**DATAWAREHOUSE**

**Estudiante : Mayorga Patroni, César A.**

**Docente : Coronel Castillo, Eric Gustavo**

**Materia : Base de Datos**

INDICE

[**DATAWAREHOUSE** 1](#_Toc501485235)

[**I.** **Introducción** 3](#_Toc501485236)

[**II.** **Definición** 3](#_Toc501485237)

[**III.** **Objetivos del Data Warehouse** 3](#_Toc501485238)

[**IV.** **Características del Data Warehouse** 4](#_Toc501485239)

[**A.** **Orientado a temas** 4](#_Toc501485240)

[**B.** **Integrado** 5](#_Toc501485241)

[**C.** **De tiempo variante (Variable en el tiempo)** 7](#_Toc501485242)

[**D.** **No volátil** 8](#_Toc501485243)

[**V.** **Arquitectura del DataWarehouse** 8](#_Toc501485244)

[**A.** **El motor del dataWarehouse** 9](#_Toc501485245)

[**B.** **Herramientas de Acceso ó Herramientas de Usuario Final** 12](#_Toc501485246)

[**VI.** **Explotación del Data Warehouse** 13](#_Toc501485247)

[**A.** **Potencial del Data Warehouse** 13](#_Toc501485248)

[**B.** **Extracción y manipulación de datos del Data Warehouse** 14](#_Toc501485249)

[**C.** **Data Mining** 15](#_Toc501485250)

1. **Introducción**

La Revolución de la información, ha provocado cambios fundamentales en la elaboración y el uso de la información producidos a finales del siglo XX. La información dispuesta en el periódico y hasta la información almacenada en las bases de datos han permitido que la información esté al alcance de todos y la tecnología asociada a todos estos logros de nuestra sociedad han conducido a que esta información empiece a adquirir importancia estratégica.

1. **Definición**

Como definición inicial para la compresión del término, Data Warehouse se traduce al idioma español, como un almacén, depósito o bodega de datos, pero aplicando el término a la vida cotidiana.

El Data Warehouse es una arquitectura que Ilena por completo los diferentes requerimientos de acceso y reporte de datos. Para algunos otros, el Data Warehouse es un proceso continuo que mezcla los datos de varias fuentes heterogéneas, incluyendo datos históricos y adquiridos para soportar la constante necesidad de consultas estructuradas y/o ad hoc, reportes analíticos y soporte de decisiones.

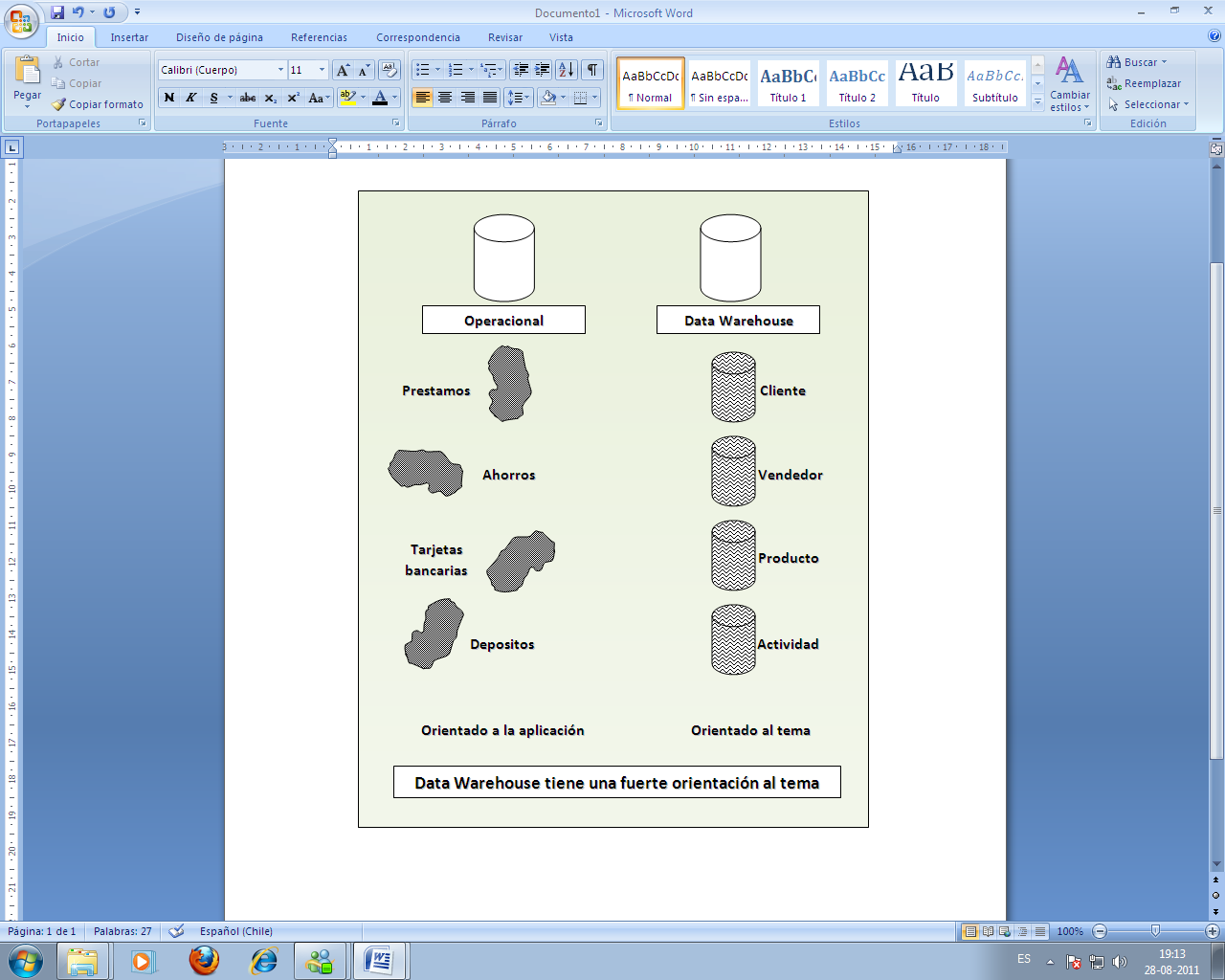
1. **Objetivos del Data Warehouse**

**Hacer que la información de la organización sea accesible:** los contenidos del Data Warehouse son entendibles y navegables, y el acceso a ellos son caracterizado por el rápido desempeño.

**La toma de decisiones:** el Data Warehouse tiene los datos correctos para soportar la toma de decisiones.

1. **Características del Data Warehouse**
2. **Orientado a temas**

Una primera característica del Data Warehouse es que la información se clasifica en base a los aspectos que son de interés para la empresa. Siendo así, los datos tomados están en contraste con los clásicos procesos orientados a las aplicaciones.

En la Figura se muestra el contraste entre los dos tipos de orientaciones. 

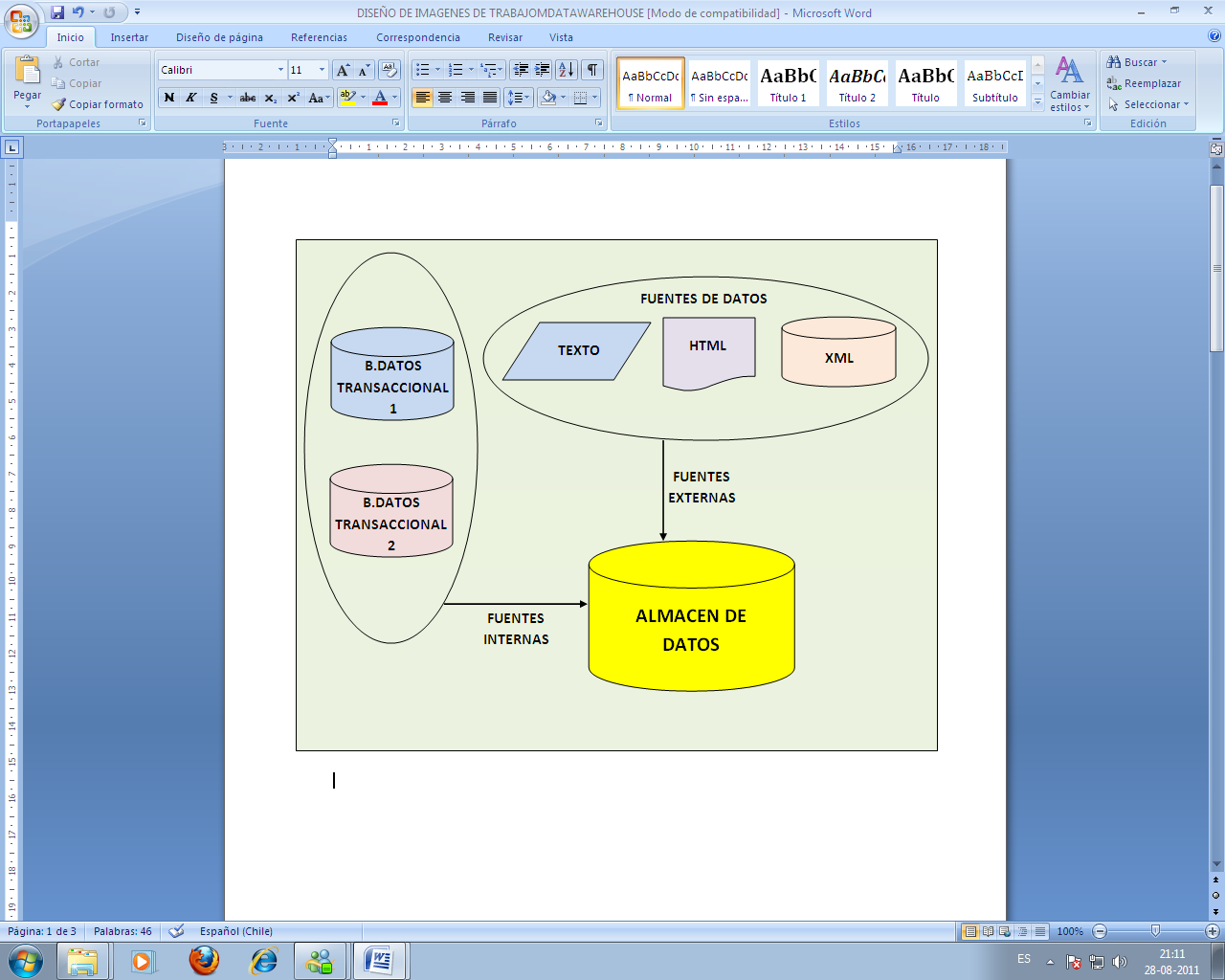
El ambiente operacional se diseña alrededor de las aplicaciones y funciones tales como préstamos, ahorros, tarjeta bancaria y depósitos para una institución financiera. Por ejemplo, una aplicación de ingreso de órdenes puede acceder a los datos sobre clientes, productos y cuentas. La base de datos combina estos elementos en una estructura que acomoda las necesidades de la aplicación.

La alineación alrededor de las áreas de los temas afecta el diseño y la implementación de los datos encontrados en el Data Warehouse. Las principales áreas de los temas influyen en la parte más importante de la estructura clave.

Las aplicaciones están relacionadas con el diseño de la base de datos y del proceso. En Data Warehousing se enfoca el modelamiento de datos y el diseño de la base de datos. El diseño del proceso (en su forma clásica) no es separado de este ambiente.

1. **Integrado**

Integra datos recolectados de diferentes sistemas operacionales de la organización y o fuentes externas.



El aspecto más importante del ambiente Data Warehousing es que la información encontrada al interior está siempre integrada.

La integración de datos se muestra de muchas maneras: en convenciones de nombres consistentes, en la medida uniforme de variables, en la codificación de estructuras consistentes, en atributos físicos de los datos consistentes, fuentes múltiples y otros.

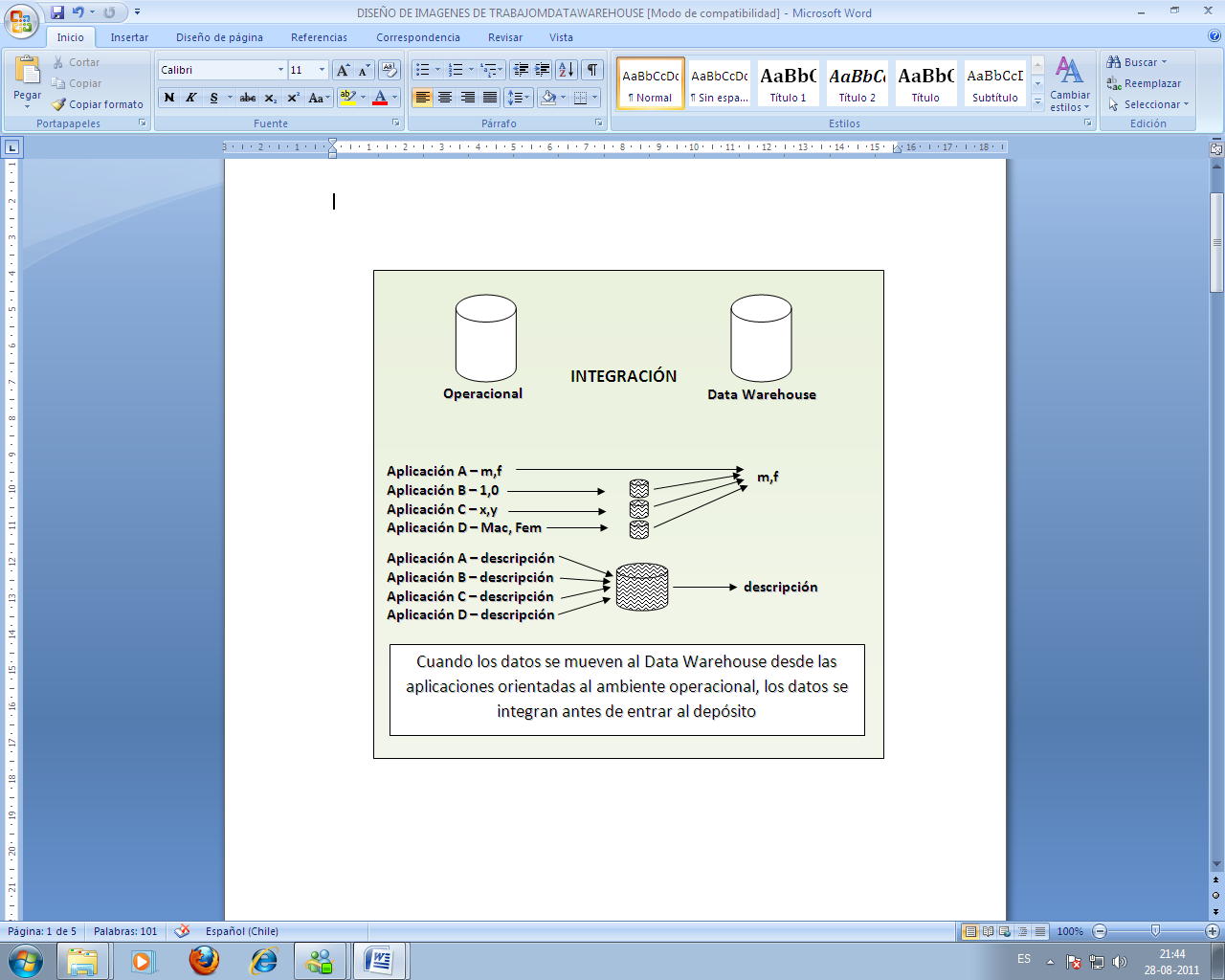
El contraste de la integración encontrada en el Data Warehouse con la carencia de integración del ambiente de aplicaciones, se muestran en la figura, con diferencias bien marcadas.

A través de los años, los diseñadores de las diferentes aplicaciones han tomado sus propias decisiones sobre cómo se debería construir una aplicación. Los estilos y diseños personalizados se muestran de muchas maneras.

**Convenciones de Nombramiento.** El mismo elemento es frecuentemente referido por nombres diferentes en las diversas aplicaciones. El proceso de transformación asegura que se use preferentemente el nombre de usuario.

**Fuentes Múltiples**. El mismo elemento puede derivarse desde fuentes múltiples. En este caso, el proceso de transformación debe asegurar que la fuente apropiada sea usada, documentada y movida al depósito.

Tal como se muestra en la figura, los puntos de integración afectan casi todos los aspectos de diseño - las características físicas de los datos, la disyuntiva de tener más de una de fuente de datos, el problema de estándares de denominación inconsistentes, formatos de fecha inconsistentes y otros.



1. **De tiempo variante (Variable en el tiempo)**

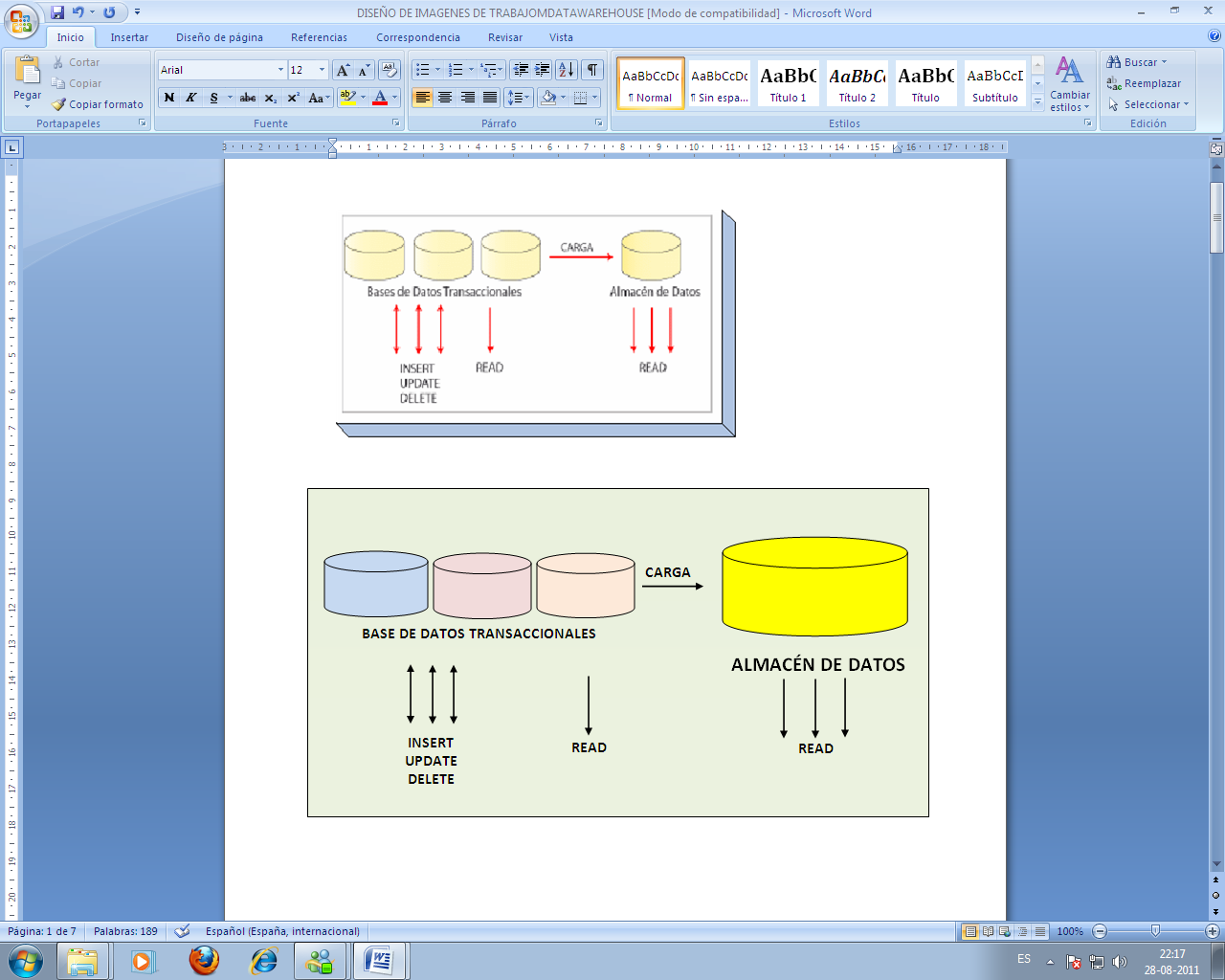
Los datos son relativos a un periodo de tiempo y estos deben ser integrados periódicamente, los mismos son almacenados como fotos que se corresponden a un periodo de tiempo.

Toda la información del data Warehouse es requerida en algún momento. Esta característica básica de los datos en un depósito, es muy diferente de la información encontrada en el ambiente operacional.

Como la información en el data Warehouse es solicitada en cualquier momento (es decir, no "ahora mismo"), los datos encontrados en el depósito se llaman de "tiempo variante".

1. **No volátil**

Los datos que son almacenados no sufren ninguna actualización solo son incrementados. El período cubierto para un DW va de 2 a 10 años.



La información es útil sólo cuando es estable. Los datos operacionales cambian sobre una base momento a momento. La perspectiva más grande, esencial para el análisis y la toma de decisiones, requiere una base de datos estable.

En la Figura se muestra que la actualización (insertar, borrar y modificar), se hace regularmente en el ambiente operacional sobre una base de registro por registro. Pero la manipulación básica de los datos que ocurre en el data Warehouse es mucho más simple. Hay dos únicos tipos de operaciones: la carga inicial de datos y el acceso a los mismos. No hay actualización de datos (en el sentido general de actualización) en el depósito, como una parte normal de procesamiento.

1. **Arquitectura del DataWarehouse**

El siguiente gráfico muestra la arquitectura clásica de un DataWarehouse, compuesto por:

* Fuentes de Datos
* Motor del DataWarehouse
  + Gestor de Carga
  + Metadatos
  + Agregaciones
  + Gestor del DataWarehouse
  + Gestor de Respaldos
  + DW Repositorio

1. **El motor del dataWarehouse**

Está integrado por los siguientes componentes

##### **Gestor de Carga**

Quizá sea uno de los elementos más importantes para el dataWarehouse, generalmente incluye las operaciones de

* **Extracción:** Es el proceso que accesa a los datos OLTP existentes, en cualquier forma que exista, desde cualquier DBMS en que exista. Típicamente, extracción y el siguiente paso, propagación, son administrados por el mismo producto. No todas las herramientas de extracción y propagación soportan todas las plataformas, de tal manera que una faceta importante de la selección de herramientas es si la herramienta soporta los sistemas operativos y las bases de datos que se esté usando para el dataWarehouse.
* **Propagación:** Es el proceso de mover datos desde los sistemas fuente hacia el sistema objetivo que contendrá el data Warehouse. El proceso de propagación toma lugar en tiempo real, o en un calendario predeterminado (batch), o sobre demanda, y puede efectuar un refresco total del Warehouse o justo un cambio neto. Cuando se selecciona una herramienta de propagación, se aspira que ésta ofrezca la gestión de cambios netos como también refresco total y permitirá tanto actualizaciones en tiempo real y calendarizadas (batch).
* **Depuración (Limpieza):** El nivel lógico cubre problemas de valores de datos que son inconsistentes dentro de la información importada (ejemplo, clientes con estado casado, pero con una edad de 3 años). El nivel técnico evalúa problemas de información tales como campos no inicializados o valores inválidos en los datos importados (ejemplo, valor de la fecha Febrero 31).
* **Transformación:** Convierte datos desde su formato OLTP al apropiado formato del dataWarehouse ejecutando funciones tales como desnormalización de datos, traduciendo códigos hacia texto significativo, convirtiendo una variedad de formatos de fechas hacia un formato estándar, convirtiendo texto tal como nombres de ciudades hacia texto estándar y renombrando campos desde nombres técnicos no significativos hacia nombres significativos que un usuario final entenderá.
* **Carga:** Los datos fuentes normalmente son extraídos y almacenados en archivos temporales tipo texto, los mismos que deben ser cargados a la base de datos del data Warehouse. La figura resume el proceso de carga, los archivos temporales finalmente son colocados en la base de data Warehouse de destino.

El módulo de Gestor de Carga también es conocido como Integrador, y es muy importante tanto en la Fase de Construcción como en la Fase de Explotación de un Data Warehouse.

FIG.18

##### **Meta Datos**

Esta área del Warehouse almacena todas las definiciones de los meta datos (datos acerca de los datos) usados por todos los procesos en el Warehouse. Los meta datos son usados para una variedad de propósitos incluyendo:

* Los procesos de extracción, transformación y carga (meta datos es usado para mapear las fuentes de datos a una vista común de la información dentro del Warehouse).
* Los procesos de gestión del Warehouse (cada tabla es descrita incluyendo su estructura, índices, vistas; meta datos es usado también para automatizar la producción de tablas resumen).
* Como parte de los procesos de gestión de consulta (meta datos es usado para dirigir una consulta a la fuente de datos más apropiada)

##### **Agregaciones**

Este componente del Warehouse almacena todos los datos agregados, predefinidos y generados por el gestor del Warehouse.

El propósito de información resumida es para mejorar el rendimiento de las consultas. Aunque hay costos operacionales incrementados asociados con la agregación inicial de los datos, esto debería ser compensado eliminando el requerimiento para ejecutar continuamente operaciones de agregación (tales como clasificación o agrupación) en las respuestas a las consultas de los usuarios. El dato agregado es actualizado continuamente en la medida que nuevos datos son cargados en al Warehouse.

##### **Gestor del DataWarehouse**

En algunos casos el gestor del Warehouse también genera perfiles de consultas para determinar qué índices y agregaciones son apropiadas. Un perfil de consulta puede ser generado para cada usuario, grupo de usuario, o el Data Warehouse y está basada en la información que describe las características de las consultas tales como la frecuencia, tablas objetivo, y tamaño de los results set.

##### **Gestor de Respaldos**

Es el componente que se encarga de respaldar constantemente la información del repositorio del Data Warehouse.

##### **Repositorio del Data Warehouse**

Es el repositorio en si o la base de datos física donde se almacena la información del Data Warehouse.

1. **Herramientas de Acceso ó Herramientas de Usuario Final**

Estas herramientas pueden reunirse en 4 grupos

* Herramientas de Reportes y Consultas
* Herramientas de Desarrollo de Aplicaciones
* Herramientas de Procesamiento Analítico en Línea (OLAP)
* Herramientas de Minería de Datos

##### **Herramientas de Reportes y Consultas**

Simplifican la generación de código SQL. Los usuarios formulan consultas a la base de datos sin tener que interactuar con el lenguaje de programación de bases de datos SQL).

##### **Herramientas de desarrollo de aplicaciones**

Los requerimientos de usuarios finales pueden ser tales que las capacidades pre-construidas de las herramientas de reportes y consultas son inadecuadas ya sea debido a que el análisis requerido no puede ser ejecutado o debido a que la interacción del usuario requiere un no razonable alto nivel de expertise por parte del usuario

En esta situación el acceso de usuarios puede requerir el desarrollo de aplicaciones usando ambientes gráficos de accedo de datos desarrollados principalmente para ambientes cliente/ servidor.

Algunas de estas plataformas de desarrollo de aplicaciones se integran con herramientas OLAP populares, y pueden accesar todos los sistemas de bases de datos principales, incluyendo Oracle, SQL Server, Sybase, e Informix. Ejemplos de estos ambientes de desarrollo de aplicaciones incluyen PowerBuilder de PowerSoft, VisualBasic de Microsoft, y Bussines Objects.

##### **Herramientas de Procesamiento Analítico en Línea (OLAP)**

Típicas aplicaciones de negocios para estas herramientas incluyen evaluación de la efectividad de una campaña de marketing, previsión de ventas de productos, y planificación de la capacidad. Estas herramientas asumen que los datos están organizados en un modelo multidimensional, el cual es soportado por una base de datos multidimensional especial (MDDBMS) o por una base de datos relacional diseñada para permitir consultas multidimensionales.

Este tipo de herramientas permiten a los usuarios ingresar a un Data Warehouse desde cualquier dimensión simple para empezar el análisis, luego navegar a otra dimensión para un mayor análisis de la información.

FIG.21

1. **Explotación del Data Warehouse**
2. **Potencial del Data Warehouse**

El Data Warehouse no produce resultados en forma mágica. Los administradores de empresas y los analistas deben acceder y recuperar los datos del Data Warehouse y convertirlos en información y en hechos. Estos hechos conforman los cimientos de una base de conocimientos que sirve para determinar la salud de la empresa y la dirección futura del negocio. Como en las granjas, los usuarios sólo cosecharán la información que se pueda derivar de los datos que sembraron en el Data Warehouse, y sólo mediante el uso de las herramientas de cosecha adecuadas. Algunas de las herramientas de cosecha necesarias son: las de acceso y recuperación, las de reportes de base de datos, las de análisis y las de data mining.

Uno de los retos al cosechar un Data Warehouse consiste en no convertir montículos de información en montañas de datos. Es fácil caer en la trampa de "entre más, mejor". No es esencial conocer todos los hechos, sólo los cruciales. Como ejemplo, una campaña de ropa para niños necesita cosechar exacta y rentablemente sólo aquellas familias que tienen niños.

1. **Extracción y manipulación de datos del Data Warehouse**

Herramientas de soporte de decisiones es el término genérico para referirse a las aplicaciones y herramientas del Data Warehouse que se emplean para recuperar, manipular y analizar los datos, y para presentar después los resultados. Estas herramientas se usan en dos modalidades: verificación y descubrimiento. En la modalidad de verificación, el usuario empresarial crea una hipótesis -una cuestión empresarial- e intenta confirmarla accediendo a los datos en el Data Warehouse. Las herramientas que implementan la modalidad de verificación son de consulta, de sistemas de reporte y de análisis multidimensional. En la modalidad de descubrimiento, las herramientas intentan descubrir características en los datos como patrones de compra o la asociación entre la adquisición de artículos diferentes. En la modalidad de descubrimiento, o eureka, el usuario empresarial no conoce ni sospecha los patrones y asociaciones descubiertos. La herramienta de Data Mining es un ejemplo de la modalidad de descubrimiento.

Desde la perspectiva de disponibilidad de herramientas, las dos modalidades de verificación y descubrimiento se clasifican en tres enfoques: Procesamiento Informático, Procesamiento Analítico y Data Mining.

1. **Data Mining**

El Data Mining apoya la modalidad de descubrimiento del soporte de decisiones. Las herramientas de Data Mining recorren los datos detallados de transacciones para desenterrar patrones y asociaciones ocultos. Por lo regular los resultados generan extensos reportes o se les analiza con herramientas de visualización de datos descubiertos.

El procesamiento informático es excelente y rentable para el despliegue masivo de consultas, análisis y reportes de datos de dos o tres dimensiones. Las herramientas de procesamiento analítico permiten diversas visualizaciones de los datos, como ventas por marca, tienda, temporada y periodos de tiempo, las cuales se pueden definir, consultar y analizar. Las herramientas de Data Mining son esenciales para comprender el comportamiento de los clientes.