UNIVERSIDAD NACIONAL DE LA AMAZONIA PERUANA

FACULTAD DE INGENIERÍA DE SISTEMAS

E INFORMÁTICA

"EL MODELO DATA WAREHOUSE – OLAP (ONLINE ANALYTICAL PROCESSING) LA MINERIA DE DATOS"
DE UNA EMPRESA EDITORIAL

INFORME DE TRABAJO PRACTICO DE SUFICIENCIA PARA OPTAR EL TITULO PROFESIONAL DE INGENIERO DE SISTEMAS E INFORMATICA

PRESENTADO POR LOS BACHILLERES

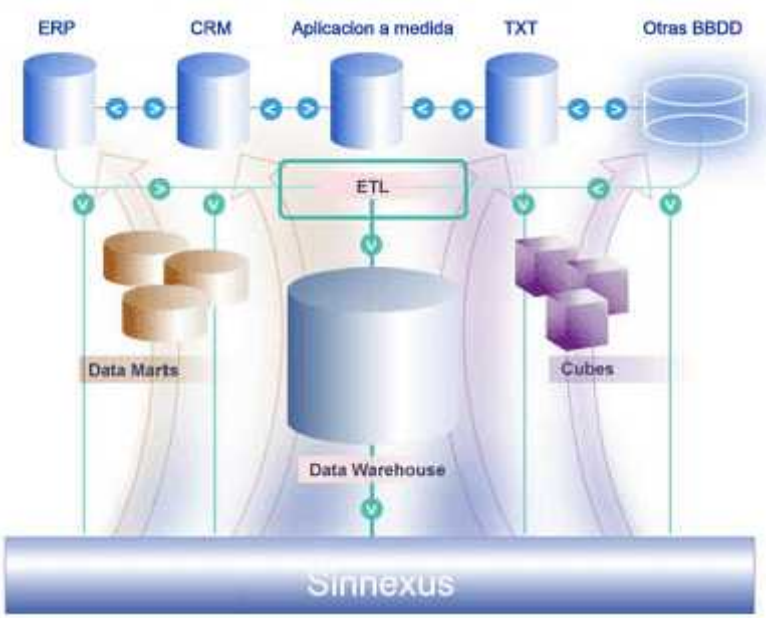
PEREA DOMPER GOLDMAN DENIS
TIBURCIO COLLANTES HUGO CLAY

ASESOR
ING. FRANCISCO MIGUEL RUIZ HDALGO

IQUITOS - PERÚ

2014

"EL MODELO DATA WAREHOUSE – OLAP (ONLINE ANALYTICAL PROCESSING) LA MINERIA DE DATOS"



RESUMEN

En este informe se presenta y describe un modelo general Olap y prototipo de un Sistema DataWareHouse para una empresa del sector público en general y se implementa en el caso Practico de una Editora, Se revisan los antecedentes, cómo se consolida la información actualmente de forma manual o con apoyo de otros sistemas, se define el problema, se muestra gráficamente la situación actual, se determina la justificación del presente trabajo y los métodos utilizados. Se detallan los objetivos generales y específicos; además se explica el concepto de Inteligencia de Negocios y Almacén de Datos. Se muestra el Modelo General OLAP de Empresa Editora, así como el prototipo desarrollado para mostrar parte de la solución al problema. Finalmente se expone las conclusiones, las recomendaciones y trabajos futuros.

INDICE

Justificación.

En el presente informe se trata de llegar a las razones fundamentales y entender el estado actual de la empresa, para poder implementar un Sistema de Data Warehouse, OLAP y minería de datos para un mejor desempeño, manejo, rapidez y seguridad de los datos.

Puntos principales de las deficiencias encontradas.

- La falta de información consolidad no permite tomar decisiones rápidas, solamente se cuenta con información mensual.
- El tiempo de procesamiento de la información ocasiona que no se cumpla con la entrega oportuna.
- No existe un solo repositorio de información dedicado exclusivamente a explotar la información histórica.
- No existe una herramienta de análisis que permita acceder a la información histórica.

También podemos decir al no tener un Data Warehouse hace que un reporte pueda tardar demasiado tiempo tanto que a lo mejor ya ni se necesite para cuando se complete o bien ya este obsoleto. Mientras que tener un Data Warehouse agiliza la generación de reporte y por ende la toma de decisiones.

Objetivo Generales.

Proponer un Modelo General OLAP en nuestro caso para Empresa Editora Perú S.A. que sirva como guía para la elaboración paulatina de los datamarts, que permitirán atender a futuro los requerimientos de información de las diferentes áreas de la empresa.

El contenido de un Data Warehouse debe ser comprensible, intuitivo y obvio para el usuario (Empresa Editora Peru S.A.). La comprensibilidad implica legibilidad, por lo que el contenido de Data Warehouse necesita ser etiquetado de manera significativa. El usuario de negocio debe estar habilitado para extraer porciones del Data Warehouse y cambiar esta información de todas las formas posibles, utilizando herramientas simples y fáciles de usar, con un tiempo de respuesta mínimo.

La información del Data Warehouse debe ser creíble. Los datos deben ser cuidadosamente reunidos de una variedad de orígenes de toda la organización, deben ser limpiados, con calidad asegurada, y liberados cuando sean aptos para el consumo del usuario.

Objetivos Específicos.

1. Realizar una herramienta que sirva de apoyo a la toma de decisiones y en las actividades de elaboración de informes.

2. Definir los requerimientos informáticos operativos y de desarrollo del modelo de Data Warehouse.

3. Fomentar el desarrollo de nuevas arquitecturas informáticas que contribuyan con los administradores a darle mayor confiabilidad a la información para la toma de decisiones.

4. Utilizar la experiencia y conocimientos adquiridos en la solución del problema para la implementación del Modelo OLAP.

III

INTRODUCCION.

En la actualidad, el dinámico mundo de los negocios plantea la necesidad de disponer de un acceso rápido y sencillo a información para la toma de decisiones. Dicha información debe estar estructurada y elaborada de acuerdo a parámetros de calidad, a fin de posibilitar una adaptación ágil y precisa a las fluctuaciones del ambiente externo.

La generación de reportes detallados, resumidos y comparativos son el medio más utilizado para explotar la información de un Sistema ERP.

Un Sistema ERP (Enterprise Resource Planning) es "Un paquete de software amplio integrado empresarial diseñado para mantener los más altos estándares de calidad de los procesos empresariales"[1].

Las empresas disponen, para la gestión de sus procesos de negocio, de sistemas transaccionales corporativos que manejan enormes cantidades de datos, organizados de forma tal que puedan ser utilizados por las aplicaciones operacionales existentes. Los niveles gerenciales necesitan a menudo tomar decisiones de alto nivel, cruciales para el funcionamiento de la empresa.

Frecuentemente se basan en su experiencia, utilizando un enfoque subjetivo del Proceso decisorio. Este enfoque no es apto para las condiciones del mundo actual en el que los sistemas de gestión de calidad vigentes han demostrado la importancia de la toma de decisiones basada en cifras, datos y hechos.

El Data Warehouse permite que los gerentes tomen decisiones siguiendo un enfoque racional, basados en información confiable y oportuna. Consiste básicamente en la transformación de los datos operacionales en información útil para decidir. El uso del Data Warehouse permite también encontrar relaciones ocultas entre los datos y predecir el comportamiento futuro bajo condiciones dadas.

La filosofía de trabajo del Data Warehouse es diferente a la de los sistemas transaccionales. Se modelan los datos a partir de dimensiones, en lugar del tradicional modelado relacional, y las herramientas de acceso a los datos se basan en una tecnología de procesamiento analítico (OLAP), distinta al procesamiento transaccional (OLTP) de los sistemas operacionales.

[1] Kwon 2001

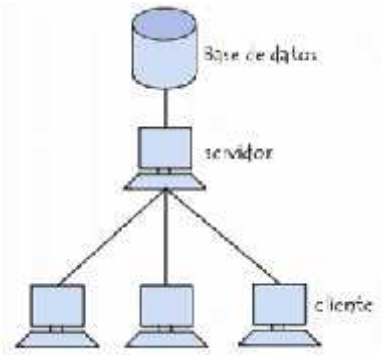
IV

1. CONCEPTOS GENERALES

1.1. Introducción a la Base de Datos (BD)

¿Qué es una Base de Datos?

Una base de datos (cuya abreviatura es *BD*) es una entidad en la cual se pueden almacenar datos de manera estructurada, con la menor redundancia posible. Diferentes programas y diferentes usuarios deben poder utilizar estos datos. Por lo tanto, el concepto de base de datos generalmente está relacionado con el de red ya que se debe poder compartir esta información. De allí el término **base**. "Sistema de información" es el término general utilizado para la estructura global que incluye todos los mecanismos para compartir datos que se han instalado.



Una base de datos proporciona a los usuarios el acceso a datos, que pueden visualizar, ingresar o actualizar, en concordancia con los derechos de acceso que se les hayan otorgado. Se convierte más útil a medida que la cantidad de datos almacenados crece.

Una base de datos puede ser local, es decir que puede utilizarla sólo un usuario en un equipo, o puede ser distribuida, es decir que la información se almacena en equipos remotos y se puede acceder a ella a través de una red.

La principal ventaja de utilizar bases de datos es que múltiples usuarios pueden acceder a ellas al mismo tiempo.

1.1.1. Componentes de una Base de Datos:

Hardware: constituido por dispositivo de almacenamiento como discos, tambores, cintas, etc.

- *Software:* que es el DBMS o Sistema Administrador de Base de Datos.
- *Datos:* los cuales están almacenados de acuerdo a la estructura externa y van a ser procesados para convertirse en información.

1.1.2. Tipos de Usuarios en Base de Datos

- **Usuario Final***:*
  Es la persona que utiliza los datos, esta persona ve datos convertidos en información:

- **Desarrollador de Aplicaciones***:*
  Es la persona que desarrolla los sistemas que interactuàn con la Base de Datos.

- **DBA***:*
  Es la persona que asegura integridad, consistencia, redundancia, seguridad este es el Administrador de Base de Datos quien se encarga de realizar el mantenimiento diario o periódico de los datos.

Las personas tienen acceso DBMS se clasifican de la siguiente manera:

- **Usuarios Ingenuos**:

  Son aquellos que interactúan con el sistema por medio de aplicaciones permanentes.

- **Usuarios Sofisticados**:

Son aquellos con la capacidad de acceder a la información por medios de lenguajes de consulta.

- **Programadores de Aplicación**:

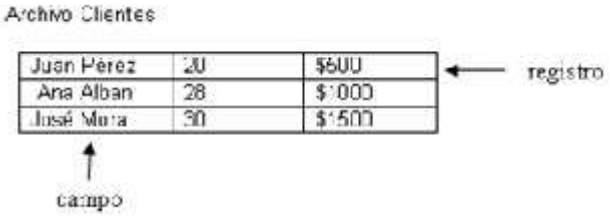
  Son aquellos con un amplio dominio del DML capaces de generar nuevos módulos o utilerías capaces de manejar nuevos datos en el sistema.

- **Usuarios Especializados**:

  Son aquellos que desarrollan módulos que no se refieren precisamente al manejo de los datos, si no a aplicaciones avanzadas como sistemas expertos, reconocimientos de imágenes, procesamiento de audio y demás.

1.1.3. Conceptos Básicos de Base de datos

Archivo: son conjuntos de registros.
Registros*:* son conjuntos de campos.
Campos*:* es la mínima unidad de referencia



1.1.4. Administración de Base de Datos

Rápidamente surgió la necesidad de contar con un sistema de administración para controlar tanto los datos como los usuarios.

Los **Sistemas Gestores de Bases de Datos** son un tipo de software muy específico, dedicado a servir de interfaz entre las bases de datos y las aplicaciones que la utilizan. Se compone de un lenguaje de definición de datos, de un lenguaje de manipulación de datos y de un lenguaje de consulta. En los textos que tratan este tema, o temas relacionados, se mencionan los términos SGBD y DBMS, siendo ambos equivalentes, y acrónimos, respectivamente, de Sistema Gestor de Bases de Datos y *DataBase Management System*, su expresión inglesa.

1.1.5. **Integridad de Datos***:*

Conjunto de seguridades que son utilizadas para mantener los datos correctos. Ocurre cuando no existe a través de todo el sistema procedimientos uniformes de validación para los datos.

Los principales sistemas de administración de bases de datos son:[2]

[2] http://es.kioskea.net/contents/66-introduccion-bases-de-datos

- Borland Paradox
- Filemaker
- IBM DB2
- Ingres
- Interbase
- Microsoft SQL server
- Microsoft Access
- Microsoft FoxPro
- Oracle
- Sybase
- MySQL
- PostgreSQL
- mSQL
- SQL Server 11

1.2. Términos Para la Inteligencia De Negocios.

La Inteligencia de Negocios o Business Intelligence (BI) se puede definir como el proceso de analizar los bienes o datos acumulados en la empresa y extraer una cierta inteligencia o conocimiento de ellos. Dentro de la categoría de bienes se incluyen las bases de datos de clientes, información de la cadena de suministro, ventas personales y cualquier actividad de marketing o fuente de información relevante para la empresa.

BI apoya a los tomadores de decisiones con la información correcta, en el momento y lugar correcto, lo que les permite tomar mejores decisiones de negocios. La información adecuada en el lugar y momento adecuado incrementa efectividad de cualquier empresa.

Con BI se puede:

- generar reportes globales o por secciones
- crear una base de datos de clientes
- crear escenarios con respecto a una decisión
- hacer pronósticos de ventas y devoluciones
- compartir información entre departamentos
- análisis multidimensionales
- generar y procesar datos
- cambiar la estructura de toma de decisiones
- mejorar el servicio al cliente.

1.3. DataWarehouse

Un **almacén de datos** (del inglés *data warehouse*) es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza.

El ingreso de datos en el Data Warehouse viene desde el ambiente operacional en casi todos los casos. El Data Warehouse es siempre un almacén de datos transformados y separados físicamente de la aplicación donde se encontraron los datos en el ambiente operacional.

Data warehousing es el centro de la arquitectura para los sistemas de información en la década de los '90. Soporta el procesamiento informático al proveer una plataforma sólida, a partir de los datos históricos para hacer el análisis. Facilita la integración de sistemas de aplicación no integrados. Organiza y almacena los datos que se necesitan para el procesamiento analítico, informático sobre una amplia perspectiva de tiempo.

Un Data Warehouse o Depósito de Datos es una colección de datos orientado a temas, integrado, no volátil, de tiempo variante, que se usa para el soporte del proceso de toma de decisiones gerenciales.

Se puede caracterizar un data warehouse haciendo un contraste de cómo los datos de un negocio almacenados en un data warehouse, difieren de los datos operacionales usados por las aplicaciones de producción.

1.3.1. Características

- **Orientada al negocio**

La primera característica del DW, es que la información se clasifica en base a los aspectos que son de interés para la empresa. Esta clasificación afecta el diseño y la implementación de los datos encontrados en el almacén de datos, debido a que la estructura del mismo difiere considerablemente a la de los clásicos procesos operacionales orientados a las aplicaciones.

- **Integrada**

La integración implica que todos los datos de diversas fuentes que son producidos por distintos departamentos, secciones y aplicaciones, tanto internas como externas, deben ser consolidados en una instancia antes de ser agregados al DW. A este proceso se lo conoce como Extracción, Transformación y Carga de Datos2 (Extraction, Transformation and Load - ETL).

La integración de datos, resuelve diferentes variados tipos de problemas relacionados con las convenciones de nombres, unidades de medidas, codificaciones, fuentes múltiples, etc, cada uno de los cuales será correctamente detallado y ejemplificado más adelante.

Esto se debe a que a través de los años los diseñadores y programadores no se han basado en ningún estándar para definir nombres de variables, tipos de datos, etc, ya sea por carecer de ellos o por no creer que sean necesarios. Por lo cual, cada uno por su parte ha dejado en cada aplicación, módulo, tabla, etc, su propio estilo personalizado, confluyendo de esta manera en la creación de modelos muy inconsistentes e incompatibles entre sí.

Los puntos de integración afectan casi todos los aspectos de diseño, y cualquiera sea su forma, el resultado es el mismo, ya que la información será almacenada en el DW en un modelo globalmente aceptable y singular, aún cuando los sistemas operacionales y demás fuentes almacenen los datos de maneras disímiles, para que de esta manera el usuario final este enfocado en la utilización de los datos del depósito y no deba cuestionarse sobre la confiabilidad o solidez de los mismos.

- **Variante en el tiempo**

Debido al gran volumen de información que se manejará en el DW, cuando se le realiza una consulta, los resultados deseados demorarán en originarse. Este espacio de tiempo que se produce desde la búsqueda de datos hasta su consecución es del todo normal en este ambiente y es, precisamente por ello, que la información que se encuentra dentro del depósito de datos se denomina de tiempo variable.

Esta característica básica, es muy diferente de la información encontrada en el ambiente Operacional, en el cual, los datos se requieren en el momento de acceder, es decir, que se espera que los valores procurados se obtengan a partir del momento mismo de acceso.

Además, toda la información en el DW posee su propio sello de tiempo:

Esto contribuye a una de las principales ventajas del almacén de datos: los datos son almacenados junto a sus respectivos históricos. Esta cualidad que no se encuentra en fuentes de datos operacionales, garantiza poder desarrollar análisis de la dinámica de la información, pues ella es procesada como una serie de instantáneas, cada una representando un periodo de tiempo. Es decir, que gracias al sello de tiempo se podrá tener acceso a diferentes versiones de la misma información.

Es importante tener en cuenta la granularidad3 de los datos, así como también la intensidad de cambio natural del comportamiento de los fenómenos del negocio, para evitar crecimientos incontrolables y desbordamientos de la base de datos.

El intervalo de tiempo y periodicidad de los datos debe definirse de acuerdo a la necesidad y requisitos de los usuarios.

Es elemental aclarar, que el almacenamiento de datos históricos, es lo que permite al DW desarrollar pronósticos y análisis de tendencias y patrones, a partir de una base estadística de información, ya que las instantáneas son actualizadas de acuerdo con las actividades del negocio.

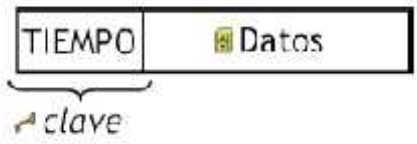


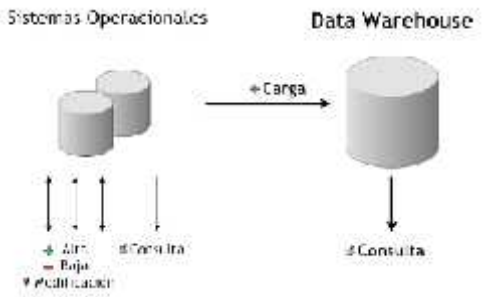
Figura 2.2

- **No volátil**

La información es útil para el análisis y la toma de decisiones solo cuando es estable. Los datos operacionales varían momento a momento, en cambio, los datos una vez que entran en el DW no cambian.

La actualización, o sea, insertar, eliminar y modificar, se hace de forma muy habitual en el ambiente operacional sobre una base, registro por registro, en cambio en el depó- sito de datos la manipulación básica de los datos es mucho más simple, debido a que solo existen dos tipos de operaciones: la carga de datos y el acceso a los mismos.

Por esta razón es que en el DW no se requieren mecanismos de control de la concurrencia y recuperación.



- **Cualidades**

Una de las primeras cualidades que se puede mencionar del DW, es que maneja un gran volumen de datos, debido a que consolida en su estructura la información recolectada durante años, proveniente de diversas fuentes, en un solo lugar centralizado. Es por esta razón que el depósito puede ser soportado y mantenido sobre diversos medios de almacenamiento.

Además, como ya se ha mencionado, el almacén de datos presenta la información sumarizada y agregada desde múltiples versiones, y maneja información histórica.

Organiza y almacena los datos que se necesitan para el procesamiento analítico e informático, con el propósito de responder a preguntas de negocios y brindarles a los usuarios finales una interface amigable, comprensible y fácil de utilizar, para que los mismos puedan tomar decisiones sobre los datos sin tener que poseer demasiados conocimientos informáticos. El DW permite un acceso más directo, es decir, la información gira en torno al negocio, y es por ello que también los usuarios pueden sentirse cómodos al explorar los datos y encontrar relaciones complejas entre los mismos.

El DW no es solo datos, sino un conjunto de herramientas para consultar, analizar y presentar información, que permiten obtener o realizar análisis, reporting, extracción y explotación de los datos, con alta performance, para transformar dichos datos en información valiosa para la organización.

Con respecto a las tecnologías empleadas, en un almacén de datos se pueden encontrar las siguientes:

- Arquitectura cliente/servidor.
- Técnicas avanzadas para replicar, refrescar y actualizar datos.
- Software front-end, para acceso y análisis de datos.
- Herramientas para extraer, transformar y cargar datos en el depósito, desde múltiples fuentes muy heterogéneas.
- Sistema de Gestión de Base de Datos4 (SGBD).

Cabe destacar, que todas las cualidades expuestas anteriormente, son imposibles de saldar en un típico ambiente operacional, y esto es una de las razones de ser del DW.

1.3.2. Ventajas

A continuación se enumerarán algunas de las ventajas más sobresalientes que trae aparejada la implementación de un DW y que ejemplifican de mejor modo sus características y cualidades:

- Transforma datos orientados a las aplicaciones en información orientada a la toma de decisiones.
- Integra y consolida diferentes fuentes de datos y departamentos empresariales, que anteriormente formaban islas, en una única plataforma sólida y centralizada.
- Provee la capacidad de analizar y explotar las diferentes áreas de trabajo y de realizar un análisis inmediato de las mismas.
- Permite reaccionar rápidamente a los cambios del mercado.
- Aumenta la competitividad en el mercado.
- Elimina la producción y el procesamiento de datos que no son utilizados ni necesarios, producto de aplicaciones mal diseñadas o ya no utilizadas.
- Mejora la entrega de información, es decir, información completa, correcta, consistente, oportuna y accesible. Información que los usuarios necesitan, en el momento adecuado y en el formato apropiado.
- Logra un impacto positivo sobre los procesos empresariales. Cuando los usuarios tienen acceso a una mejor calidad de información, la empresa puede lograr por sí misma: aprovechar el enorme valor potencial de sus recursos de información y transformarlo en valor verdadero; eliminar los retardos de los procesos empresariales que resultan de información incorrecta, inconsistente y/o inexistente; integrar y optimizar procesos a través del uso compartido e integrado de las fuentes de información; permitir al usuario adquirir mayor confianza acerca de sus propias decisiones y de las del resto, y lograr así, un mayor entendimiento de los impactos ocasionados.
- Aumento de la competitividad de los encargados de tomar decisiones.
- Los usuarios pueden acceder directamente a la información en línea, lo que contribuye a su capacidad para operar con mayor efectividad en las tareas rutinarias o no. Además, pueden tener a su disposición una gran cantidad de valiosa información multidimensional, presentada coherentemente como fuente única, confiable y disponible en sus estaciones de trabajo. Así mismo, los usuarios tienen la facilidad de contar con herramientas que les son familiares para manipular y evaluar la información obtenida en el DW, tales como: hojas de cálculo, procesadores de texto, software de análisis de datos, software de análisis estadístico, reportes, etc.
- Permite la toma de decisiones estratégicas y tácticas.

1.3.3. Desventajas

A continuación se consignarán algunas de las desventajas encontradas en la implementación de un DW:

- Requiere una gran inversión, debido a que su correcta construcción no es tarea sencilla y consume muchos recursos, además, su misma implementación implica desde la adquisición de herramientas de consulta y análisis, hasta la capacitación de los usuarios.
- Existe resistencia al cambio por parte de los usuarios.
- Los beneficios del almacén de datos son apreciados en el mediano y largo plazo. Este punto deriva del anterior, y básicamente se refiere a que no todos los usuarios confiarán en el DW en una primera instancia, pero sí lo harán una vez que comprueben su efectividad y ventajas. Además, su correcta utilización surge de la propia experiencia.
- Si se incluyen datos propios y confidenciales de clientes, proveedores, etc, el depósito de datos atentará contra la privacidad de los mismos, ya que cualquier usuario podrá tener acceso a ellos.
- Infravaloración de los recursos necesarios para la captura, carga y almacenamiento
- de los datos.
- Infravaloración del esfuerzo necesario para su diseño y creación.

Incremento continuo de los requerimientos del usuario.

1.4. Tecnología OLAP.

OLAP es el acrónimo en inglés de procesamiento analítico en línea (On-Line Analytical Processing). Es una solución utilizada en el campo de la llamada Inteligencia empresarial (o Business Intelligence) cuyo objetivo es agilizar la consulta de grandes cantidades de datos. Para ello utiliza estructuras multidimensionales (o Cubos OLAP) que contienen datos resumidos de grandes Bases de datos o Sistemas Transaccionales (OLTP). Se usa en informes de negocios de ventas, marketing, informes de dirección, minería de datos y áreas similares.

1.4.1. Funcionalidad

En la base de cualquier sistema OLAP se encuentra el concepto de cubo OLAP (también llamado cubo multidimensional o hipercubo). Se compone de hechos numéricos o **medidas**, que se clasifican por dimensiones. El cubo de metadatos es típicamente creado a partir de un esquema en estrella o copo de nieve, esquema de las tablas en una base de datos relacional. Las medidas se obtienen de los registros de una tabla de hechos y las dimensiones se derivan de la dimensión de los cuadros.

La razón de usar OLAP para las consultas es la rapidez de respuesta. Una base de datos relacional almacena entidades en tablas discretas si han sido normalizadas. Esta estructura es buena en un sistema OLTP pero para las complejas consultas multitabla es relativamente lenta. Un modelo mejor para búsquedas (aunque peor desde el punto de vista operativo) es una base de datos multidimensional.

La principal característica que potencia a OLAP, es que es lo más rápido a la hora de ejecutar sentencias SQL de tipo SELECT, en contraposición con OLTP que es la mejor opción para operaciones de tipo INSERT, UPDATE Y DELETE.

2. Como Se Definen Los Requerimientos Y Los Procesos De Negocio Para Modelar un DataWarehouse Con Base En La Estrategia En La Organización

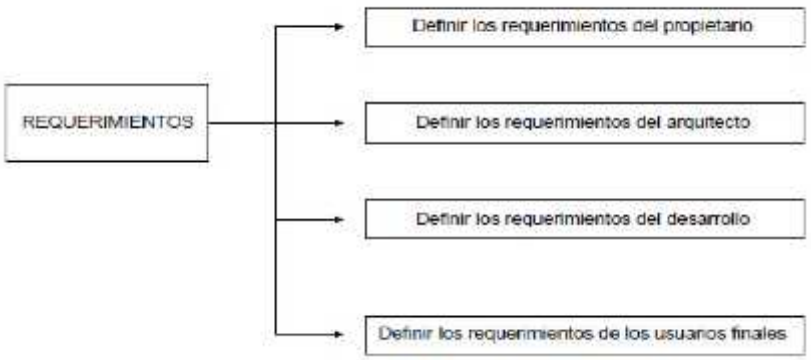
- **Definición de los requerimientos del negocio.**

Antes de plantear el data warehouse, debemos encontrar en la empresa nuestro sistema nervioso digital. ¿Y qué es el sistema nervioso digital?, es como el sistema nervioso humano, en el cual todas sus partes se interrelacionan haciendo que la empresa funcione productivamente controlando sus procesos. Por ello un sistema nervioso digital transforma 3 elementos de un negocio:

- Su relación con los clientes y asociado (comercio).
- El flujo de información y la relación entre sus empleados (administración del conocimiento).
- Procesos de negocios internos (operaciones de negocios).

2.1. Definición de Los Requerimientos de Negocio (Alto Nivel).

La fase en mención es una especificación precisa de las funciones que se obtendrán del data warehouse, para ello se debe definir los requerimientos que se necesitará, como se muestra en la Figura.



Requerimientos para la solución de un Datawarehouse.

2.2. Prioridad En Los Procesos.

Esta fase significa convertir los requerimientos agrupados en un conjunto de especificaciones que puedan apoyar el diseño. En este análisis debe considerarse 3 tipos de especificaciones:

1. Requerimientos de enfoque empresarial que delinean las fronteras de la información que debe comprender el data warehouse. El enfoque empresarial determinará también la audiencia y sus requerimientos de información.
2. Especificación de requerimientos de fuentes de datos que delinean las fronteras de información disponible en las fuentes de datos actuales.
3. Especificaciones de requerimientos de usuario final y acceso, las cuales definen cómo se utilizará la información del data warehouse. Junto con éstas se encuentra la especificación de los tipos de herramientas y técnicas de exhibición que se usarán.

Luego del análisis viene el diseño que tiene dos actividades principales:

- Diseño detallado de la arquitectura de datos: Es el desarrollo del modelo físico de datos para la base de datos de almacenamiento del datwarehouse y mercado de datos.
- Diseño detallado de la arquitectura de aplicaciones: Es la Correspondencia de los modelos físicos de datos de la fuente de datos con los modelos físicos data warehouse y mercado de datos.

2.3. Elementos en la Planeación del Proyecto.

La Figura muestra la planeación que se tiene que realizar en un datawarehouse. Algunos de los pasos se pueden efectuar al mismo tiempo (en paralelo), lo cual acorta la duración de esta fase.

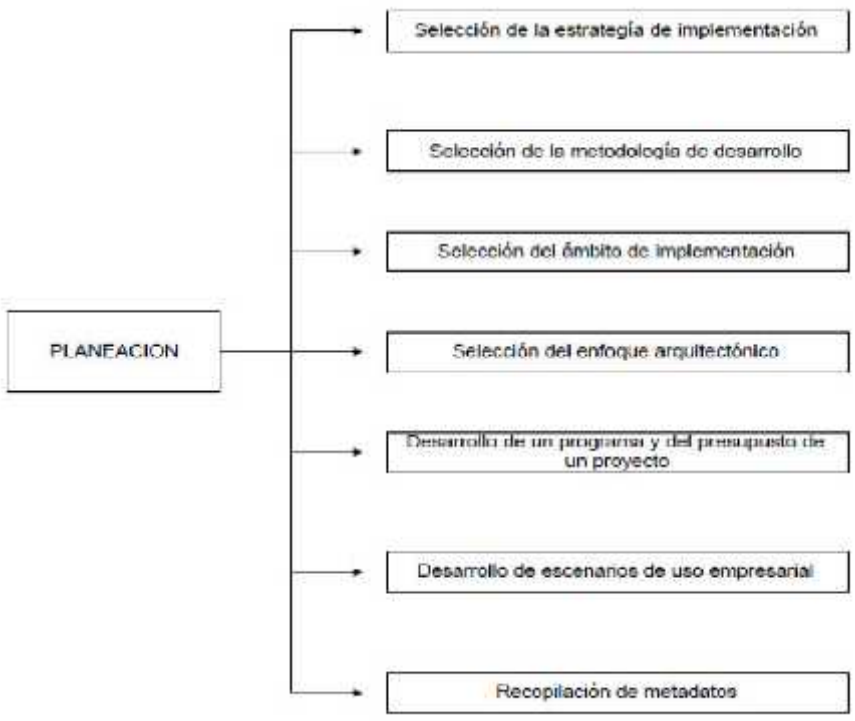


Fig. Planeación necesaria para el sistema del DatawareHouse

Uno de los primeros pasos más importantes consiste en decidir la estrategia general de implementación. La decisión tiene mucho que ver con la cultura de la organización y se basa en cómo se llevan a cabo las tareas dentro de la organización.
Se debe tener en cuenta la metodología a utilizar, las más conocidas son: Método en Cascada y Método Espiral, se define el método arquitectónico, el desarrollo del programa y los escenarios que la empresa va a tener cuando se implemente el datawarehouse, para ello se define claramente los metadatos, que son elementos que se va a utilizar para la planeación efectiva del datawarehouse

- **Construcción**

En esta fase se realiza la implementación física de los diseños desarrollados durante la fase de diseño. Las aplicaciones que se necesitan construir son las siguientes:

- Programas que creen y modifiquen la base de datos para el datawarehouse.
- Programas que traigan datos de fuentes relacionadas y no relacionadas.
- Programas que realicen transformación de datos.
- Programas que realicen actualización de base de datos.
- Programas que efectúen búsquedas en base de datos muy grandes.

- **Despliegue**

Los requerimientos de despliegue para un data warehouse son:

- La información contenida en el data warehouse debe estar en términos y lenguajes que comprendan los usuarios ya que ellos no son técnicos.
- Debe existir una necesidad de que la información que proporcione el data warehouse debe de ser precisa para los usuarios finales.

- **Expansión**

En esta etapa se prevé algunas de las siguientes áreas de mejora:

- Consultas empresariales que no pueden formularse o satisfacerse debido a la limitación del data warehouse.
- Consultas empresariales que comprenden fuente de datos externas que no formaron parte de la implementación inicial.
- Desempeño no satisfactorio de componentes del data warehouse

3. Integración De Datos

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 N° 2, 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.

Se diferencian en la codificación, en las estructuras claves, en sus características físicas, en las convenciones de nombramiento y otros. La capacidad colectiva de muchos de los diseñadores de aplicaciones, para crear aplicaciones inconsistentes, es fabulosa. La Figura N° 2 mencionada, muestra algunas de las diferencias más importantes en las formas en que se diseñan las aplicaciones.

3.1. Los Procesos Extract Tranform and Load (ETL)

Extract, Transform and Load («extraer, transformar y cargar», frecuentemente abreviado **ETL**) es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, data mart, o data warehouse para analizar, o en otro sistema operacional para apoyar un proceso de negocio.

Los procesos *ETL* también se pueden utilizar para la integración con sistemas heredados.

- **Extraer.**

La primera parte del proceso ETL consiste en extraer los datos desde los sistemas de origen. La mayoría de los proyectos de almacenamiento de datos fusionan datos provenientes de diferentes sistemas de origen. Cada sistema separado puede usar una organización diferente de los datos o formatos distintos. Los formatos de las fuentes normalmente se encuentran en bases de datos relacionales o ficheros planos, pero pueden incluir bases de datos no relacionales

u otras estructuras diferentes. La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación.

Una parte intrínseca del proceso de extracción es la de analizar los datos extraídos, de lo que resulta un chequeo que verifica si los datos cumplen la pauta o estructura que se esperaba. De no ser así los datos son rechazados.

Un requerimiento importante que se debe exigir a la tarea de extracción es que ésta cause un impacto mínimo en el sistema origen. Si los datos a extraer son muchos, el sistema de origen se podría ralentizar e incluso colapsar, provocando que éste no pueda utilizarse con normalidad para su uso cotidiano. Por esta razón, en sistemas grandes las operaciones de extracción suelen programarse en horarios o días donde este impacto sea nulo o mínimo.

- **Transformar**

La fase de transformación aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados. Algunas fuentes de datos requerirán alguna pequeña manipulación de los datos. No obstante en otros casos pueden ser necesarias aplicar algunas de las siguientes transformaciones:

- Seleccionar sólo ciertas columnas para su carga (por ejemplo, que las columnas con valores nulos no se carguen).
- Traducir códigos (por ejemplo, si la fuente almacena una "H" para Hombre y "M" para Mujer pero el destino tiene que guardar "1" para Hombre y "2" para Mujer).
- Codificar valores libres (por ejemplo, convertir "Hombre" en "H" o "Sr" en "1").
- Obtener nuevos valores calculados (por ejemplo, total_venta = cantidad * precio, o Beneficio = PVP - Coste).
- Unir datos de múltiples fuentes (por ejemplo, búsquedas, combinaciones, etc.).
- Calcular totales de múltiples filas de datos (por ejemplo, ventas totales de cada región).
- Generación de campos clave en el destino.
- Transponer o pivotar (girando múltiples columnas en filas o viceversa).

- Dividir una columna en varias (por ejemplo, columna "Nombre: García López, Miguel Ángel"; pasar a dos columnas "Nombre: Miguel Ángel", "Apellido1: García" y "Apellido2: López").
- La aplicación de cualquier forma, simple o compleja, de validación de datos, y la consiguiente aplicación de la acción que en cada caso se requiera:
- Datos OK: Entregar datos a la siguiente etapa (Carga).
- Datos erróneos: Ejecutar políticas de tratamiento de excepciones (por ejemplo, rechazar el registro completo, dar al campo erróneo un valor nulo o un valor *centinela*).

- **Carga**

La fase de carga es el momento en el cual los datos de la fase anterior (**transformación**) son cargados en el sistema de destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes. En algunas bases de datos se sobrescribe la información antigua con nuevos datos. Los **data warehouse** mantienen un historial de los registros de manera que se pueda hacer una auditoría de los mismos y disponer de un rastro de toda la historia de un valor a lo largo del tiempo.

Existen dos formas básicas de desarrollar el proceso de carga:

3.1.1. Herramientas ETL

- Los volúmenes de datos en el sistema se incrementan de forma exponencial, y se deben procesar y granular grandes cantidades de datos (productos vendidos, llamadas telefónicas, transacciones bancarias) para su posterior auditoría.
- Se fusionan empresas y se debe hacer un matching de los datos de los distintos sistemas legacy para generar una única salida de datos equivalentes.
- La inteligencia de negocio requiere que se almacenen las estructuras de datos históricos en aplicaciones OLAP, para el análisis, notificación y cuadros de mando operacional y táctico (dashboarding) y estratégico (scorecarding).

Tipos de Herraminetas.

- XMLoader
- CoSort+Fact por IRI http://www.iri.com/
- SynchroDB https://synchrodb.com
- Pentaho Data Integration

- SAS Data Integration Studio & DataFlux http://www.sas.com/technologies/dw/index.html
- Ab Initio
- Barracuda Software (Integrator)
- MakeWare Soluciones Tecnológicas http://www.makeware.net
- Benetl
- Biable http://www.visiontecnologica.com
- BITool - ETL Software http://www.bitool.com/
- BOPOS TLOG-4690 rhiscom (back-office POS)
- CloverETL [1]
- Cognos Decisionstream
- Data Integrator (herramienta de Business Objects)
- Data Migraton Toolset de Backoffice Associates (BoA) http://www.boaweb.com/migrationtoolset.htm
- Genio, Hummingbird
- IBM Websphere DataStage (Previously Ascential DataStage)
- Microsoft DTS (incluido en SQL-Server 2000)
- Microsoft SQL Server Integration Services (SSIS) (a partir de MS SQL Server 2005)
- MySQL Migration Toolkit
- Scriptella ETL - Libre, Apache-licensed ETL
- Oracle Warehouse Builder
- WebFocus-iWay DataMigrator Server

Con licencia libre

- Kettle (Pentaho)
- Scriptella Open Source ETL Tool
- Talend Open Studio
- CloverETL Community

Con licencia propietaria.

- Benetl (freeware)
- Datastage d'IBM (suite au rachat d'Ascential en 2005)
- Integrator
- DataStudio

- Informática PowerCenter
- Oxio Data Intelligence ETL full web
- SmartDB Workbench
- Sunopsis

3.2. La Limpieza De Los Datos

El ***data cleansing***, ***data scrubbing*** o **limpieza de datos**, es el acto de descubrimiento, corrección o eliminación de datos erróneos de una base de datos. El proceso de data cleansing permite identificar datos incompletos, incorrectos, inexactos, no pertinentes, etc. y luego substituir, modificar o eliminar estos datos sucios ("data duty"). Después de la limpieza, la base de datos podrá ser compatible con otras bases de datos similares en el sistema.

Las inconsistencias descubiertas, modificadas o eliminadas pueden haber sido causado por: las definiciones de diccionario de datos diferentes de entidades similares, errores de entrada del usuario y corrupción en la transmisión o el almacenaje.

La Limpieza de datos se diferencia de la validación de datos ("data validation"), en que la validación de datos cumple la función de rechazar los registros erróneos durante la entrada al sistema. El proceso de data cleansing incluye la validación y además la corrección de datos, para alcanzar datos de calidad ("Data quality").

- **Pasos para la limpieza de Datos.**

1. Analizar sus datos corporativos para descubrir inexactitudes, anomalías y otros problemas.
2. Transformar los datos para asegurar que sean precisos y coherentes.
3. Asegurar la integridad referencial, que es la capacidad del data warehouse, para identificar correctamente al instante cada objeto del negocio, tales como un producto, un cliente o un empleado.
4. Validar los datos que usa la aplicación del data warehouse para realizar las consultas de prueba.
5. Producir la metadata, una descripción del tipo de datos, formato y el significado relacionado al negocio de cada campo.
6. Finalmente, viene el paso crucial de la documentación del proceso completo para que se pueda ampliar, modificar y arreglar los datos en el futuro con más facilidad.
7. En la práctica, se tendría que realizar múltiples pasos como parte de una operación única o cuando use una sola herramienta. En particular, limpiar la data y asegurar la integridad referencial son procesos interdependientes.

Las herramientas comerciales pueden ayudar en cada uno de estos pasos. Sin embargo, es posible escribir sus propios programas para hacer el mismo trabajo.

Los programas de limpieza de datos no proporcionan mucho razonamiento, por lo que las compañías necesitan tomar sus decisiones en forma manual, basados en información importante y reportes de auditoría de datos.

Cada vez que se carga un nuevo conjunto de datos, la limpieza de datos comúnmente constituye cerca del 25 por ciento de lo que puede ser un proceso de cuatro semanas.

A continuación, se darán algunos ejemplos de las experiencias de las empresas que han realizado limpieza de datos para un ambiente data warehousing.

Ejemplo:

CompuCom Systems, un gran integrador de sistemas basados en Dallas, implementó un registro de 12 millones, en un depósito de 10 Gb para el soporte de decisiones internas y de los clientes, según el orden y la condición y producir información por medio del Web.

CompuCom implementó algunas rutinas de mejoramiento de datos en lenguajes de cuarta generación (4GL), asociado con su base de datos Progress, la cual corre sobre un HP 9000. El incremento incluye desciframiento de valores de columnas en descripciones inglesas cortas o mnemotecnia. El código de limpieza de datos, tales como las conversiones de fecha y datos, están escritas en lenguaje C.

La ventaja de ésto es que CompuCom ahora posee estas rutinas y puede usarlas en otras aplicaciones.

Los usuarios ayudaron a definir los requerimientos de limpieza de datos, ya que son ellos los que mejor conocen los datos y pueden informar sobre qué tipo de datos sucios deben salir y cómo limpiarlos.

La compañía no usa una herramienta de limpieza comercial porque gran parte de sus datos está en la misma forma básica. Así, la compañía puede fácilmente usar de nuevo las rutinas escritas.

La desventaja principal ha sido la cantidad de tiempo de desarrollo (alrededor de una semana) que se necesitó para crear las rutinas. Aunque tienen cierta dificultad de tiempo para mantenerse al día con la demanda y han buscado paquetes de software [comercial], no han encontrado aún, en el mercado, algo que se ajuste mejor a sus requerimientos.

- **Calidad de datos**

  La calidad de datos debe cumplir con los siguientes requisitos:

  - Exactitud: Los datos deben cumplir los requisitos de integridad, consistencia y densidad.
  - Integridad: Los datos deben cumplir los requisitos de Entereza y validez.
    - Entereza: Alcanzado por la corrección de datos que contienen anomalías.
    - Validez: Alcanzado por la cantidad de datos que satisfacen las restricciones de integridad.
  - Consistencia: Alcanzado por la corrección de contradicciones y anomalías sintácticas.
  - Uniformidad: Relacionado con irregularidades.
  - Densidad: Conocer el cociente de valores omitidos sobre el número de valores totales.
  - Unicidad: Relacionado con datos duplicados.

- **Procesos de limpieza de datos**
  - Auditoria de Datos: Los datos son revisados con el empleo de métodos estadísticos de descubrir anomalías y contradicciones. Esto tarde o temprano da una indicación de las características de las anomalías y sus posiciones.
  - Definición de Workflow (Flujo de Trabajo): La detección y el retiro de anomalías son realizados por una secuencia de operaciones sobre los datos sabidos como el workflow. Para alcanzar un workflow apropiado, se debe identificar las causas de las anomalías y errores. Si por ejemplo encontramos que una anomalía es un resultado de errores de máquina en etapas de entrada de datos, la disposición del teclado puede ayudar en la solución de posibles problemas.
  - Ejecución de Workflow: En esta etapa, el workflow es ejecutado después de que su especificación es completa y su corrección es verificada. The implementación del workflow debería ser eficiente aún sobre los juegos

grandes de los datos que inevitablemente plantean una compensación, porque la ejecución de la operación limpiadora puede ser cara.

- Post-Proceso y Control: Los datos que no podían ser corregidos durante la ejecución del workflow deberán ser corregidos manualmente, de ser posible. El resultado es un nuevo ciclo en el proceso de limpieza de datos donde los datos son revisados nuevamente para ajustarse a las especificaciones de un workflow adicional y realizar un tratamiento automático.

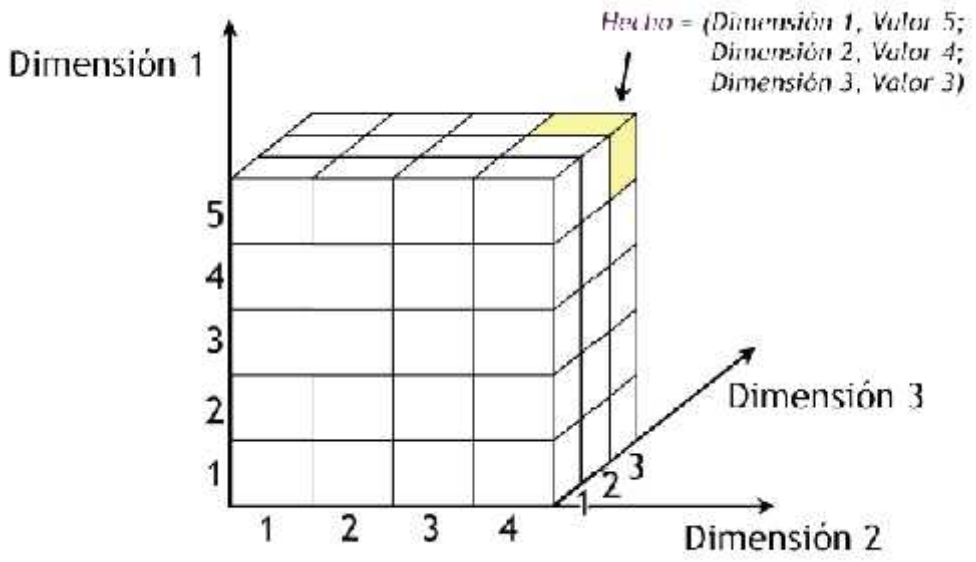
- **Métodos más usados**
  - Análisis: El análisis en la limpieza de datos, es realizado para la detección de errores de sintaxis. Un analizador gramatical decide si una cuerda de datos es aceptable dentro de la especificación de datos permitida. Esto es similar al modo que un analizador gramatical trabaja con gramáticas y lenguas.
  - Transformación de Datos: La Transformación de Datos permite al trazar un mapa de datos, en el formato esperado. Esto incluye conversiones de valor o funciones de traducción así como normalización de valores numéricos para conformarse a valores mínimos y máximos.

  - Eliminación de duplicados: La detección de duplicados requiere un algoritmo para determinar si los datos contienen representaciones dobles de la misma entidad. Por lo general, los datos son ordenados por un dato "llave" o "pivote" que permite la identificación más rápida.
  - Método Estadístico: Incluye analizar los datos usando promedios, desviación estándar, rangos, o algoritmos de cluster, este análisis se realiza por expertos que identificar errores. Aunque la corrección de datos sea difícil ya que no saben el valor verdadero, pueden ser resueltos poniendo los valores a un promedio u otro valor estadístico. Los métodos estadísticos también pueden ser usados para manejar los valores que fallan, que pueden ser substituidos por uno o varios valores posibles que por lo general son obtenidos por algoritmos de aumento de datos extensos...

4. Diseño Dimensional del Proceso Del Negocio

Las bases de datos multidimensionales, proveen una estructura que permite tener acceso flexible a los datos, para explorar y analizar sus relaciones, y resultados consiguientes. Estas se pueden visualizar como un cubo multidimensional, en donde las variables asociadas existen a lo largo de varios ejes o dimensiones, y la intersección de las mismas representa la medida, indicador o el hecho que se está evaluando.

En la siguiente representación matricial se puede ver más claramente lo que se acaba de decir



En el cubo de la figura anterior, existen tres dimensiones, “Dimensión 1”, “Dimensión 2” y “Dimensión 3”, cada una con sus respectivos valores asociados. También, se ha seleccionado un hecho al azar para demostrar su correspondencia con los valores de las dimensiones. En este caso, la medida a la que se hace referencia, representa el cruce del Valor “5” de “Dimensión 1”, con el Valor “4” de “Dimensión 2” y con el Valor “3” de “Dimensión 3”. Se puede observar que el resultado del análisis está dado por los cruces matriciales de acuerdo a los valores de las dimensiones seleccionadas.

Las bases de datos multidimensionales implican tres variantes posibles de modelamiento, que permiten realizar consultas de soporte de decisión:

Esquema en estrella5 (Star Scheme).
Esquema copo de nieve6 (Snowflake Scheme).
Esquema constelación7 o copo de estrellas (Starflake Scheme).

Los mencionados esquemas pueden ser implementados de diversas maneras, que, independientemente al tipo de arquitectura, requieren que toda la estructura de datos este desnormalizada o semi desnormalizada, para evitar desarrollar uniones (Join) complejas para acceder a la información, con el fin de agilizar la ejecución de consultas. Los diferentes tipos de implementación son los siguientes:

Relacional – ROLAP.
Multidimensional – MOLAP.
Híbrido – HOLAP.

4.1. Concepto Del Modelado Dimensional.

Es una técnica para diseñar el modelo lógico de la bodega e datos, que permite alto rendimiento en el momento de acceder a los datos (orientados a consultas).

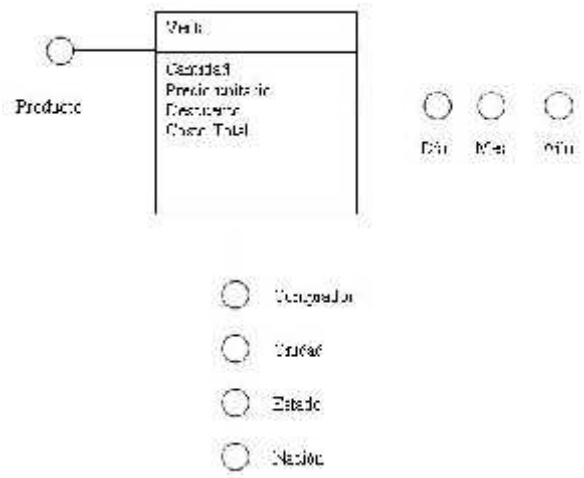
4.2. El Proceso Del Modelado Dimensional.

El análisis de los datos en un tiempo finito, ha traído consigo estudios sobre la mejor forma de almacenar y representar estos datos para que puedan ser consultados de una forma más rápida. El uso del Modelo Multidimensional es una de las aproximaciones más acertadas y seguidas por los especialistas en estos días. Este se basa en el estudio de los eventos del negocio analizados desde sus distintas dimensiones. Así:

Definición 1:

Llamamos evento o **Hecho** a una operación que se realiza en el negocio en un tiempo determinado. Son objeto de análisis para la toma de decisiones. Se Representan en una caja con su nombre y las medidas que lo caracterizan. (Robert Wrembel & Christian Concilia, 2007)

Ej: Figura 1: Representación gráfica de un Hecho y sus dimensiones



Los **Hechos** están estrechamente relacionados con el tiempo. Los eventos que son estáticos no tiene objetivo de análisis para este modelo, aunque son muy pocos los hechos que no ocurren con determinada periodicidad en un negocio. Los hechos están caracterizados por medidas numéricas como se muestra en el ejemplo de la figura 1: la cantidad, el precio unitario, el descuento, etc, son las medidas del Hecho (VENTA).

Nota: Fíjese que el producto que se vende, su costo y la fecha de la venta no son características de esta como lo podrían ser en cualquier diseño relacional. En este caso, esos serían dimensiones de ese Hecho, por las que, puede ser analizado más adelante.

Definición 2:

Una **Medida** es una propiedad de un Hecho (casi siempre numérica), que es usada para su análisis. (Robert Wrembel & Christian Concilia, 2007)

Nota: Un hecho puede no poseer ninguna medida. En ese caso se dice que el Hecho es vacío y solo se usa para contar la aparición de este en el tiempo.

Definición 3:

Una **Dimensión** es una característica de un hecho que permite su análisis posterior, en el proceso de toma de decisiones. (Robert Wrembel & Christian Concilia, 2007)

Nota: Un hecho debe estar relacionado al menos con una dimensión: "El tiempo".

Es un interés del negocio tomar decisiones sobre los hechos que ocurren en este, pero para esto se necesita su análisis. Por ejemplo Las ventas en la semana antes del 14 de Febrero, puede ser un objeto de análisis para un negocio comercial. Para esto se necesita tener el Hecho Ventas analizado en la dimensión Tiempo. En este caso en los Días:

7 <= d <= 14. Si se quisiera saber que productos fueron los más vendidos en esos días entonces tendríamos que adicionar una nueva dimensión de análisis, Producto. Así adicionando dimensiones a nuestro estudio se pudieran llegar a conclusiones sobre si el siguiente año en esa época debería comprarse más objetos de un producto o menos de otro. Elemento este muy importante para la futura estrategia de la empresa.

Definición 4:

En una empresa pueden existir varios hechos que sean analizados por dimensiones iguales. En este caso se les llama a estas dimensiones: Dimensiones Compartidas. (Robert Wrembel & Christian Concilia, 2007). Un ejemplo de esto es El Hecho Ventas puede ser analizado en las dimensiones Tiempo y Producto. Lo mismo ocurre con el Hecho Compras.

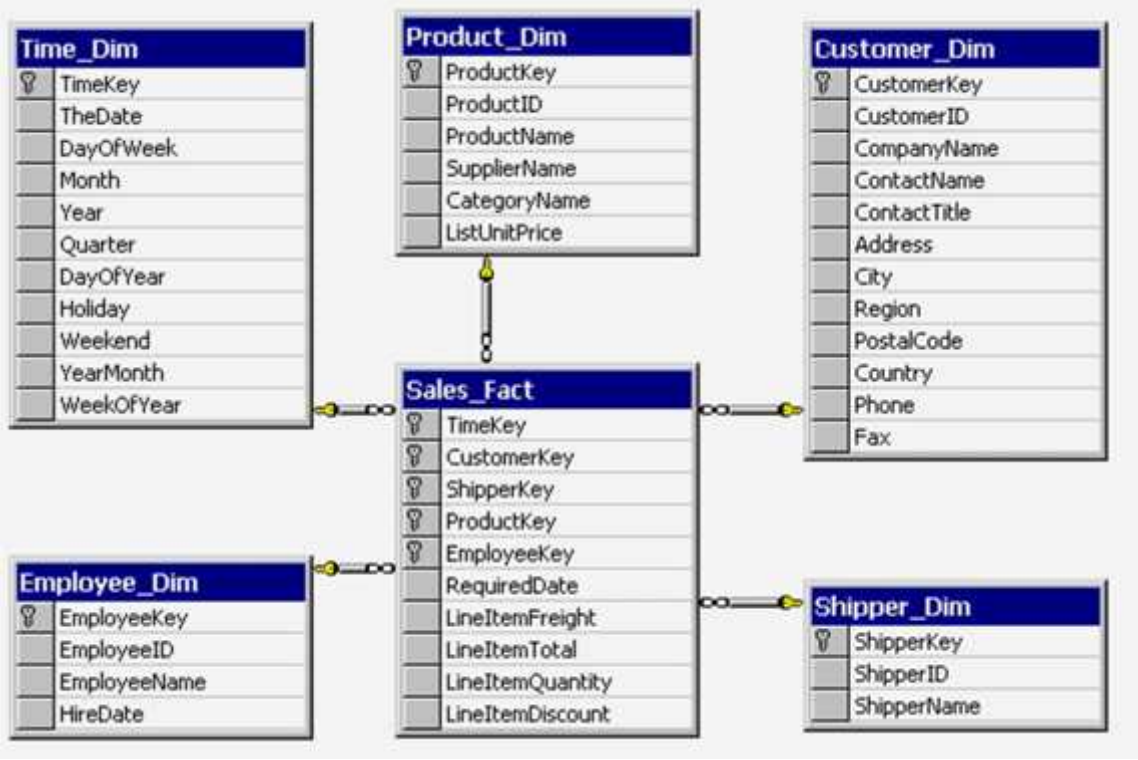
Las dimensiones deben ser atómicas y las relaciones entre estas crean jerarquías que permiten un análisis jerárquico de los datos. Un ejemplo de esto es el Tiempo, que es dividido en tres dimensiones. Día, Mes y Año. Cada uno es una dimensión distinta, pero relacionadas jerárquicamente en una relación de 1 a muchos, que permite el análisis del Hecho, por días, meses o años, o la combinación de ellos. Esto da al traste con las acostumbradas (OLTP) que manejan el Tiempo como una propiedad de una entidad, y lo tratan como un todo. Por lo que, como podemos inferir de aquí: en muchos casos hará falta convertir las bases de datos de estos sistemas a la nueva filosofía. SQL Server tiene facilidades para esto llamadas DTS (Data Transformation Services) que

permite leer datos desde cualquier SGBDR que posea un driver ODBC o implemente la nueva tecnología OLE DB de Microsoft.

Diagrama en Estrella

Uno de los tipos de consultas más usadas en las OLAP es la llamada Estrella. Su nombre lo adquiere debido a que su implementación en un ambiente relacional (MOLAP Multidimentional Online Analitical Processing) está dado por varias tablas que almacenan las jerarquías dimensionales y una tabla que contiene el hecho con una relación 1:m con estas tablas de dimensiones. Veamos un ejemplo gráfico:

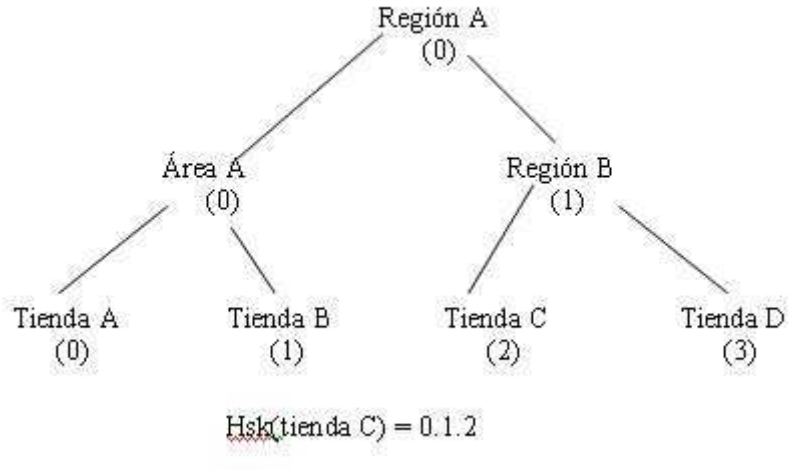
Figura 2: Diagrama en estrella del Hecho, Ventas. (Microsoft Data Warehouse Training Kit, 2000)



Como podemos ver en la figura las tablas de dimensiones están ligadas a la tabla Hecho, por relaciones. La integridad referencial es llevada a cabo por la creación de llaves foráneas en la tabla Hecho, que a su ves forman parte de la llave principal de la esta tabla. Es importante destacar que las jerarquías completas son guardadas en una sola tabla dimensión. Este es el formato no normalizado, existe

otro formato que intenta normalizar estas tablas dimensión. Ejemplo (Time_Dim). Cada tabla dimensión tiene su propia llave que es mantenida por el sistema Data Warehouse. A estas llaves se les llama "Surrogate Key". Las llaves Surrogate Jerárquicas, no son más que una codificación de cada elemento de la jerarquía almacenado en la tabla dimensión. Veamos la figura 3 de cómo se logran estas llaves.

Figura 3: Formación de una llave Surrogate Jerárquica (Robert Wrembel & Christian Concilia, 2007)



Vamos a ver ahora como sería una consulta sobre este tipo de diagrama en estrella:

Figura 3:

Plantilla de consulta para una consulta en estrella (ad hoc star query) (Robert Wrembel & Christian Concilia, 2007)

```
SELECT    <grouping attributes and/or aggregation functions>
FROM      <fact table>, D1, D2, ..., Dk
WHERE     <star join conditions: equalities on key-f.key> AND
          LP1 AND LP2 AND ... AND LPk AND
          <restrictions on attributes of the fact table>
GROUP BY  <grouping attributes>
HAVING    <group selection predicate>
ORDER BY  <sorting attributes>
```

Nota: En la figura las D1, D2, .. , Dk significan tablas de dimensión y los LP1, LP2, …,LPk son los predicados usados para simplificar la consulta. El ejemplo

siguiente muestra mejor como sería esta consulta: (Robert Wrembel & Christian Concilia, 2007)

```
SELECT L.area, D.month, SUM(F.sales)
FROM SALES_FACT F, LOCATION L, DATE D, PRODUCT P
WHERE F.day = D.day AND F.store_id = L.store_id AND
        F.product_id = P.item_id AND D.year = 1999 AND
L.population>1000000 AND P.category = "air condition"
GROUP BY L.area, D.month
```

En este tipo de procesamiento el mayor de los problemas es el super join que se crea al procesar las tablas de dimensiones con los datos de la tabla Hecho, para esto se han hecho varios estudios sobre la mejor forma de hacer este tipo de consultas de forma que sean lo más óptimas posibles, una de las técnicas mejores probadas es la de reescribir la consulta como lo muestra el siguiente ejemplo que mostramos:

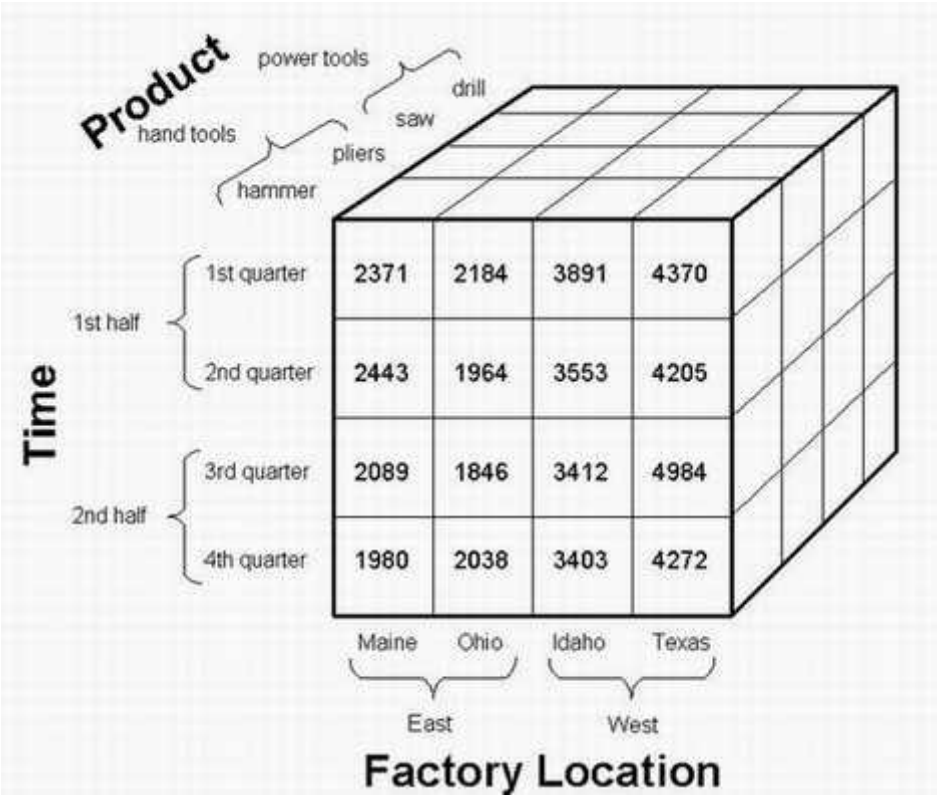
Ejemplo: Optimizar la consulta en el Data Warehouse (Robert Wrembel & Christian Concilia, 2007)

```
SELECT dim2.dim2_attr, dim3.dim3_attr, dim5.dim5_attr, fact.fact1
FROM fact, dim2, dim3, dim5
WHERE fact.dim2_key = dim2.dim2_key /* joins */
AND fact.dim3_key = dim3.dim3_key
AND fact.dim5_key = dim5.dim5_key
AND dim2.dim2_attr IN ('c','d') /* dimension restrictions */
AND dim3.dim3_attr IN ('e','f')
AND dim5.dim5_attr IN ('l','m')
is rewritten in the following form:
SELECT ... FROM fact
WHERE fact.dim2_key IN (SELECT dim2.dim2_key FROM dim2 WHERE dim2.dim2_attr IN
('c','d'))
AND fact.dim3_key IN (SELECT dim3.dim3_key FROM dim3 WHERE dim3.dim3_attr IN
('e','f'))
AND fact.dim5_key IN (SELECT dim5.dim5_key FROM dim5 WHERE AND dim5.dim5_attr
('l','m'))
```

Para cerrar con broche dorado este tema es necesario hacer alusión a los llamados Cubos de datos: Estos no son más que el conjunto formado por todas las tablas Dimensión y la tabla Hecho que al final dan una vista en forma de Cubo cuyas celdas están compuestas por las medidas de la tabla Hecho. Esta es la base de las aplicaciones OLAP. El cubo de datos es lo que hace que los reportes sean

obtenidos con un bajo tiempo de respuesta y que el análisis de los datos pueda ser tan diverso, pues cada cara del cubo se refiere a un análisis distinto de las medidas almacenadas. Veamos el ejemplo gráfico del cubo:

Figura 4: Cubo de datos (Microsoft Books Online, 2000)



Como podemos ver en el ejemplo la cantidad de producción puede ser analizada por producto, teniendo en cuenta la dimensión Producto, Por Tiempo, por Localización de las Industrias o en su conjunto por todas ellas a la vez o cualquier combinación de estas. Esto le da al analista o al sistema experto una amplia gama de posibilidades de las cuales puede tomar ventaja. En nuestro caso de estudio de las ventas. El cubo de datos formado por la Tabla Sales_Fac en conjunto con las restantes tablas de Dimensión nos permite analizar las ventas por Empleado, por Consumidor, por Tiempo, Etc.

5. Minería De Datos

) Historia de la minería a de datos

La minería de datos, entendida como la búsqueda de patrones dentro de grandes bases de datos utilizando para ello métodos estadísticos y de aprendizaje basado en telecomunicaciones, financiero y de autoservicio están en el proceso de adquirir alguna

solución tecnológica en este campo, por lo que surge una demanda por recursos humanos con conocimientos en minería de datos.

Además, al enfrentar un ambiente mas competitivo las empresas requieren de tecnologías que les permitan pronosticar, dentro de un marco probabilística, el comportamiento de sus clientes y prospectos a fin de desarrollar estrategias de atracción o retención.

La idea de data mining no es nueva. Ya desde los años sesenta (1960) los estadísticos manejaban el termino como data fishig, data mining o data archaeology con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido. A principios de los años ochenta (1980), Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky – Shapiro, entre otros, empezaron a consolidar los términos de data mining y KDD. A finales de los ochenta solo existía un par de empresas dedicadas a esta tecnología; en 2002 existen mas de 100 empresas en el mundo que ofrecen alrededor de 300 soluciones.

- **¿Por qué surge la Minería de Datos?**

El análisis e interpretación manual de los datos se torna impráctico (lento, caro y subjetivo) en la medida que los volúmenes de datos crecen exponencialmente.

Distintos factores influyen en la acumulación de datos:

- Dispositivos de almacenamiento más baratos.
- Transacciones comerciales son almacenadas mayoritariamente en formato electrónico.
- Captura automática de actividades realizadas en Internet.
- Desarrollo de algoritmos eficientes y robustos para el procesamiento de estos datos.
- Poder computacional más barato) métodos computacional/ intensivos para el análisis de datos.

Ventajas comerciales y científicas

- **¿Qué es la Minería de Datos (MD)?**

"... proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos" [Witten y Frank, 2000]
"... uso de datos históricos para descubrir regularidades generales y mejorar las decisiones futuras" [Mitchell, 1999]
"... proceso que tiene como objetivo convertir datos en conocimiento" [Hernández Orallo, 2004]

"... es un paso *particular en el proceso de KDD que consiste en la aplicación de algoritmos específicos para extraer patrones (o modelos) desde los datos"* [Fayyad, 1996]

- **Principales características y objetivos de la minería de datos**

  - Explorar los datos se encuentran en las profundidades de las bases de datos, como los almacenes de datos, que algunas veces contienen información almacenada durante varios años.
  - En algunos casos, los datos se consolidan en un almacén de datos y en mercados de datos; en otros, se mantienen en servidores de Internet e Intranet.
  - El entorno de la minería de datos suele tener una arquitectura cliente servidor.
  - Las herramientas de la minería de datos ayudan a extraer el mineral de la información enterrado en archivos corporativos o en registros públicos, archivados.
  - El minero es, muchas veces un usuario final con poca o ninguna habilidad de programación, facultado por barrenadoras de datos y otras poderosas herramientas indagatorias para efectuar preguntas adhoc y obtener rápidamente respuestas.
  - Hurgar y sacudir a menudo implica el descubrimiento de resultados valiosos e inesperados.
  - Las herramientas de la minería de datos se combinan fácilmente y pueden analizarse y procesarse rápidamente.
  - Debido a la gran cantidad de datos, algunas veces resulta necesario usar procesamiento en paralelo para la minería de datos.
  - La minería de datos produce cinco tipos de información:
    - Asociaciones.
    - Secuencias.
    - Clasificaciones.
    - Agrupamientos.
    - Pronósticos.

- **Los Principales Modelos De Análisis De Datos.**

  Podemos decir que "en Minería de Datos cada caso es un caso". Sin embargo, en términos generales, el proceso se compone de cuatro etapas principales:

- Determinación de los objetivos. Trata de la delimitación de los objetivos que el cliente desea.
- Preprocesamiento de los datos. Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Es la etapa que consume más de la mitad del tiempo del proyecto.

- Determinación del modelo. Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.
- Análisis de los resultados: Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

- **Tipología de Patrones de Minería de Datos.**

Existen diferentes Tipos de Conocimiento los cuales son:

- Asociaciones:

  Una asociación entre dos atributos ocurre cuando la frecuencia de que se den dos valores determinados de cada uno conjuntamente es relativamente alta.

Ejemplo:

En un supermercado se analiza si los pañales y los potitos de bebé se compran conjuntamente.

- Dependencias:

  Una dependencia funcional (aproximada o absoluta) es un patrón en el que se establece que uno o más atributos determinan el valor de otro.
  Existen muchas dependencias nada interesantes (causalidades inversas).

Ejemplo:

Que un paciente haya sido ingresado en maternidad determina su sexo.

- Clasificación:

  Una clasificación se puede ver como el esclarecimiento de una dependencia, en la que el atributo dependiente puede tomar un valor entre varias clases, ya conocidas.

Ejemplo:

Se sabe (por un estudio de dependencias) que los atributos edad, grado de miopías y astigmatismo han determinado los pacientes para los que su operación

de cirugía ocular ha sido satisfactoria. Podemos intentar determinar las reglas exactas que clasifican un caso como positivo o negativo a partir de esos atributos.

- Agrupamiento / Segmentación:

  El agrupamiento (o clustering) es la detección de grupos de individuos. Se diferencia de la clasificación en el que no se conocen ni las clases ni su número (aprendizaje no supervisado), con lo que el objetivo es determinar grupos o racimos (clusters) diferenciados del resto.

- Tendencias/Regresión:

  El objetivo es predecir los valores de una variable continua a partir de la evolución sobre otra variable continua, generalmente el tiempo, o sobre un conjunto de variables.

Ejemplo.

Se intenta predecir el número de clientes o pacientes, los ingresos, llamadas, ganancias, costes, etc. a partir de los resultados de semanas, meses o años anteriores.

- Información del Esquema: (descubrir claves primarias alternativas, R.I.).
- Reglas Generales: patrones no se ajustan a los tipos anteriores. Recientemente los sistemas incorporan capacidad para establecer otros patrones más generales.

6. Caso Practico Editorial Perú S.A.

Trabajos Previos

Business Intelligence (Inteligencia de negocios)

Según [Microsoft, 2004] vivimos en una época en que la información es la clave para obtener una ventaja competitiva en el mundo de los negocios. Para mantenerse competitiva una empresa, los gerentes y tomadores de decisiones requieren de un acceso rápido y fácil a información útil y valiosa para la empresa. Una forma de solucionar este problema es por medio del uso de Business Intelligence o Inteligencia de Negocios. La Inteligencia de Negocios o Business Intelligence (BI) se puede definir como el proceso de analizar los bienes o datos acumulados en la empresa y extraer una cierta inteligencia o conocimiento de ellos. Dentro de la categoría de bienes se incluyen las bases de datos de clientes, información de la cadena de suministro, ventas personales y cualquier actividad de marketing o fuente de información relevante para la empresa. La clave para BI es la información y uno de sus mayores beneficios es la posibilidad de utilizarla en la toma de decisiones. Tal vez le ayude a comprender mejor el concepto por medio de un ejemplo. Una franquicia de hoteles a nivel nacional que utiliza aplicaciones de BI para llevar un registro estadístico del porcentaje promedio de ocupación del hotel, así como los días promedio de estancia de cada huésped, considerando las diferencias entre temporadas.
Con esta información se puede:

- Calcular la rentabilidad de cada hotel en cada temporada del año
- Determinar quién es su segmento de mercado
- Calcular la participación de mercado de la franquicia y de cada hotel
- Identificar oportunidades y amenazas

Estas son sólo algunas de las formas en que una empresa u organización se puede beneficiar por la implementación de software de BI, hay una gran variedad de aplicaciones o software que brindan a la empresa la habilidad de analizar de una forma rápida por qué pasan las cosas y enfocarse a patrones y amenazas.
¿Qué se puede hacer con Business Intelligence (BI)? Con Business Intelligence (BI) se puede:

- Generar reportes globales o por secciones
- Crear una base de datos de clientes

- Crear escenarios con respecto a una decisión
- Hacer pronósticos de ventas y devoluciones
- Compartir información entre departamentos
- Análisis multidimensionales
- Generar y procesar datos

- Cambiar la estructura de toma de decisiones
- Mejorar el servicio al cliente

La siguiente es una lista de las áreas más comunes en las que las soluciones de inteligencia de negocios son utilizadas:

- Ventas: Análisis de ventas; Detección de clientes importantes; Análisis de productos, líneas, mercados; Pronósticos y proyecciones.
- Marketing: Segmentación y análisis de clientes; Seguimiento a nuevos productos.
- Finanzas: Análisis de gastos; Rotación de cartera; Razones financieras.
- Manufactura: Productividad en líneas; Análisis de desperdicios; Análisis de calidad; Rotación de inventarios y partes críticas.
- Embarques: Seguimiento de embarques; Motivos por los cuales se pierden pedidos.

Componentes de Business Intelligence Multidimensionalidad:

La información multidimensional se puede encontrar en hojas de cálculo, bases de datos, etc. Una herramienta de BI debe ser capaz de reunir información dispersa en toda la empresa e incluso en diferentes fuentes para así proporcionar a los departamentos la accesibilidad, poder y flexibilidad que necesitan para analizar la información. Por ejemplo, un pronóstico de ventas de un nuevo producto en varias regiones no está completo si no se toma en cuenta también el comportamiento histórico de las ventas de cada región y la forma en que la introducción de nuevos productos se ha desarrollado en cada región en cuestión.

Data Mining (Minería de Datos):

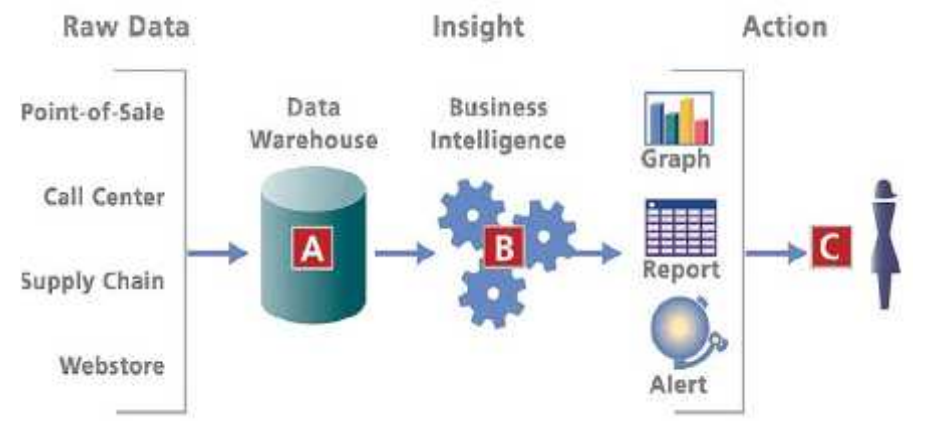
Las empresas suelen generar grandes cantidades de información sobre sus procesos productivos, desempeño operacional, mercados y clientes. Pero el éxito de los negocios depende por lo general de la habilidad para ver nuevas tendencias o cambios en las tendencias. Las aplicaciones de data mining pueden identificar tendencias y comportamientos, no sólo para extraer información, sino también para descubrir las relaciones en bases de datos que pueden identificar comportamientos que no son muy evidentes.

Agentes:

Los agentes son programas que "piensan". Ellos pueden realizar tareas sin necesidad de intervención humana. Por ejemplo, un agente pueden realizar tareas complejas, como elaborar documentos, establecer diagramas de flujo, etc.

Data Warehouse (Almacén de Datos):

Es la respuesta de la tecnología de información a la descentralización en la toma de decisiones. Coloca información de todas las áreas funcionales de la organización en manos de quien toma las decisiones. También proporciona herramientas para búsqueda y análisis.



A: Data centralizada desde múltiples fuentes dentro de un datawarehouse
B: Herramientas BI que analizan la data para entender mejor el negocio
C: Reportes inteligentes para la toma de decisiones
Figura 2 : Business intelligence

El Futuro:

La Figura 2, describe la Inteligencia de Negocios (BI) que ya no puede ser ignorada por ninguna organización que reconoce que estamos en la era de la información. En el futuro cercano debemos esperar lo siguiente:

- ✓ Proyectos más frecuentes y más largos
- ✓ Barreras de entrada para

los primeros
- ✓ La infraestructura será estándar
- ✓ Se establecerán centros de información
- ✓ Convergencia de tecnologías (acceso por internet)
- ✓ Cambiar actividades consideradas periféricas en Core Business

Soluciones Datawarehouse

Según [Cognos, 2002] en AFP Nueva Vida las soluciones cognos cambiaron la manera como se hacían los negocios. se permitió efectuar seguimiento más a detalle del comportamiento de los afiliados y las empresas aportadoras. asimismo, se identificaron segmentos importantes de clientes y empresas.
Según [Nakasone, 2004] las empresas han optado por utilizar la inteligencia de negocios y el primer escalón es construyendo un almacén de datos (datawarehouse), para luego avanzar en la minería de datos (datamining).

Modelo Olap para una empresa pública

A fin de alcanzar los objetivos planteados en el presente trabajo se ha planteado el modelamiento de una solución general para toda empresa del estado, tomamos como base la empresa pública Editora Perú, pero puede extenderse la solución a cualquier otra empresa tal como Sedapal, Essalud, Ministerios, etc.

Figura 3: Modelo General

En la Figura 3 se muestra el Modelo General en este caso mostrando a Editora Perú, pero se puede generalizar para toda empresa del estado, todas las empresas tienen la obligación legal de reportar a las siguientes entidades:

- Sunat (Superintendencia Nacional de Administración Tributaria)
- Fonafe (Fondo Nacional de Financiamiento de la Actividad Empresarial del Estado)
- Produce (Ministerio de la Producción)
- INEI (Instituto Nacional de Estadística e Informática)
- Contaduría General de la República
- Contraloría General de la República
- Consucode (Consejo Superior de Contrataciones y Adquisiciones del Estado)

Además internamente Editora Perú debe generar información interna para sus órganos de supervisión y control: Recursos Humanos, Costos, Presupuesto, Indicadores, Gestión Contable, Clientes, Proveedores, Tesorería, Activo Fijo, Manufactura, Distribución.

En la Figura 4 se muestra el Modelo General Detalle, donde se puede observar que cada entidad exige información detallada para fines de control, las que se deben entregar en tiempo y plazos establecidos legalmente.

En la Figura 4 en el extremo superior derecho se muestra el Modelo OLAP Sunat, se resalta el Registro de Ventas que será lo que se desarrollará como parte de este trabajo, en este caso se tiene la base legal, los formularios, formatos de archivo a informar, archivo de transferencia, luego a través del Sistema PDT la información será enviada a la Sunat.
El presente trabajo permite realizar un "reciclaje" o "reuso" de ésta información para alimentar los datamart y conformar el datawarehouse de una empresa pública.

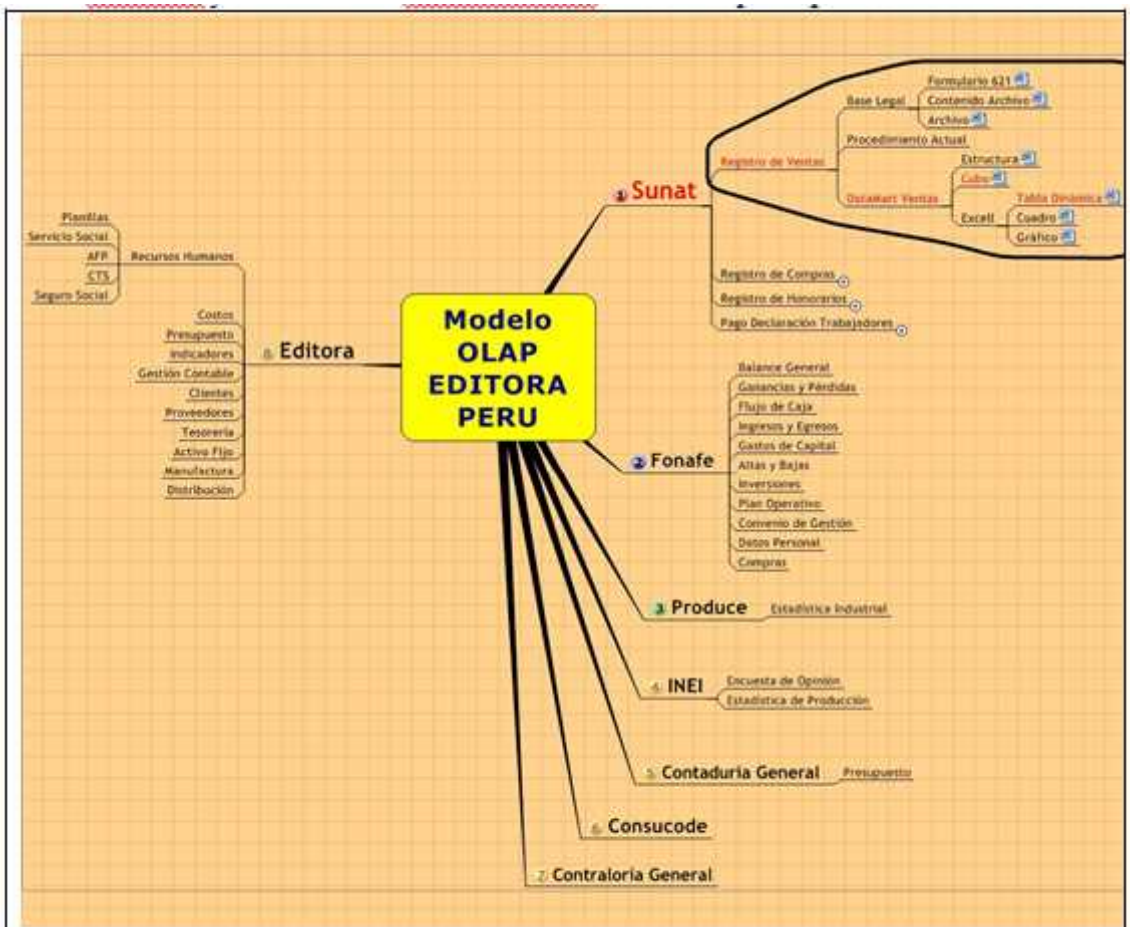


Figura 4: Modelo General Detalle

Experimentos y Resultados

Organización e instancias de prueba

Se ha considerado los años del 2002 al 2005. Los componentes de hardware con los que trabajará el sistema son: Servidor compatible, Servidor IBM X Series Modelo H70, SAN Storage HP.Mientras que los componentes de software con los que trabajará el sistema son: Sistema Operativo: Windows 2000 para el servidor, Explorador de Internet (Internet Explorer), Manejador de base de datos MSSQL 2000, Sistema ERP, Sistema Operativo AIX v 4.3, Manejador de Base de Datos Oracle v.9, Sistema ERP BAAN IV C2.

Procesa
miento

Se ha desarrollado procedimientos almacenados que permiten transferir los datos del datamart de ventas de la base de datos Oracle v9.0 a Ms Sql 2000, luego otro procedimiento almacenado genera el cubo o datamart de ventas dentro del Analysis Server, en ese momento el usuario tiene disponible la información que se muestra.

Resul
tados

El usuario debe acceder a la hoja de cálculo y mediante un procedimiento seleccionar el cubo de ventas.

La Figura 5 nos muestra las ventas totales de los años 2002 al 2005 donde podemos visualizar
cada mes como se comportan las ventas en Editora Perú.

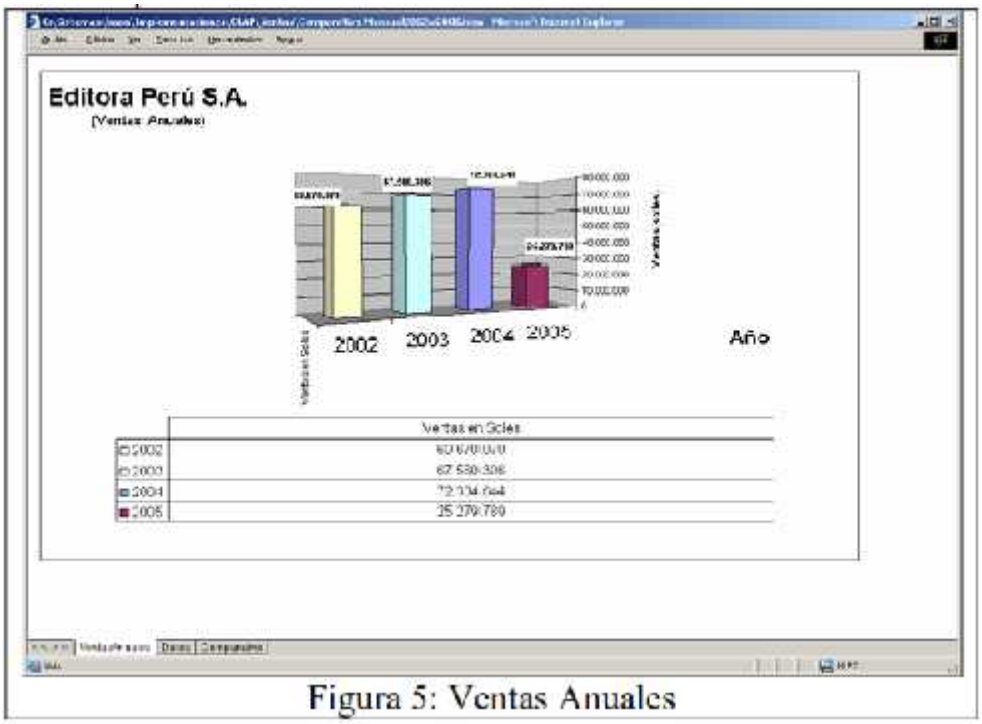


Figura 5: Ventas Anuales

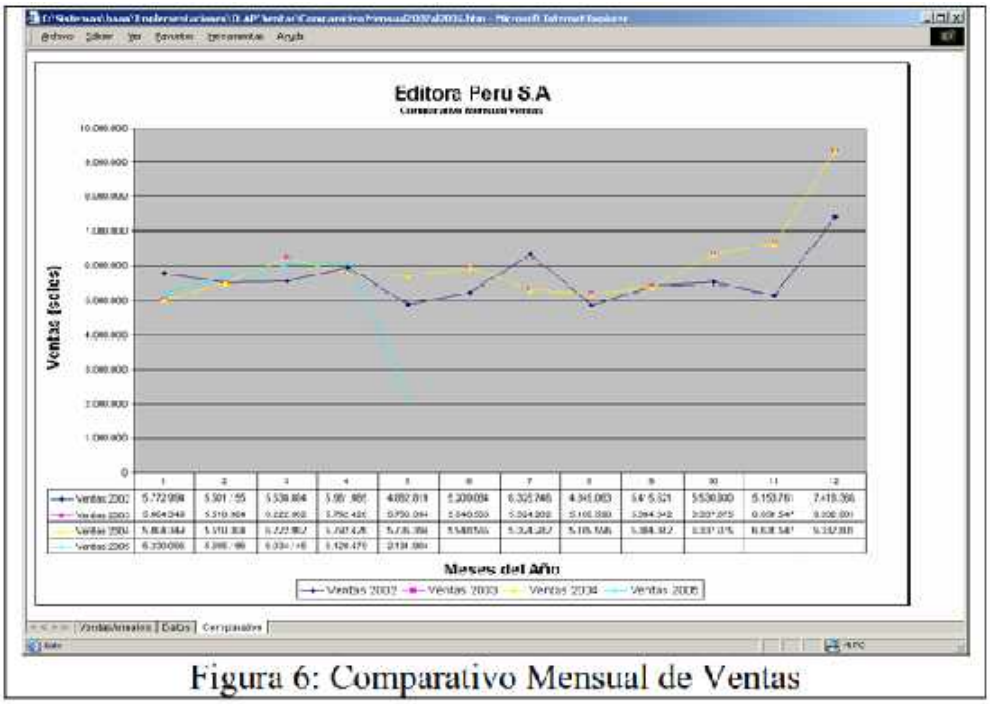


Figura 6: Comparativo Mensual de Ventas

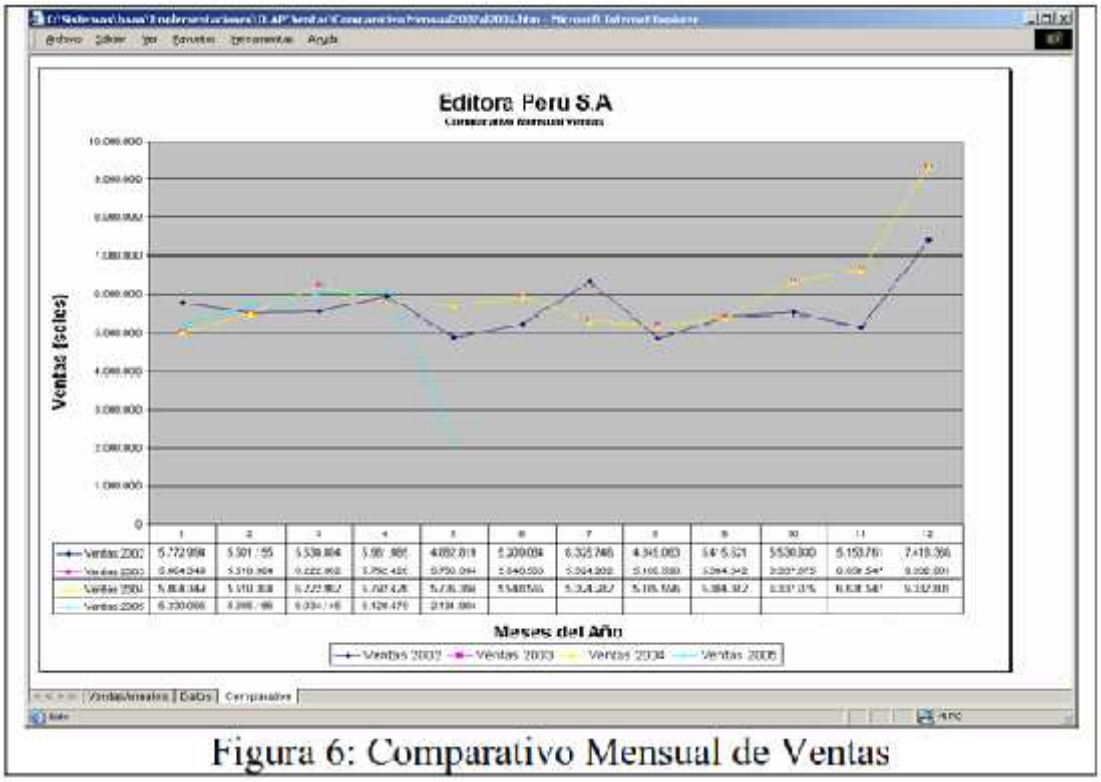


Figura 6: Comparativo Mensual de Ventas

La Figura 6 nos muestra el cuadro comparativo mensual de ventas, en él se puede comparar la venta mensual como anual, de modo que se muestren las tendencias y las expectativas de venta de acuerdo al pronóstico realizado al inicio del año. Esta información está disponible para toda la empresa.

Discusión de los Experimentos

En la Figura 7 se muestra gráficamente la solución al problema, así como las ventajas que se obtienen con la implementación de esta solución:

Ahora se cuenta con un servidor OLAP diferente al servidor OLTP por lo que ya no se compite con los recursos diarios, el acceso es más rápido y directamente de la hoja de cálculo Microsoft Office.

El procesamiento es totalmente automatizado, se procesa el periodo deseado, en este caso se ha considerado hacerlo quincenal y mensualmente.

Al existir un solo repositorio OLAP ahora todos los usuarios de la empresa pueden acceder simultáneamente a la información sin problemas de lentitud manteniéndose la integridad.

Se ha estandarizado el uso de plantillas para la presentación de la información tanto de los cuadros como de los gráficos.

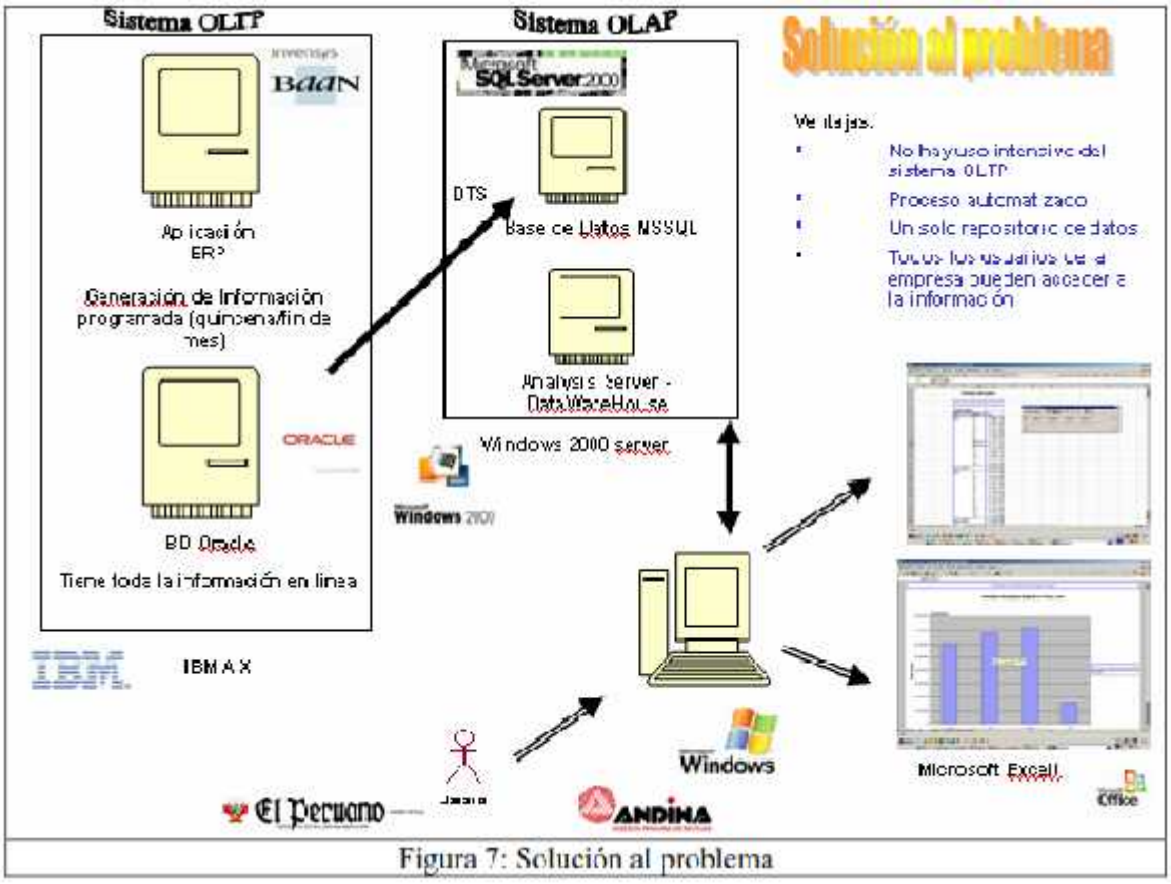


Figura 7: Solución al problema

7. Conclusiones.

La implementación del datamart de ventas permitirá que el usuario pueda contar con una herramienta en línea totalmente automatizada de fácil uso, que le permita disminuir el tiempo de procesamiento y dedique mayor tiempo a la etapa de análisis de la información

- ✓ Se eliminan los errores o diferencias por migración de datos y formateo.
- ✓ Se disminuye el tiempo de procesamiento por procedimiento manuales
- ✓ Por decisión de usuario el procesamiento se realiza quincenal y mensual, quedando abierta la posibilidad de efecuralo diariamente.
- ✓ El acceso a la información residente ahora en el servidor OLAP es mas rápido en comparación con el acceso a la base de datos del servidor OLTP utilizando en el proceso manual.

Según [Microsoft, 2004], los usuarios pueden acceder a un solo repositorio de datos (servidor OLAP) directamente desde la hoja de cálculo Microsoft Office, utilizando plantillas estándar. El único costo para la implementación de este sistema consiste en la capacitación de la herramienta Analysis Server de Microsoft producto que viene como parte de la Base de Datos MS SQL .

Recomendaciones o trabajos futuros

Como se muestra en la Figura 7 existen las siguientes tareas a futuro:
Generar información de los datamarts para informar en el Portal de Editora Perú (www.editoraperu.com.pe). Implementar todos los cubos faltantes que se muestran en la Figura 4.

Modelo General OLAP Detallado. Que todos los usuarios de la empresa tengan acceso a la información de la empresa vía Intranet

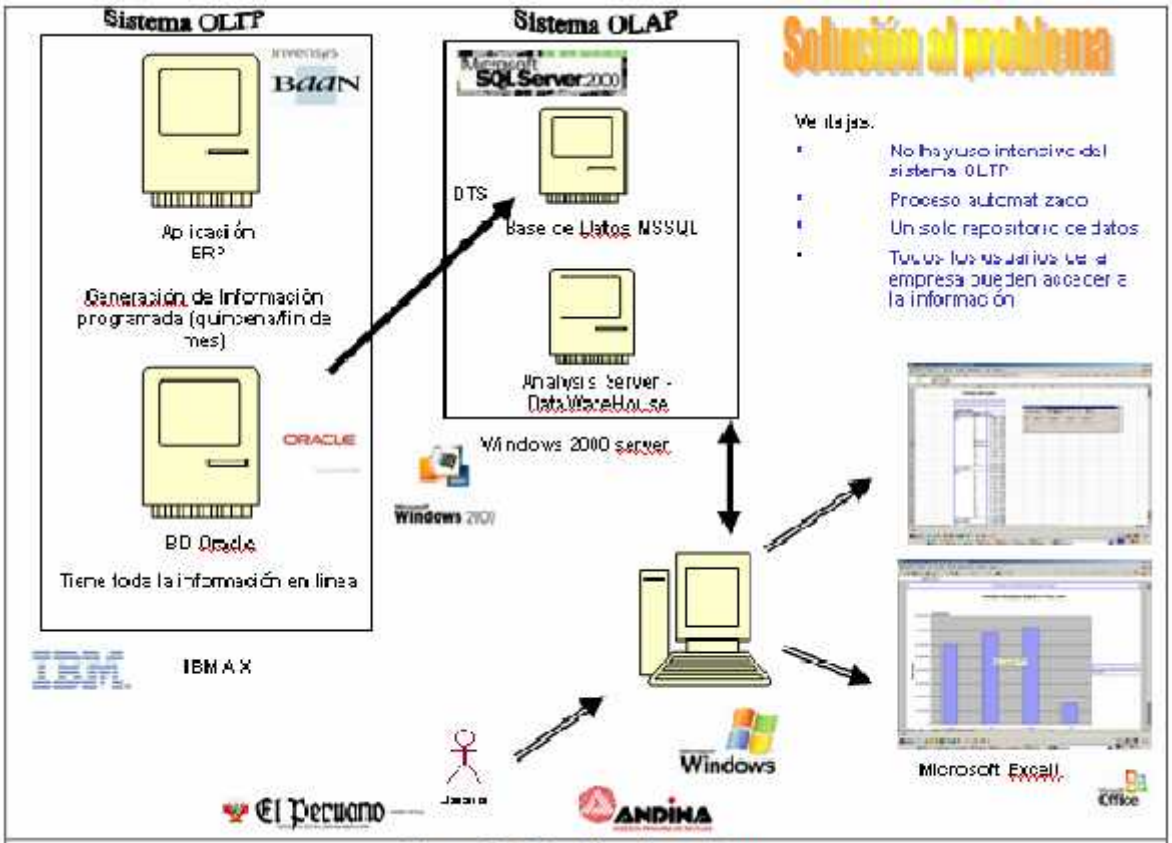


Figura 7: Solución al problema

8. Referencia Bibliográfica.

- Han, J., Kamber, M. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001. ISBN 1-55860-489-8.
- Kimball, R., Caserta, J. The Data Warehouse ETL Toolkit, Wiley and Sons, 2004. ISBN 0-7645-6757-8.
- Muller H., Freytag J., Problems, Methods, and Challenges in Comprehensive Data Cleansing, Humboldt-Universitat zu Berlin, Germany.
- Rahm, E., Hong, H. Data Cleaning: Problems and Current Approaches, University of Leipzig, Germany.
- [Microsoft, 2004] Microsoft (2004). Guía de Estrategia de Business Intelligence. 24 pags, © 2004
- Microsoft Corporation.
- [Kwon 01] Kwon, O. B. and Leeb, J. J. A multi-agent intelligent system for efficient ERP maintenance, Instituto Handong de Informacion Tecnologica, Expert System and Application, Volume 21, Issue 4, November 2001, Pages 191-202.
- [Cognos, 2002] Cognos, Casos de éxito; AFP Unión Vida - Perú, 2003
- [Nakasone, 2004] Nakasone Nicolas; OLAP Essential - Línea de Talleres Inteligencia de
- Negocios, MugPerú, Lima 2004
- [Nakasone, 2004] Nakasone Nicolas; OLAP Advanced - Línea de Talleres Inteligencia de
- Negocios, MugPerú, Lima 2004
- [Microsoft, 2004] Microsoft; Designing and Implementing OLAP Solutions with Microsoft SQL Server; 2004, Workbook, Microsoft Training and Certification Microsoft Official
- Microsoft 2004 Crear inteligencia empresarial con Analysis Services de Office XP y SQL 2000

- [Diane Larsen, 2003] Servicios de transformación de datos (DTS) en Microsoft SQL Server 2000,
- Microsoft Corporation Site : http://www.microsoft.com/spanish/msdn/articulos/archivo/010201/voices/dts_overview.asp, accesado en Enero 2007.
- [Cognos, 2005] Delivering Warehouse Return On Investment with Business Intelligence From
- Cognos®, http://www.dmreview.com/whitepaper/dwi.pdf, Datawarehouse Review
- Whitepaper from Sybase, Inc
- [Cognos, 2005] Data Warehousing for Healthcare: The Greatest Weapon in Your Competitive
- Arsenal, http://www.dmreview.com/whitepaper/WID242.pdf , Datawarehouse Review
- OLAP Train and Reed Jacobson

Web

http://businessintelligence.lifetips.com/es/tip/134745/performance-management-reporting/performance-management-reporting/glosario-de-inteligencia-de-negocios.html

http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MonoAdsDiseno.pdf