX (II) - Minería de datos (I)

OBJETIVOS ESPECÍFICOS

- Entender la construcción del MDX
- Comprender el proceso de minería de datos.

CONTENIDO

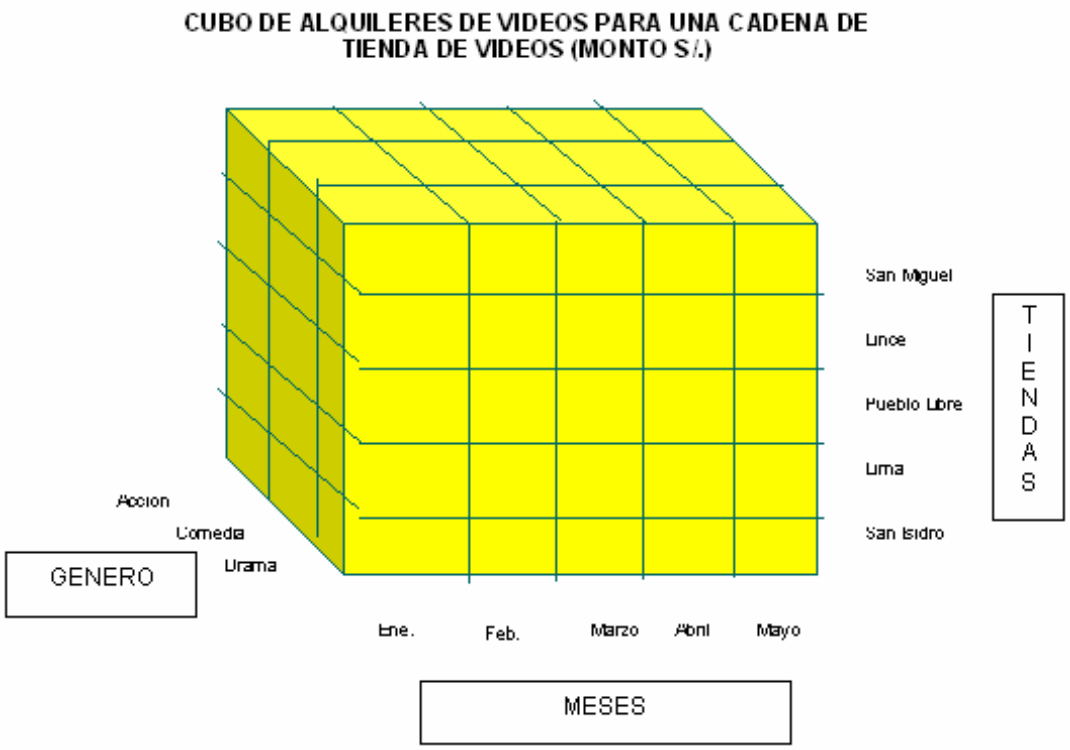
- Construcción de MDX.
- Introducción a la minería de datos.
- Las herramientas
- Algoritmos estadísticos.
- Algoritmos de KDD(Knowledge Discover in Databases)..

ACTIVIDADES

- Contestar las preguntas del cuestionario de autoevaluación.

29. Construcción de MDX.

Vamos a mostrar algunos ejemplos mas, que serán de utilidad en la construcción de MDX



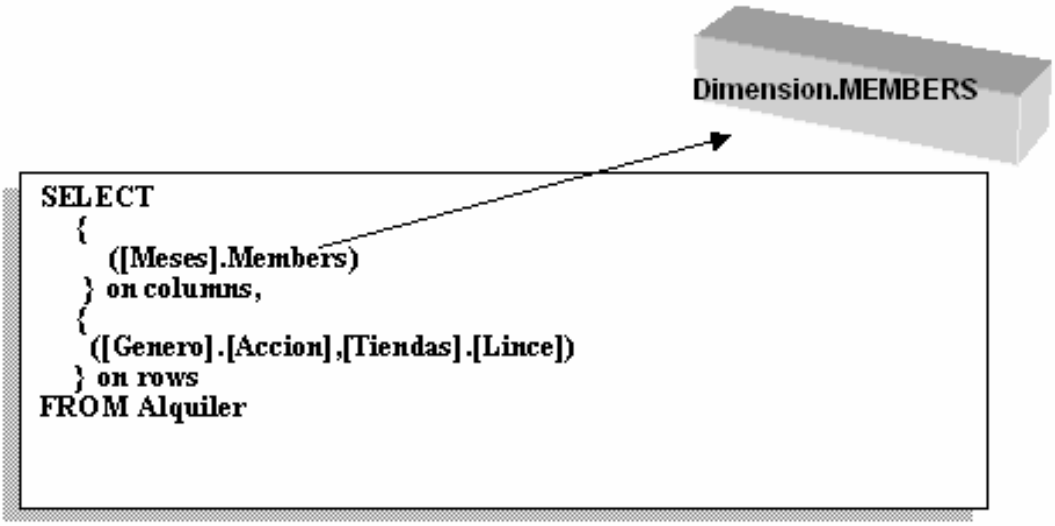
Ejemplo 1: Ubicar el monto de alquiler para el genero “Drama”, en el distrito de “San Isidro”, para todos los meses del año. (Asumir solo data para el 2004)

```
SELECT
   {
      ([Measures].[Monto Alquiler])
   } on columns,
   {
    ([Genero].[Accion],[Tiendas].[Lince])
   } on rows
FROM Alquiler
```

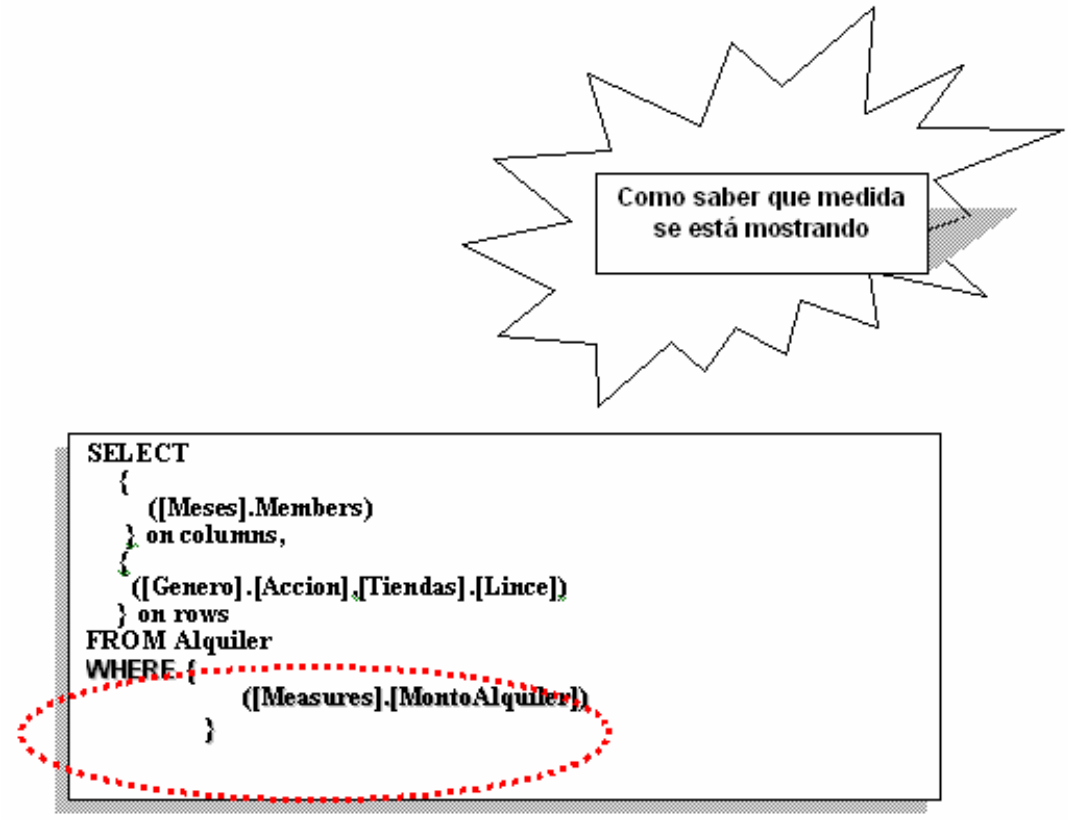
La salida del MDX, seria:

		Monto_Alquiler
Acción	Lince	50030

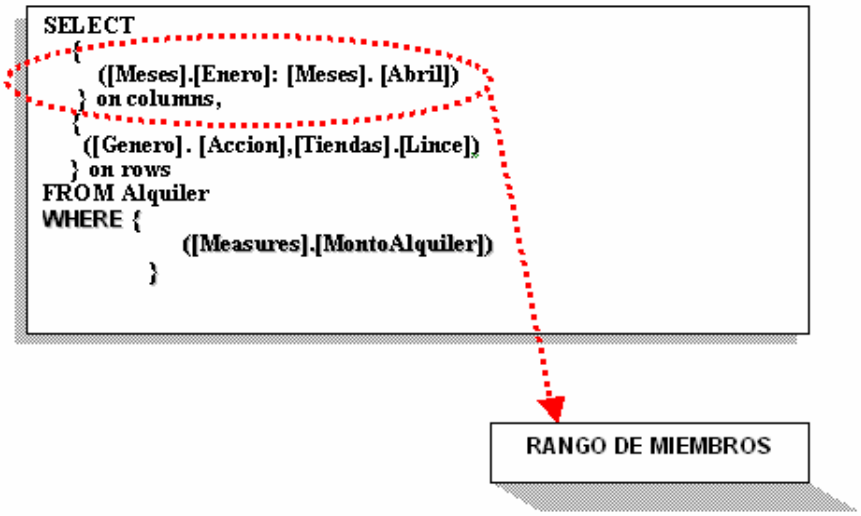
Pero no podemos saber cuanto le corresponde a cada mes



		Enero	Febrero	Marzo	Abril	Mayo
Acción	Lince	1000	2000	40000	2304	10033



Ejemplo 2 : Ubicar el monto de alquiler para el genero “Drama”, en el distrito de “San Isidro” , para los meses de Enero a Abril. (Asumir solo data para el 2004)



Ejemplo 3: Ubicar los distritos donde el monto de alquiler para el mes de Enero es mayor a 1000, en el género Drama.

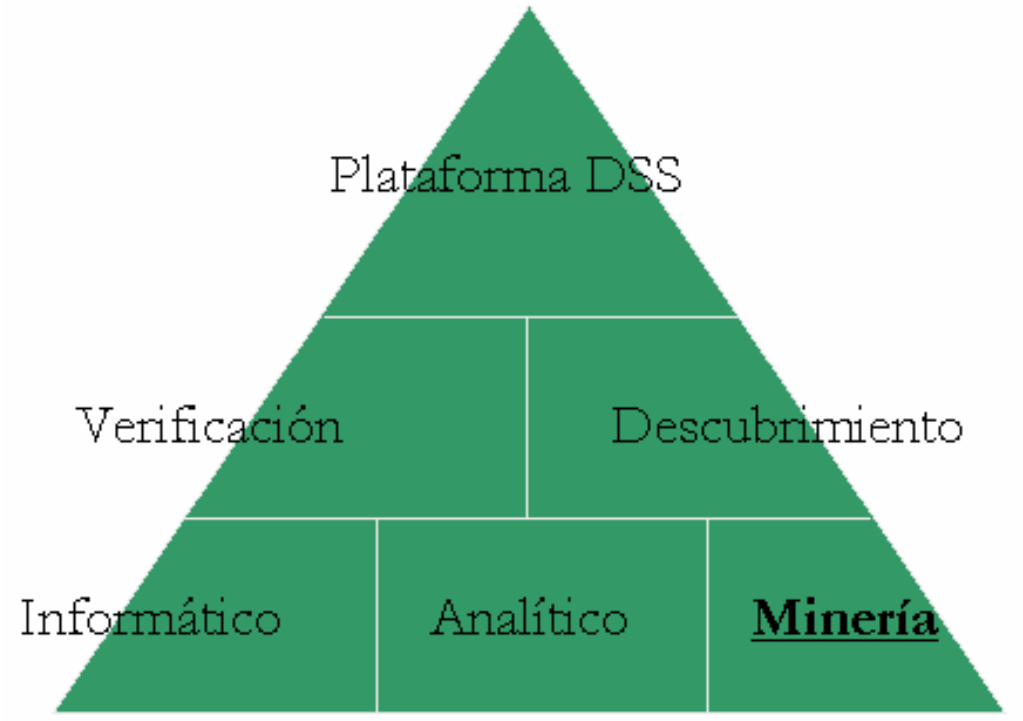
```
Select
  Filter
      (
       [Tiendas].members,
       ([Measures]. [MontoAlquiler],[Genero].[Drama])>1000
      )

   on columns
From Alquiler
```

30. Introducción a la minería de datos.

Una parte fundamental de la plataforma de soporte a decisiones son los procesos de minería de datos basados en algoritmos estadísticos y en algoritmos de descubrimiento de información en Bases de datos (KDD).

Este tipo de procesos están orientados a descubrir patrones, tendencias, relaciones, agrupamientos relevantes para el negocio que hasta este momento eran desconocidos.

¡Error!



Los roles participantes en un proceso de minería de datos son los siguientes:

- Análisis del negocio

- Análisis estadístico y matemático
- Análisis de la información.

El análisis del negocio determina las variables a analizar y los temas a analizar, luego en la interpretación de los resultados que se obtengan con los algoritmos.

El análisis estadístico y matemático es la correcta aplicación de los algoritmos a los problemas reales de la empresa.

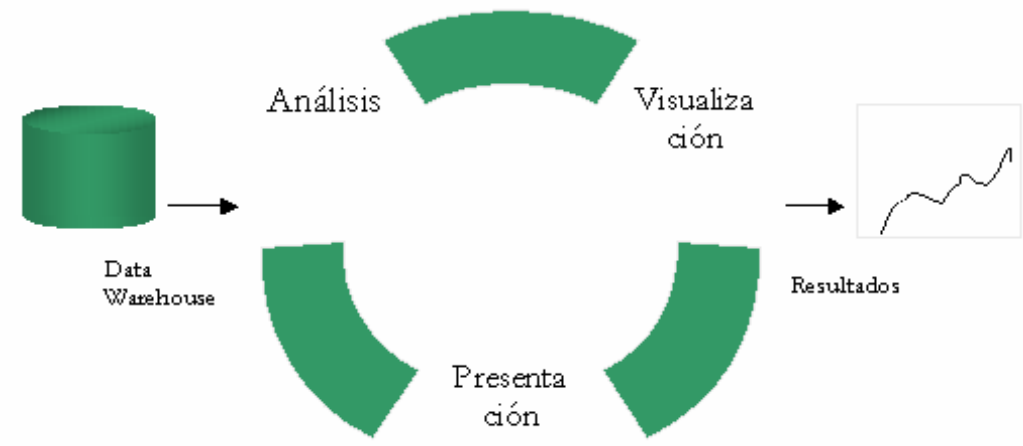
El análisis informático se encarga del soporte tecnológico y la provisión de información para los algoritmos, se enfrentan los problemas de calidad y de normalización de la información.

Los factores que propiciaron el crecimiento de la minería de datos son:

- El Data warehouse, que proporciona la información integrada y se ha enfrentado a los problemas de calidad de los datos.
- La reducción de los costos del hardware, lo que permite procesar grandes volúmenes de información aplicando algoritmo complejos.
- La evolución de las herramientas como los algoritmos, que facilitan al usuario la tarea de análisis.

31. Las herramientas

En el siguiente esquema se puede apreciar las funcionalidades básicas de las herramientas de minería de datos.



32. Algoritmos estadísticos

La aplicación del análisis estadístico tiene las siguientes fases:

- Se utilizan para detectar patrones no usuales de datos.
- Estos patrones se explican mediante modelos estadísticos o matemáticos.

Las funciones incorporadas en una herramienta de análisis estadístico son las siguientes:

- Funciones de visualización.
- Funciones exploratorias.
- Funciones estadísticas.
- Funciones de administración de datos.
- Funciones de grabación y reproducción.
- Herramientas de presentación.
- Herramientas de desarrollo.
- Tiempo de respuesta razonable.

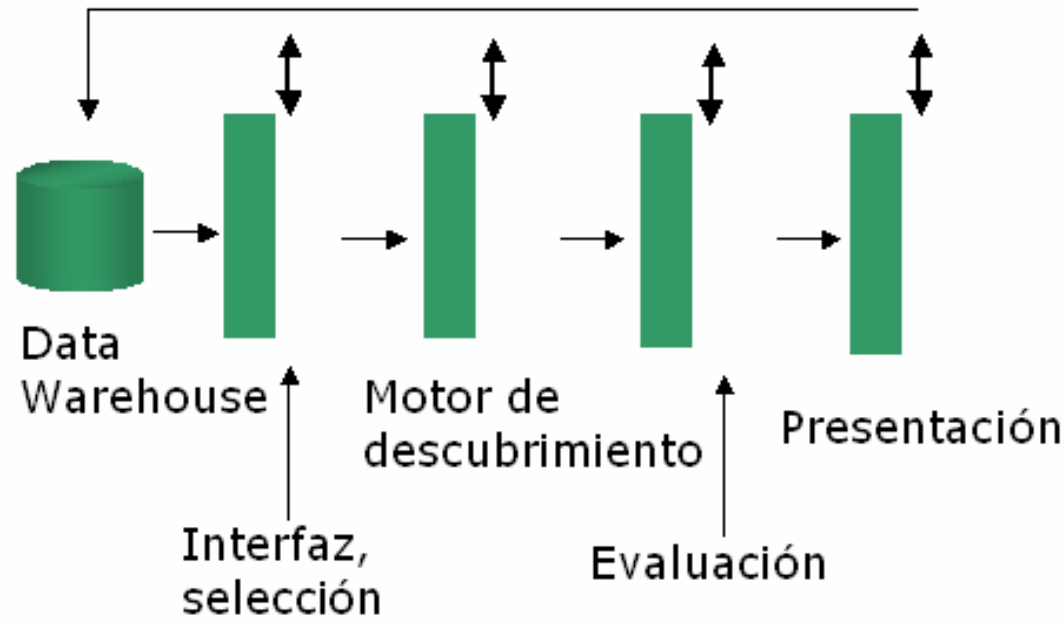
Los retos del trabajo con algoritmos estadísticos son los siguientes:

- Trabajo intenso.
- Los resultados dependen de la habilidad del analista.
- Muchas veces no se sabe qué buscar.
- Es complicado trabajar con datos no-numéricos.

33. Algoritmos de descubrimiento de conocimientos.

El proceso KDD busca extraer información implícita no trivial de las bases de datos, que no era conocida y que sea de utilidad. Para lograrlo se procesa la información con algoritmos neuronales, árboles de decisión, entre otros.

El proceso KDD tiene las fases que se indican en la figura adjunta



Tecnología del descubrimiento de conocimientos

- Basada en algoritmos para patrones y relaciones.
- Tareas genéricas:
  - Análisis de dependencias.

- Clasificación
- Descripción de conceptos.
- Redes neuronales.
- Detección de desviaciones.

Retos

- Calidad de datos.
- Bases de datos muy grandes.
- Desempeño y costos.
- Técnicas de analistas empresariales.
- Calidad de datos.
- Bases de datos muy grandes.
- Desempeño y costos.
- Técnicas de analistas empresariales.

Autoevaluación

1. ¿Por qué usar members?
2. ¿Por qué se dice que la minería de datos esta orientada a “descubrir”?
3. ¿Qué factores promovieron el desarrollo de la minería de datos?
4. ¿Cuáles son los roles en un proceso de minería de datos?
5. ¿Cuáles son las características de un proceso estadístico?
6. ¿Cuáles son las características del KDD?

Para recordar

1. Se pretende descubrir, en los datos, cosas que no son evidentes y que sean útiles para el negocio.
2. La minería de datos ayuda a descubrir estas “relaciones insospechadas”.
3. Las tecnologías de minería se categorizan en:
    a. -Análisis estadístico.
    b. -Descubrimiento de conocimiento.
4. La minería de datos es un componente esencial del paquete DSS.
5. Uso de FILTER, dentro de expresiones SELECT.

SEMANA

16

Minería de datos (II)

Sesión de Integración (2)

OBJETIVOS ESPECÍFICOS

- Identificar las aplicaciones de la minería de datos.
- Integrar conceptos del curso

CONTENIDO

- Identificar las técnicas de minería de datos.
- Ejemplo: Datamining en la Web.
- Uso actual de Datamining.

ACTIVIDADES

- Contestar las preguntas del cuestionario de autoevaluación.

34. Identificar las técnicas de minería de datos

Dentro de las técnicas de minería de datos podemos encontrar los siguientes:

- El problema de la extracción de patrones
- Métodos estadísticos
- Reglas de asociación y dependencias
- Métodos basados en casos, en densidad o distancia
- Métodos bayesianos
- Árboles de decisión y sistemas de aprendizaje de Reglas
- Métodos relacionales y otros métodos declarativos
- Redes neuronales artificiales
- Métodos basados en núcleo y máquinas de soporte vectorial
- Métodos estocásticos

Adaptado de
Fuente: : Introducción a la minería de datos
Autor : - César Ferri Ramírez; José Hernández Orallo; María José Ramírez Quintana

35. Ejemplo de Datamining en la web

La pantalla anterior registraría la información del usuario. Es posible poder recoger datos referentes a:

- Datos personales
- Datos laborales
- Datos financieros
- Datos preferencias
- Datos hábito de consumo
- Datos académicos

Generalmente, no le damos uso a dicha información, nuestros sistemas transaccionales no profundizan en ellos.

Inclusive las soluciones OLAP, sólo nos permiten reportear y analizar la información que ellos no muestran, como cantidad de visitantes al site de un determinado país, cuántos profesionales nos visitan anualmente, etc.

La posibilidad de poder encontrar asociaciones escondidas entre los datos generados y el perfil de los usuarios no es factible obtenerla con procedimientos OLAP, la solución es el empleo de técnicas datamining.

Por ejemplo:

- Las personas de sexo masculino del departamento de Lima consumen mucho limón en los meses de verano.
- La venta de papas fritas se incrementa, al consumir habas y gaseosas los sábados por las noches.
- Identificar a nuestros compradores compulsivos

Estas asociaciones y muchas otras se dificultan si intentáramos encontrarlas sin el uso de las herramientas datamining.

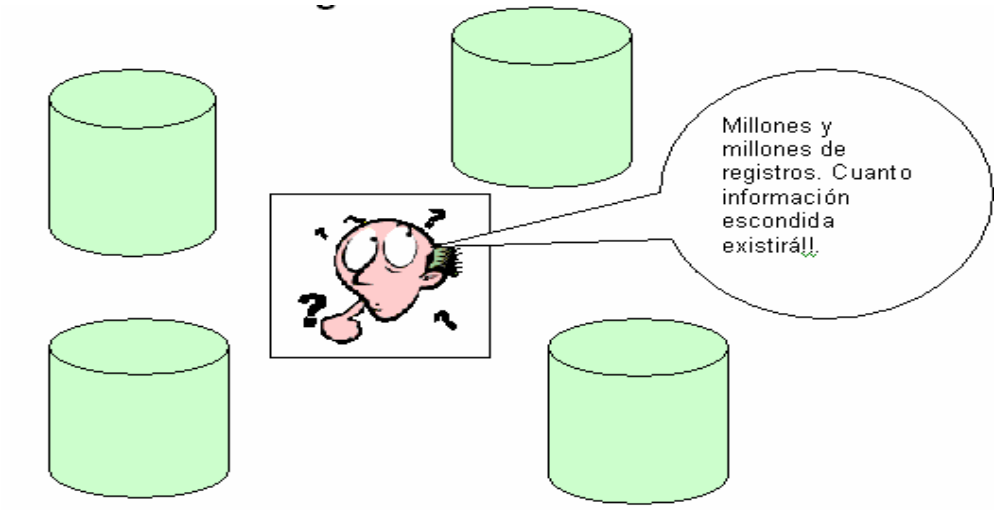
Una de las facilidades para realizar datamining es el tener definidos con anterioridad los objetivos del negocio.

Tras el datamining, es factible conseguir:

- Mercado objetivo
- Personalización
- Asociaciones
- Segmentación
- Administrar conocimiento
- Clasificación

Adaptado de
Fuente: : Data Mining on the Web There's Gold in that Mountain of Data
Autor : - Dan R. Greening
Url: http://psychology.about.com

36. Uso actual del Datamining



Muy útil en empresas que manejan gran cantidad de información, como compañías de seguro, bancos, etc.
Una de las aplicaciones mas comunes es en la evaluación de tarjetas de crédito, es factible colectar datos de los aplicantes los cuales servirán para poder determinar si son propietarios o no de una vivienda, el cantidad de años laborando, etc.; con la finalidad de poder emitir un juicio de comportamiento y compromiso con la tarjeta de crédito.

En la industria de ventas de artículos, es factible implementar una solución datamining, al asociar los diversos items que son vendidos o adquiridos en conjunto por los clientes. (Shorts deportivos y medias deportivas, generalmente son adquiridos al mismo tiempo.).

Autoevaluación

1. ¿Cuándo usar datamining?
2. Proponga un caso de datamining
3. ¿Qué es clasificación?

Para recordar

1. Las soluciones de datawarehouse deben complementarse con el datamining.
2. Datamining permite oportunidades de negocio y ventajas competitivas.
3. Conocer características de comportamiento de nuestros clientes en relación al negocio.