UNIVERSIDAD NACIONAL DE LA AMAZONIA PERUANA

FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA

“DATA WAREHOUSING”

INFORME DE TRABAJO PRACTICO DE SUFICIENCIA

PARA OBTAR EL TITULO PROFESIONAL DE:

INGENIERO DE SISTEMAS E INFORMATICA

PRESENTADO POR EL BACHILLER:

JESUS MARTIN RAMIREZ PANDURO

ASESOR:

DR. LUIS BENJAMÍN IRIGOIN SÁNCHEZ

IQUITOS – PERU

2014

Informe Técnico de examen de suficiencia previa actualización académica, aprobado en sustentación pública, por el jurado examinador designado por el coordinador de la Facultad de Ingeniería de Sistemas e Informática, de la Universidad Nacional de la Amazonía Peruana.

Lic. Ángel Enrique López Rojas
PRESIDENTE

Ing. José Edgar García Díaz
PRIMER MIEMBRO

Lic. Richard Alex López Albiño
SEGUNDO MIEMBRO

Dr. Luis Benjamín Irigoin Sánchez
ASESOR

DEDICADO al mis padres Don Jesús Ramírez Enrique y Doña Leovina Panduro Cardenas, por su apoyo incondicional para formarme como profesional, y a mi nueva familia, Cheryll Marcelina Armas Viteris y Maria Jesús Ramírez Armas mi motor y motivo para seguir creciendo.

Agradezco a Dios por iluminarme en mi camino cada día; por haberme dado una familia unida.

DATA WAREHOUSING

PRESENTACIÓN

Desde que se inició la era de la computadora, las organizaciones han usado los datos desde sus sistemas operacionales para atender sus necesidades de información. Algunas proporcionan acceso directo a la información contenida dentro de las aplicaciones operacionales. Otras, han extraído los datos desde sus bases de datos operacionales para combinarlos de varias formas no estructuradas, en su intento por atender a los usuarios en sus necesidades de información.

Ambos métodos han evolucionado a través del tiempo y ahora las organizaciones manejan una data no limpia e inconsistente, sobre las cuales, en la mayoría de las veces, se toman decisiones importantes.

La gestión administrativa reconoce que una manera de elevar su eficiencia está en hacer el mejor uso de los recursos de información que ya existen dentro de la organización. Sin embargo, a pesar de que esto se viene intentando desde hace muchos años, no se tiene todavía un uso efectivo de los mismos.

La razón principal es la manera en que han evolucionado las computadoras, basadas en las tecnologías de información y sistemas. La mayoría de las organizaciones hacen lo posible por conseguir buena información, pero el logro de ese objetivo depende fundamentalmente de su arquitectura actual, tanto de hardware como de software

El data warehouse, es actualmente, el centro de atención de las grandes instituciones, porque provee un ambiente para que las organizaciones hagan un mejor uso de la información que está siendo administrada por diversas aplicaciones operacionales.

Un data warehouse es una colección de datos en la cual se encuentra integrada la información de la Institución y que se usa como soporte para el proceso de toma de decisiones gerenciales. Aunque diversas organizaciones y personas individuales logran comprender el enfoque de un Warehouse, la experiencia ha demostrado que existen muchas dificultades potenciales.

Reunir los elementos de datos apropiados desde diversas fuentes de aplicación en un ambiente integral centralizado, simplifica el problema de acceso a la información y en consecuencia, acelera el proceso de análisis, consultas y el menor tiempo de uso de la información.

La innovación de la Tecnología de Información dentro de un ambiente data warehousing, puede permitir a cualquier organización hacer un uso más óptimo de los datos, como un ingrediente clave para un proceso de toma de decisiones más efectivo. Las organizaciones tienen que aprovechar sus recursos de información para crear la información de la operación del negocio, pero deben considerarse las estrategias tecnológicas necesarias para la implementación de una arquitectura completa de data warehouse.

RESUMEN

En el presente trabajo, se sistematizarán todos los conceptos inherentes al Data Warehousing, haciendo referencia a cada uno de ellos en forma ordenada, en un marco conceptual claro, en el que se desplegarán sus características y cualidades, y teniendo siempre en cuenta su relación o interrelación con los demás componentes del ambiente.

Inicialmente, se definirá el concepto de Business Intelligence y sus respectivas características. Seguidamente, se introducirá al Data Warehousing y se expondrán sus aspectos más relevantes y significativos. Luego, se precisarán y detallarán todos los componentes que intervienen en su arquitectura, de manera organizada e intuitiva, atendiendo su interrelación. Finalmente, se describirán algunos conceptos complementarios que deben tenerse en cuenta.

El principal objetivo de esta investigación, es ayudar a comprender el complejo ambiente del Data Warehousing, sus respectivos componentes y la interrelación entre los mismos, así como también cuáles son sus ventajas, desventajas y características propias. Es por ello, que se hará énfasis en la sistematización de todos los conceptos de la estructura del Data Warehousing, debido a que la documentación existente se enfoca en tratar temas independientes sin tener en cuenta su vinculación y referencias a otros componentes del mismo.

ÍNDICE

Tabla de contenido

JUSTIFICACIÓN

El presente trabajo adquiere una notoria importancia pues en nuestro país y de modo muy particular en la región Loreto; y más aún en nuestra ciudad no existe organización alguna sea del ámbito gubernamental o privado de origen local; que tenga implementada una solución tecnología de Inteligencia de Negocio.

Las que las coloca en una clara desventaja frente a empresas que si las tienen y que hacen un uso intensivo de la información propia de ellos, que se encuentran debidamente estructuradas (base de datos); además de información no estructurada y que se encuentra en diferentes medios magnéticos como hojas de cálculos, tablas aisladas (dbf´s y demas), o la Web.

La información que aquí se resumen y sistematiza pretende ser una guía para aquellas organizaciones que tomen la decisión de implementar una Arquitectura TIC, sin importar si la solución sea para una gran empresa o pequeña o si la solución será comercial o de tendencia libre, esto por el lado de la empresa. Pero además pretende brindar a los primerizos una puerta de inicio a los temas técnicos que los profesionales en TIC’s tendrán que asumir.

OBJETIVOS

GENERAL:

Aplicar los beneficios de DATA WAREHOSING en los proceso de toma de decisiones organizacionales.

ESPECÍFICO:

a) Reducir el tiempo mínimo que se requiere para recoger toda la información relevante de un tema en particular.
b) Proporcionar herramientas de análisis para establecer comparaciones y tomar decisiones.

Desarrollo del Tema

1. INTRODUCCIÓN AL DATA WAREHOUSING: DEFINICIONES Y CONCEPTOS

Los activos de información son inmensamente valiosos para cualquier empresa, y debido a esto, dichos activos deberán ser almacenados adecuadamente y de fácil acceso cuando se necesitan. Sin embargo, la disponibilidad de demasiados datos hace que la extracción de la información más importante sea difícil, si no imposible. Ver resultados de cualquier búsqueda de Google, y verá que la ecuación de datos = información no siempre es correcta, es decir, demasiados datos es simplemente demasiado. El almacenamiento de datos es un fenómeno que surgió de la gran cantidad de datos electrónicos almacenados en los últimos años y de la urgente necesidad de utilizar esos datos para lograr objetivos que van más allá de las tareas de rutina vinculados al procesamiento diario. En un escenario típico, una gran corporación tiene muchas sucursales, y los altos directivos tienen que cuantificar y evaluar cómo cada sucursal contribuye al rendimiento empresarial global. Las bases de datos corporativas detallan datos sobre las tareas realizadas por las sucursales. Para satisfacer las necesidades de los administradores, las consultas hechas a la medida pueden ser emitidas para recuperar los datos requeridos. Para que este proceso funcione, los administradores de bases de datos primero deben formular la consulta deseada (normalmente una consulta SQL agregada) después de estudiar de cerca los catálogos de base de datos. Entonces la pregunta es procesada. Esto puede tomar un par de horas debido a la enorme cantidad de datos, la complejidad de la consulta, y los efectos concurrentes de otras consultas regulares sobre la carga de trabajo de datos. Por último, se genera un informe y se pasa a los altos directivos en forma de una hoja de cálculo. Hace muchos años, los diseñadores de bases de datos se dieron cuenta de que este enfoque es poco factible, ya que es muy exigente en términos de tiempo y recursos, y no siempre logra los resultados deseados. Por otra parte, una mezcla de consultas analíticas con las consultas rutinarias de transacciones disminuye inevitablemente el sistema, y esto no se ajusta a las necesidades de los usuarios de cualquier tipo de consulta. Los Data Warehouse de hoy realizan procesamiento analítico en línea separada (OLAP) de procesamiento transaccional en línea (OLTP) mediante la creación de un nuevo repositorio de información que integra los datos básicos a partir de diversas fuentes, organiza correctamente los formatos de datos, y luego hace que los datos estén disponibles para el análisis y evaluación orientados a la planificación y los procesos de toma de decisiones.

Repasemos algunos campos de aplicación para los que se utilizan con éxito las tecnologías de Data Warehousing:

1. **Ventas y reclamos**; para análisis comerciales, envío y control de inventario, atención al cliente y relaciones públicas.
2. **Control de costes de producción** de bienes, suministrando y ayuda de la orden.
3. **Servicios financieros** de análisis de riesgos y tarjetas de crédito, detección de fraudes.
4. **Industria del transporte** para la gestión de vehículos.

5. **Los servicios de telecomunicaciones** de llamadas análisis de flujo y análisis de perfil de cliente.
6. **Servicio de atención médica** de admisión del paciente.

El campo de aplicación de los sistemas de Data Warehousing no sólo se limita a las empresas, sino que también va de la epidemiología a la demografía, de la ciencia natural y a la educación. Una propiedad que es común a todos los campos es la necesidad de herramientas de almacenamiento y consulta para recuperar resúmenes de información con facilidad y rapidez de la enorme cantidad de datos almacenados en bases de datos o puestos a disposición por la Internet. Este tipo de información nos permite estudiar fenómenos de negocios, aprendemos acerca de las correlaciones significativas, y adquirir conocimiento útil para apoyar los procesos de toma de decisiones.

Sistema de apoyo a las decisiones; Un sistema de soporte de decisiones (DSS) es un conjunto de técnicas informáticas interactivas extensibles y herramientas diseñadas para tratar y analizar datos y para apoyar a los gerentes en la toma de decisiones.

Data warehousing; es una colección de métodos, técnicas y herramientas que se utilizan para apoyar a gestores del conocimiento, trabajadores de alto nivel, directores, gerentes y analistas para llevar a cabo los análisis de datos que ayuda con la realización de los procesos de toma de decisiones y la mejora de los recursos de información.

Data Warehouse; es una colección de datos que soporta los procesos de toma de decisiones. Ofrece las siguientes características (Inmon, 2005):
Es orientado a un tema.
Está integrada y coherente.
Se muestra su evolución en el tiempo y no es volátil.

Data Marts; es un subconjunto o una agregación de los datos almacenados a un almacén de datos principal. Incluye un conjunto de piezas de información pertinentes a un área específica del negocio, dirección corporativa, o categoría de usuarios.

2. MODELOS DE DATA WAREHOUSING Y OPERACIONES EN CUBOS OLAP

2.1. BASE DE DATOS MULTIDIMENSIONALES

Una base de datos multidimensional es una base de datos en donde su información se almacena en forma multidimensional, es decir, a través de tablas de **hechos** y tablas de **dimensiones**.

Proveen una estructura que permite, a través de la creación y consulta a una estructura de datos determinada (cubo multidimensional, Business Model, etc), tener acceso flexible a los datos, para explorar y analizar sus relaciones, y consiguientes resultados.

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

- Esquema en estrella.
- Esquema copo de nieve.
- Esquema constelación o copo de estrellas.

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 semidesnormalizada, 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:

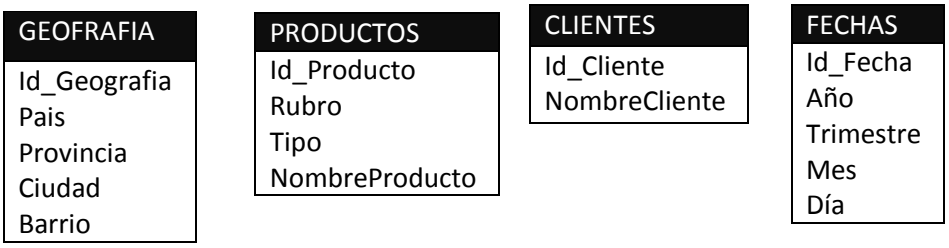
- Relacional - ROLAP
- Multidimensional - MOLAP
- Hibrido - HOLAP

2.1.1. TABLAS DE DIMENSIONES

Las tablas de dimensiones definen como están los datos organizados lógicamente y proveen el medio para analizar el contexto del negocio. Contienen datos cualitativos.

Representan los aspectos de interés, mediante los cuales los usuarios podrán filtrar y manipular la información almacenada en la tabla de hechos.

En la siguiente figura se pueden apreciar algunos ejemplos:



Como se puede observar, cada tabla posee un identificador único y al menos un campo o dato de referencia que describe los criterios de análisis relevantes para la organización, estos son por lo general de tipo texto.

Los datos dentro de estas tablas, que proveen información del negocio o que describen alguna de sus características, son llamados datos de referencia.

Más detalladamente, cada tabla de dimensión podrá contener los siguientes campos:

- Clave principal o identificador único
- Clave foránea

- Datos de referencia primarios: Datos que identifican la dimensión. Por ejemplo nombre del cliente.
- Datos de referencia secundarios: Datos que complementan la descripción de la dimensión. Por ejemplo email del cliente, fax del cliente, etc.

Usualmente la cantidad de tablas de dimensiones, aplicadas a un tema de interés en particular, varían entre tres y quince.

Debe tenerse en cuenta, que no siempre la clave primaria del OLTP, se corresponde con la clave primaria de la tabla de dimensión relacionada. Es recomendable manejar un sistema de claves en el DW (Claves Subrogadas) totalmente diferente al de los OLTP, ya que si estos últimos son recodificados, el almacén quedaría inconsistente y debería ser poblado nuevamente en su totalidad.

2.1.1.1. LA TABLA DIMENSIÓN TIEMPO

En un Data Warehouse, la creación y el mantenimiento de una tabla de dimensión **tiempo** es obligatoria, y la definición de granularidad y estructuración de la misma depende de la dinámica del negocio que se esté analizando. Toda la información dentro del depósito, como ya se ha explicado, posee su propio sello de tiempo que determina la ocurrencia de un hecho específico, representando de esta manera diferentes versiones de una misma situación.

Es importante tener en cuenta que la dimensión tiempo no es solo una secuencia cronológica representada de forma numérica, sino que mantiene niveles jerárquicos especiales que inciden notablemente en las actividades de la organización. Esto se debe a que los usuarios podrán por ejemplo analizar las ventas realizadas teniendo en cuenta el día de la semana en que se produjeron, quincena, mes, trimestre, semestre, año, estación, etc.

Existen muchas maneras de diseñar esta tabla, y en adición a ello no es una tarea sencilla de llevar a cabo. Por estas razones se considera una buena práctica evaluar con cuidado la temporalidad de los datos, la forma en que trabaja la organización, los resultados que se esperan obtener del almacén de datos relacionados con una unidad de tiempo y la flexibilidad que se desea obtener de dicha tabla.

Así mismo, si se requiere analizar los datos por Fecha (año, mes, día, etc) y por Hora (hora, minuto, segundo,

etc), lo más recomendable es confeccionar dos tablas de dimensión Tiempo; una contendrá los datos referidos a la Fecha y la otra los referidos a la Hora.

Si bien, el lenguaje SQL ofrece funciones del tipo DATE, en la tabla de dimensión Tiempo, se modelan y presentan datos temporales que no pueden calcularse a través de consultas SQL, lo cual le añade una ventaja más.

Es conveniente mantener en la tabla de dimensión **Tiempo** un campo que se refiera al día **Juliano**. El día juliano se representa a través de un número secuencial e identifica unívocamente cada día. Mantener este campo permitirá la posibilidad de realizar consultas que involucren condiciones de filtrado de fechas desde-hasta, mayor que, menor que, etc, de manera sencilla e intuitiva; ya que si por ejemplo a partir de tal fecha se desea analizar los datos de los 81 días siguientes, el valor "desde" sería el día Juliano de la fecha en cuestión y el valor "hasta" sería igual a "desde" más 81.

2.1.2. TABLAS DE HECHOS

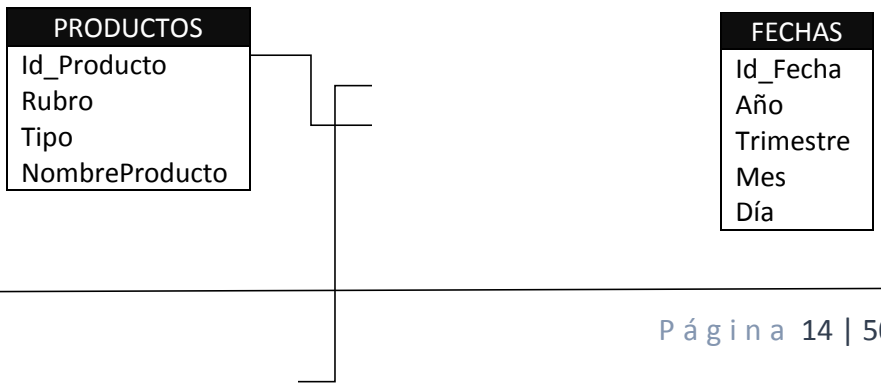
Las tablas de hechos contienen, precisamente, los hechos que serán utilizados por los analistas de negocio para apoyar el proceso de toma de decisiones. Contienen datos cuantitativos.

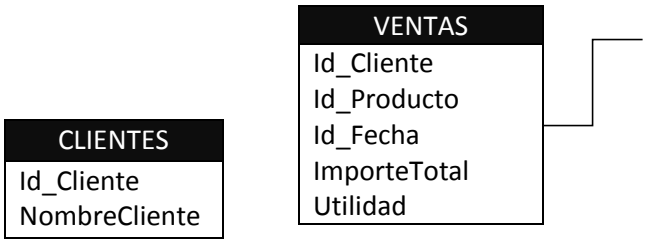
Los hechos son datos instantáneos en el tiempo, que son filtrados, agrupados y explorados a través de condiciones definidas en las tablas de dimensiones.

Los datos presentes en las tablas de hechos constituyen el volumen de la bodega, y pueden estar compuestos por millones de registros dependiendo de su granularidad y antigüedad de la organización. Los más importantes son los de tipo numérico.

El registro del hecho posee una clave primaria que está compuesta por las claves primarias de las tablas de dimensiones relacionadas a este.

En la siguiente imagen se puede apreciar un ejemplo de lo antes mencionado:





Como se muestra en la figura anterior, la tabla de hechos "VENTAS" se ubica en el centro, e irradiando de ella se encuentran las tablas de dimensiones "CLIENTES", "PRO- DUCTOS" y "FECHAS", que están conectadas mediante sus claves primarias. Es por ello que la clave primaria de la tabla de hechos es la combinación de las claves primarias de sus dimensiones. Los hechos en este caso son "ImporteTotal" y "Utilidad".

A continuación, se entrará más en detalle sobre la definición de un hecho, también llamado dato agregado:

Los hechos son aquellos datos que residen en una tabla de hechos y que son utilizados para crear indicadores, a través de sumarizaciones preestablecidas al momento de crear un cubo multidimensional, Business Model, etc. Debido a que una tabla de hechos se encuentra interrelacionada con sus respectivas tablas de dimensiones, permite que los hechos puedan ser accedidos, filtrados y explorados por los valores de los campos de estas tablas de dimensiones, obteniendo de este modo una gran capacidad analítica.

Las sumarizaciones no están referidas solo a sumas, sino también a promedios, mínimos, máximos, totales por sector, porcentajes, fórmulas predefinidas, etc, dependiendo de los requerimientos de información del negocio.

Para ejemplificar este nuevo concepto de hechos, se enumerarán algunos que son muy típicos y fáciles de comprender:

ImporteTotal = precioProducto X cantidadVendida

Rentabilidad = utilidad / PN

CantidadVentas = cantidad

PromedioGeneral = AVG(notasFinales)

A la izquierda de la igualdad se encuentran los hechos; a la derecha los campos de los OLTP que son utilizados para representarlos. En el último ejemplo se realiza un precálculo para establecer el hecho.

Existen dos tipos de hechos, los básicos y los derivados, a continuación se detallarán cada uno de ellos, teniendo en cuenta para su ejemplificación la siguiente tabla de hechos:

VENTAS
Id_Dimensión1
Id_Dimensión2
Id_DimensiónN
Precio
Cantidad
Total

Hechos básicos: son los que se encuentran representados por un campo de una tabla de hechos. Los campos "precio" y "cantidad" de la tabla anterior son hechos básicos.

Hechos derivados: son los que se forman al combinar uno o más hechos con alguna operación matemática o lógica y que también residen en una tabla de hechos. Estos poseen la ventaja de almacenarse previamente calculados, por lo cual pueden ser accedidos a través de consultas SQL sencillas y devolver resultados rápidamente, pero requieren más espacio físico en el DW, además de necesitar más tiempo de proceso en los ETL que los calculan. El campo "total" de la tabla anterior en un hecho derivado, ya que se conforma de la siguiente manera:

total = precio * cantidad

Los campos "precio" y "cantidad", también pertenecen a la tabla "HECHOS". Cabe resaltar, que no es necesario que los hechos derivados se compongan únicamente con hechos pertenecientes a una misma tabla.

Los hechos son gestionados con el principal objetivo de que se construyan indicadores basados en ellos, a través de la creación de un cubo multidimensional, Business Model, u otra estructura de datos.

2.2.1. TABLA DE HECHOS AGREGADAS Y PREAGREGADAS

Las tablas de hechos agregadas y preagregadas se utilizan para almacenar un resumen de los datos, es decir, se guardan los datos en niveles de granularidad superior a los que inicialmente fueron obtenidos y/o gestionados.

Para obtener tablas agregadas o preagregadas, es necesario establecer un criterio por el cual realizar el resumen. Por ejemplo, esto ocurre cuando se desea obtener información de ventas sumarizadas por mes.

Cada vez que se requiere que los datos en una consulta se presenten en un nivel de granularidad superior al que se encuentran alojados en el Data Warehouse, se debe llevar a cabo un proceso de agregación.

El objetivo general de las tablas de hechos agregadas y preagregadas es similar, pero cada una de ellas tiene una manera de operar diferente:

Tablas de hechos agregadas: se generan luego de que se procesa la consulta correspondiente a la tabla de hechos que se resumirá. En general, la agregación se produce dinámicamente a través de una instrucción SQL que incluya sumarizaciones.

Tablas de hechos preagregadas: se generan antes de que se procese la consulta correspondiente a la tabla de hechos que se resumirá. De esta manera, la consulta se realiza contra una tabla que ya fue previamente sumarizada. Habitualmente, estas sumarizaciones se calculan a través de procesos ETL.

Las tablas de hechos preagregadas cuentan con los siguientes beneficios:

- Reduce la utilización de recursos de hardware que normalmente son incurridos en el cálculo de las sumarizaciones.
- Reduce el número de registros que serán analizados por el usuario.
- Reduce el tiempo utilizado en la generación de consultas por parte del usuario.

Las tablas de hechos preagregadas son muy útiles en los siguientes casos generales:

- Cuando los datos a nivel detalle (menor nivel granular) son innecesarios y/o no son requeridos.
- Cuando una consulta sumarizada a determinado nivel de granularidad es solicitado con mucha frecuencia.
- Cuando los datos son muy abundantes, y las consultas demoran en ser procesadas demasiado tiempo.

Como contrapartida, las tablas de hechos preagregadas presentan una serie de desventajas:

- Requieren que se mantengan y gestionen nuevos procesos ETL.
- Demandan espacio de almacenamiento extra en el depósito de datos.

2.2. CUBO MULTIDIMENSIONAL

Si bien existen diversas estructuras de datos, a través de las cuales se puede representar los datos del Data Warehouse, solamente se entrará en detalle acerca de los cubos multidimensionales, por considerarse que esta estructura de datos es una de las más utilizadas y cuyo funcionamiento es el más complejo de entender.

Un cubo multidimensional o hipercubo, representa o convierte los datos planos que se encuentran en filas y columnas, en una matriz de N dimensiones.

Los objetos más importantes que se pueden incluir en un cubo multidimensional, son los siguientes:

Indicadores:
Sumarizaciones que se efectúan sobre algún hecho o expresiones basadas en sumarizaciones, pertenecientes a una tabla de hechos.

Atributos:
Campos o criterios de análisis, pertenecientes a tablas de dimensiones.

Jerarquías:
Representa una relación lógica entre dos o más atributos.

De esta manera en un cubo multidimensional, los atributos existen a lo largo de varios ejes o dimensiones, y la intersección de las mismas representa el valor que tomará el indicador que se está evaluando.

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

Para la creación del cubo de la figura anterior, se definieron tres Atributos (“Atributo 1”, “Atributo 2” y “Atributo 3”) y se definió un Indicador (“Indicador 1”). Entonces el cubo quedo compuesto por 3 dimensiones o ejes (una por cada Atributo), cada una con sus respectivos valores asociados. También, se ha seleccionado una intersección al azar para demostrar la correspondencia con los valores de las Atributos. En este caso,

el indicador “Indicador 1”, representa el cruce del Valor “5” de “Atributo 1”, con el Valor “4” de “Atributo 2” y con el Valor “3” de “Atributo 3”.

Se puede observar, que el resultado del análisis está dado por los cruces matriciales de acuerdo a los valores de las dimensiones seleccionadas.

Más específicamente, para acceder a los datos del DW, se pueden ejecutar consultas sobre algún cubo multidimensional previamente definido. Dicho cubo debe incluir entre otros objetos: indicadores, atributos, jerarquías, etc, basados en los campos de las tablas de dimensiones y de hechos, que se deseen analizar. De esta manera, las consultas son respondidas con gran performance, minimizando al máximo el tiempo que se hubiese incurrido en realizar dicha consulta sobre una base de datos transaccional.

2.2.1. Indicadores

Los indicadores son sumarizaciones efectuadas sobre algún hecho o expresiones basadas en sumarizaciones, que serán incluidos en algún cubo multidimensional, con el fin de analizar los datos almacenados en el Data Wareouse. El valor que estos adopten estará condicionado por los atributos/jerarquías que se utilicen para analizarlos.

Los indicadores, además de hechos, pueden estar compuestos por otros indicadores, pero no ambos simultáneamente. Pueden utilizarse para su creación funciones de sumarización (suma, conteo, promedio, etc), funciones matemáticas, estadísticas, operadores matemáticos y lógicos.

2.2.2. Atributos

Los atributos constituyen los criterios de análisis que se utilizarán para analizar los indicadores dentro de un cubo multidimensional. Los mismos se basan, en su gran mayoría, en los campos de las tablas de dimensiones y/o expresiones.

Dentro de un cubo multidimensional, los atributos son los ejes del mismo.

2.2.3. Jerarquías

Una jerarquía representa una relación lógica entre dos o más atributos pertenecientes a un cubo multidimensional; siempre y cuando posean su correspondiente relación “padre-hijo”.

- Las jerarquías poseen las siguientes características:
- Pueden existir varias en un mismo cubo.
- Están compuestas por dos o más niveles.
  Se tiene una relación “1-n” o “padre-hijo” entre atributos consecutivos de un nivel superior y uno inferior.
  Por lo general, las jerarquías pueden IDENTIFICARSE fácilmente, debido a que existen relaciones “1-n” o “padre-hijo” entre los propios atributos de un cubo.

La principal ventaja de manejar jerarquías, reside en poder analizar los datos desde su nivel más general al más detallado y viceversa, al desplazarse por los diferentes niveles.
La siguiente figura muestra un pequeño ejemplo de lo recién expuesto:

Aquí, se puede apreciar claramente cómo se construye una jerarquía:

1. Se crearon dos atributos, ”FECHA - Año” y ”FECHA - Mes”, los cuales están constituidos de la siguiente manera:
   ”FECHA - Año” = FECHA.Año “FECHA - Mes” = FECHA.Mes
   A la izquierda de la igualdad se encuentra el nombre del atributo; a la derecha el nombre del campo de la tabla de dimensión “FECHA”.
   2. Se creó la jerarquía llamada “Jerarquía Fechas”, en la cual se colocó el atributo más general en la cabecera y se comenzó a disgregar los niveles hacia abajo. En este caso, la figura se explica como sigue:
   Un mes del año pertenece solo a un año. Un año puede poseer uno o más meses del año.

2.2.4. Relación

Una relación representa la forma en que dos atributos interactúan dentro de una jerarquía. Existen básicamente dos tipos de relaciones:

a) Explícitas:

Son las más comunes y se pueden modelar a partir de atributos directos y están en línea continua de una jerarquía, por ejemplo, un país posee una o más provincias y una provincia pertenece solo a un país.

b) Implícitas:

Son las que ocurren en la vida real, pero su relación no es de vista directa, por ejemplo, una provincia tiene uno o más ríos, pero un río pertenece a una o más provincias. Como se puede observar, este caso se trata de una relación muchos a muchos.

2.2.5. Granularidad

La granularidad representa el nivel de detalle al que se desea almacenar la información sobre el negocio que se esté analizando. Por ejemplo, los datos referentes a ventas o compras realizadas por una empresa, pueden registrarse día a día, en cambio, los da- tos pertinentes a pagos de sueldos o cuotas de socios, podrán almacenarse a nivel de mes.

Mientras mayor sea el nivel de detalle de los datos, se tendrán mayores posibilidades analíticas, ya que los mismos podrán ser resumidos o sumarizados. Es decir, los datos que posean granularidad fina (nivel de detalle) podrán ser resumidos hasta obtener una granularidad media o gruesa. No sucede lo mismo en sentido contrario, ya que por ejem- plo, los datos almacenados con granularidad media podrán resumirse, pero no tendrán la facultad de ser analizados a nivel de detalle. O sea, si la granularidad con que se guardan los registros es a nivel de día, estos datos podrán sumarizarse por semana, mes, semes- tre y año, en cambio, si estos registros se almacenan a nivel de mes, podrán sumarizarse por semestre y año, pero no lo podrán hacer por día y semana.

2.3. TIPOS DE MODELAMIENTO DE UN DATAWAREHOUSE

2.3.1. ESQUEMA EN ESTRELLA

El esquema en estrella, consta de una tabla de hechos central y de varias tablas de dimensiones relacionadas a esta, a través de sus respectivas claves. En la siguiente figura se puede apreciar un esquema en estrella estándar:

El modelo ejemplificado cuando se abordó el tema de las tablas de hechos, es un esquema en estrella, por lo cual se lo volverá a mencionar para explicar sus cualidades.

Este modelo debe estar totalmente desnormalizado, es decir que no puede presentarse en tercera forma normal (3ra FN), es por ello que por ejemplo, la tabla de dimensión “PRODUCTOS” contiene los campos “Rubro”, “Tipo” y “NombreProducto”. Si se normaliza esta tabla, se obtendrá el siguiente resultado:

Cuando se normaliza, se pretende eliminar la redundancia, la repetición de datos y que las claves sean independientes de las columnas, pero en este tipo de modelos se requiere no evitar precisamente esto.

Las ventajas que trae aparejada la desnormalización, son las de obviar uniones (Join) entre las tablas cuando se realizan consultas, procurando así un mejor tiempo de respuesta y una mayor sencillez con respecto a su utilización. El punto en contra, es que se genera un cierto grado de redundancia, pero el ahorro de espacio no es significativo.

El esquema en estrella es el más simple de interpretar y optimiza los tiempos de respuesta ante las consultas de los usuarios. Este modelo es soportado por casi todas las herramientas de consulta y análisis, y los metadatos son

fáciles de documentar y mantener, sin embargo es el menos robusto para la carga y es el más lento de construir.

A continuación se destacarán algunas características de este modelo, que ayudarán a comprender mejor el por qué de sus ventajas:

- Posee los mejores tiempos de respuesta.
- Su diseño es fácilmente modificable.
- Existe paralelismo entre su diseño y la forma en que los usuarios visualizan y manipulan los datos.
- Simplifica el análisis.
- Facilita la interacción con herramientas de consulta y análisis.

2.3.2. ESQUEMA COPO DE NIEVE

Este esquema representa una extensión del modelo en estrella cuando las tablas de dimensiones se organizan en jerarquías de dimensiones.

Como se puede apreciar en la figura anterior, existe una tabla de hechos central que está relacionada con una o más tablas de dimensiones, quienes a su vez pueden estar relacionadas o no con una o más tablas de dimensiones.

Este modelo es más cercano a un modelo de entidad relación, que al modelo en estrella, debido a que sus tablas de dimensiones están normalizadas.

Una de los motivos principales de utilizar este tipo de modelo, es la posibilidad de segregar los datos de las tablas de dimensiones y proveer un esquema que sustente los requerimientos de diseño. Otra razón es que es muy flexible y puede implementarse después de que se haya desarrollado un esquema en estrella.

Se pueden definir las siguientes características de este tipo de modelo:

- Posee mayor complejidad en su estructura.
- Hace una mejor utilización del espacio.
- Es muy útil en tablas de dimensiones de muchas tuplas.
- Las tablas de dimensiones están normalizadas, por lo que requiere menos esfuerzo de diseño.
- Puede desarrollar clases de jerarquías fuera de las tablas de dimensiones, que per- miten realizar análisis de lo general a lo detallado y viceversa.

A pesar de todas las características y ventajas que trae aparejada la implementación del esquema copo de nieve, existen dos grandes inconvenientes de ello:

- Si se poseen múltiples tablas de dimensiones, cada una de ellas con varias jerarquías, se creará un número de tablas bastante considerable, que pueden llegar al punto de ser inmanejables.
- Al existir muchas uniones y relaciones entre tablas, el desempeño puede verse reducido.

La existencia de las diferentes jerarquías de dimensiones debe estar bien fundamentada, ya que de otro modo las consultas demorarán más tiempo en devolver los resultados, debido a que se deben realizar las uniones entre las tablas.

2.3.3. ESQUEMA CONSTELACION

Este modelo está compuesto por una serie de esquemas en estrella, y tal como se puede apreciar en la siguiente figura, está formado por una tabla de hechos principal ("HECHOS_A") y por una o más tablas de hechos auxiliares ("HECHOS_B"), las cuales pueden ser sumarizaciones de la principal. Dichas tablas yacen en el centro del modelo y están relacionadas con sus respectivas tablas de dimensiones.

No es necesario que las diferentes tablas de hechos compartan las mismas tablas de dimensiones, ya que, las tablas de hechos auxiliares pueden vincularse con solo algunas de las tablas de dimensiones asignadas a la tabla de hechos principal, y también pueden hacerlo con nuevas tablas de dimensiones.

Su diseño y cualidades son muy similares a las del esquema en estrella, pero posee una serie de diferencias con el mismo, que son precisamente las que lo destacan y caracterizan.

Entre ellas se pueden mencionar:

- Permite tener más de una tabla de hechos, por lo cual se podrán analizar más as- pectos claves del negocio con un mínimo esfuerzo adicional de diseño.
- Contribuye a la reutilización de las tablas de dimensiones, ya que una misma tabla de dimensión puede utilizarse para varias tablas de hechos.
- No es soportado por todas las herramientas de consulta y análisis.

2.4. OPERACIONES EN CUBOS OLAP

Las operaciones que se pueden realizar sobre modelos multidimensionales y que son las que verdaderamente les permitirán a los usuarios explorar e investigar los datos en busca de respuestas, son:

- Drill-down.
- Drill-up.
- Drill-across.
- Roll-across.
- Pivot.
- Page.
- Drill-through.

A continuación, se explicará cada una de ellas y se ejemplificará su utilización, para lo cual se utilizará como guía el siguiente esquema en estrella.

El mismo posee cuatro tablas de dimensiones y una tabla de hechos central, en la cual el hecho "Venta" representa las ventas a un cliente, de un producto en particular, de una marca específica en un año dado.

Sobre este modelo, entonces, se creará un cubo llamado "Cubo - Query Manager". El mismo contiene los siguientes objetos:

De la tabla de hechos "VENTAS", se sumarizará el hecho "Venta" para crear el indicador denominado:

• "VENTAS - Venta".

La fórmula utilizada para crear este indicador es la siguiente:

• "VENTAS - Venta" = SUM(VENTAS.Venta).

De la tabla de dimensión "MARCAS", se tomará el campo "Marca" para la creación del atributo denominado:

• "MARCAS - Marca".

De la tabla dimensión "TIEMPO", se tomará el campo "Año" para la creación del atributo denominado:

• "TIEMPO - Año".

De la tabla dimensión "PRODUCTOS", se tomará el campo "Producto" para la creación del atributo denominado:

• "PRODUCTOS - Producto".

De la tabla dimensión "PRODUCTOS", se tomará el campo "Clase" para la creación del atributo denominado:

• "PRODUCTOS - Clase".

Se definió la jerarquía “Jerarquía PRODUCTOS”, que se aplicará sobre los atributos recientemente creados, “PRODUCTOS - Producto” y “PRODUCTOS - Clase”, en donde:

• Una clase de producto pertenece solo a un producto. Un producto puede ser de una o más clases.

2.4.1. Drill-Down:

Permite apreciar los datos en un mayor detalle, bajando por una jerarquía definida en un cubo. Esto brinda la posibilidad de introducir un nuevo nivel o criterio de agregación en el análisis, disgregando los grupos actuales. Drill-Down es ir de lo general a lo especifico. Gráficamente:

Para explicar esta operación se utilizará la siguiente representación tabular:

Como puede apreciarse, en la cabecera de la tabla se encuentran los atributos y el indicador (destacado con color de fondo diferente) definidos anteriormente en el cubo multidimensional; y en el cuerpo de la misma se encuentran los valores correspondientes. Se ha resaltado la primera fila, ya que es la que se analizará más en detalle.
En este caso, se realizará Drill-Down sobre la jerarquía “Jerarquía PRODUCTOS”, entonces:

Tal y como puede apreciarse en los ítems resaltados de la tabla, se agregó un nuevo nivel de detalle (“PRODUCTOS – Clase”) a la lista inicial, y el valor “40” que pertenecía a las ventas del “Producto1”, de la marca “M1”, en el año “2007”, se dividió en dos filas. Esto se debe a que ahora se tendrá en cuenta el atributo “PRODUCTOS - Clase” para realizar las sumarizaciones del indicador “VENTAS - Venta”.
La siguiente imagen muestra este mismo proceso pero, representado matricialmente:

De aquí en más se utilizará esta forma para explicar cada operación.

2.4.2. Drill-Up

Permite apreciar los datos en menor nivel de detalle, subiendo por una jerarquía definida en un cubo. Esto brinda la posibilidad de quitar un nivel o criterio de agregación en el análisis, agregando los grupos actuales.
Drill-up es ir de lo específico a lo general.

2.4.3. Drill-Across

Funciona de forma similar a drill-down, con la diferencia de que drill-across no se realiza sobre una jerarquía, sino que su forma de ir de lo general a lo específico es agregar un atributo a la consulta como nuevo criterio de análisis.

2.4.4. Roll-across

Funciona de forma similar a drill-up, con la diferencia de que roll-across no se hace sobre una jerarquía, sino que su forma de ir de lo específico a lo general es quitar un atributo de la consulta, eliminando de esta manera un criterio de análisis.

2.4.5. Pivot

Permite seleccionar el orden de visualización de los atributos e indicadores, con el objetivo de analizar la información desde diferentes perspectivas.

Pivot permite realizar las siguientes acciones:

- Mover un atributo o indicador desde el encabezado de fila al encabezado de columna.
- Mover un atributo o indicador desde el encabezado de columna al encabezado de fila.
- Cambiar el orden de los atributos o indicadores del encabezado de columna.
- Cambiar el orden de los atributos o indicadores del encabezado de fila.

2.4.6. Page

Presenta el cubo dividido en secciones, a través de los valores de un atributo, como si se tratase de páginas de un libro.
Page es muy útil cuando las consultas devuelven muchos registros y es necesario desplazarse por los datos para poder verlos en su totalidad.
Cuando existe más de un criterio por el cual realizar Page, debe tenerse en cuenta el orden en que estos serán procesados, ya que dependiendo de esto, de podrán obtener diferentes resultados sobre una misma consulta. Para ejemplificar este concepto se utilizará como base la tabla expuesta al inicio.

2.4.7. Drill-through

Permite apreciar los datos en su máximo nivel de detalle. Esto brinda la posibilidad de analizar cuáles son los datos

relacionados al valor de un Indicador, que se ha sumarizado dentro del cubo multidimensional.

2.5. PROYECTOS DE DATA WAREHOUSE

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

Con el fin de que se llegue a una total comprensión de cada paso o etapa, se acompañará con la implementación en una empresa real, para demostrar los resultados que se deben obtener y ejemplificar cada concepto.

2.5.1. ANALISIS DE REQUERIMIENTOS

Lo primero que se hará será identificar los requerimientos de los usuarios a través de preguntas que expliciten los objetivos de su organización. Luego, se analizarán estas preguntas a fin de identificar cuáles serán los indicadores y perspectivas que serán tomadas en cuenta para la construcción del Data Warehouse. Finalmente se confeccionará un modelo conceptual en donde se podrá visualizar el resultado obtenido en este primer paso.

2.5.1.1. IDENTIFICAR PREGUNTAS

El primer paso comienza con el acopio de las necesidades de información, el cual puede llevarse a cabo a través de muy variadas y diferentes técnicas, cada una de las cuales poseen características inherentes y específicas, como por ejemplo entrevistas, cuestionarios, observaciones, etc.

El análisis de los requerimientos de los diferentes usuarios, es el punto de partida, ya que ellos son los que deben, en cierto modo, guiar la investigación hacia un desarrollo que refleje claramente lo que se espera del depósito de datos, en relación a sus funciones y cualidades.

El objetivo principal de esta fase, es la de obtener e identificar las necesidades de información clave de alto nivel, que es esencial para llevar a cabo las metas y estrategias de la empresa, y que facilitará una eficaz y eficiente toma de decisiones.

Debe tenerse en cuenta que dicha información, es la que proveerá el soporte para desarrollar los pasos sucesivos, por lo cual, es muy importante que se preste especial atención al relevar los datos.

Una forma de asegurarse de que se ha realizado un buen análisis, es corroborar que el resultado del mismo haga explícitos los objetivos estratégicos planteados por la empresa que se está estudiando.

Otra forma de encaminar el relevamiento, es enfocar las necesidades de información en los procesos principales que desarrolle la empresa en cuestión.

La idea central es, que se formulen preguntas complejas sobre el negocio, que incluyan variables de análisis que se consideren relevantes, ya que son estas las que permitirán estudiar la información desde diferentes perspectivas.

Un punto importante que debe tenerse muy en cuenta, es que la información debe estar soportada de alguna manera por algún OLTP, ya que de otra forma, no se podrá elaborar el Data Warehouse.

2.5.1.2. IDENTIFICAR INDICADORES Y PERSPECTIVAS

Una vez que se han establecido las preguntas de negocio, se debe proceder a su descomposición para descubrir los indicadores que se utilizarán y las perspectivas de análisis que intervendrán.

Para ello, se debe tener en cuenta que los indicadores, para que sean realmente efectivos son, en general, valores numéricos y representan lo que se desea analizar concretamente, por ejemplo: saldos, promedios, cantidades, sumatorias, fórmulas, etc.

En cambio, las perspectivas se refieren a los objetos mediante los cuales se quiere examinar los indicadores, con el fin de responder a las preguntas planteadas, por ejemplo: clientes, proveedores, sucursales, países, productos, rubros, etc. Cabe destacar, que el Tiempo es muy comúnmente una perspectiva.

2.5.1.3. MODELO CONCEPTUAL

En esta etapa, se construirá un modelo conceptual a partir de los indicadores y perspectivas obtenidas en el paso anterior.

A través de este modelo, se podrá observar con claridad cuáles son los alcances del proyecto, para luego poder trabajar sobre ellos, además al poseer un alto nivel de definición de los datos, permite que pueda ser presentado ante los usuarios y explicado con facilidad.

A la izquierda se colocan las perspectivas seleccionadas, que serán unidas a un óvalo central que representa y lleva el nombre de la relación que existe entre ellas. La relación, constituye el proceso o área de estudio elegida. De dicha relación y entrelazadas con flechas, se desprenden los indicadores, estos se ubican a la derecha del esquema.

Como puede apreciarse en la figura anterior, el modelo conceptual permite de un solo vistazo y sin poseer demasiados conocimientos previos, comprender cuáles serán los resultados que se obtendrán, cuáles serán las variables que se utilizarán para analizarlos y cuál es la relación que existe entre ellos.

2.5.2. ANALISIS DE LOS OLTP

Seguidamente, se analizarán las fuentes OLTP para determinar cómo serán calculados los indicadores y para establecer las respectivas correspondencias entre el modelo conceptual creado en el paso anterior y las fuentes de datos. Luego, se definirán qué campos se incluirán en cada perspectiva. Finalmente, se ampliará el modelo conceptual con la información obtenida en este paso.

2.5.2.1. CONFORMAR INDICADORES

En este paso se deberán explicitar cómo se calcularán los indicadores, definiendo los siguientes conceptos para cada uno de ellos:

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

2.5.2.2. ESTABLECER CORRESPONDENCIAS

El objetivo de este paso, es el de examinar los OLTP disponibles que contengan la información requerida, como así también sus características, para poder

identificar las correspondencias entre el modelo conceptual y las fuentes de datos.

La idea es, que todos los elementos del modelo conceptual estén correspondidos en los OLTP.

2.5.2.3. NIVEL DE GRANULARIDAD

Una vez que se han establecido las relaciones con los OLTP, se deben seleccionar los campos que contendrá cada perspectiva, ya que será a través de estos por los que se examinarán y filtrarán los indicadores.

Para ello, basándose en las correspondencias establecidas en el paso anterior, se debe presentar a los usuarios los datos de análisis disponibles para cada perspectiva. Es muy importante conocer en detalle que significa cada campo y/o valor de los datos encontrados en los OLTP, por lo cual, es conveniente investigar su sentido, ya sea a través de diccionarios de datos, reuniones con los encargados del sistema, análisis de los datos propiamente dichos, etc.

Luego de exponer frente a los usuarios los datos existentes, explicando su significado, valores posibles y características, estos deben decidir cuales son los que consideran relevantes para consultar los indicadores y cuales no.

Con respecto a la perspectiva “Tiempo”, es muy importante definir el ámbito mediante el cual se agruparán o sumarizarán los datos. Sus campos posibles pueden ser: día de la semana, quincena, mes, trimestres, semestre, año, etc.

Al momento de seleccionar los campos que integrarán cada perspectiva, debe prestarse mucha atención, ya que esta acción determinará la granularidad de la información encontrada en el DW.

2.5.2.4. EL MODELO CONCEPTUAL AMPLIADO

En este paso, y con el fin de graficar los resultados obtenidos en los pasos anteriores, se ampliará el modelo conceptual, colocando bajo cada perspectiva los campos seleccio- nados y bajo cada indicador su respectiva fórmula de cálculo.

2.5.3. MODELO LOGICO DEL DATA WAREHOUSE

A continuación, se confeccionará el modelo lógico de la estructura del Data Warehouse, teniendo como base el modelo conceptual que ya ha sido creado. Para ello, primero se definirá el tipo de modelo que se utilizará y luego se llevarán a cabo las acciones propias al caso, para diseñar las tablas de dimensiones y de hechos. Finalmente, se realizarán las uniones pertinentes entre estas tablas.

2.5.3.1. TIPO DE MODELO LOGICO DEL DATA WAREHOUSE

Se debe seleccionar cuál será el tipo de esquema que se utilizará para contener la estructura del depósito de datos, que se adapte mejor a los requerimientos y necesidades de los usuarios. Es muy importante definir objetivamente si se empleará un esquema en estrella, constelación o copo de nieve, ya que esta decisión afectará considerablemente la elaboración del modelo lógico.

2.5.3.2. TABLAS DIMENSIONALES

En este paso se deben diseñar las tablas de dimensiones que formaran parte del Data Warehouse.

Para los tres tipos de esquemas, cada perspectiva definida en el modelo conceptual constituirá una tabla de dimensión. Para ello deberá tomarse cada perspectiva con sus campos relacionados y realizarse el siguiente proceso: Se elegirá un nombre que identifique la tabla de dimensión. Se añadirá un campo que represente su clave principal. Se redefinirán los nombres de los campos si es que no son lo suficientemente intuitivos.

Para los esquemas copo de nieve, cuando existan jerarquías dentro de una tabla de dimensión, esta tabla deberá ser normalizada.

2.5.3.3. TABLAS DE HECHOS

En este paso, se definirán las tablas de hechos, que son las que contendrán los hechos a través de los cuales se construirán los indicadores de estudio.

Para los esquemas en estrella y copo de nieve, se realizará lo siguiente:

- Se le deberá asignar un nombre a la tabla de hechos que represente la información

analizada, área de investigación, negocio enfocado, etc.

- Se definirá su clave primaria, que se compone de la combinación de las claves primarias de cada tabla de dimensión relacionada.
- Se crearán tantos campos de hechos como indicadores se hayan definido en el modelo conceptual y se les asignará los mismos nombres que estos. En caso que se prefiera, podrán ser nombrados de cualquier otro modo.

2.5.3.4. UNIONES

Para los tres tipos de esquemas, se realizarán las uniones correspondientes entre sus tablas de dimensiones y sus tablas de hechos.

2.5.4. INTEGRACION DE LOS DATOS

Una vez construido el modelo lógico, se deberá proceder a poblarlo con datos, utilizando técnicas de limpieza y calidad de datos, procesos ETL, etc.; luego se definirán las reglas y políticas para su respectiva actualización, así como también los procesos que la llevarán a cabo.

2.5.4.1. CARGA INICIAL

Una vez construido el modelo lógico, se deberá proceder a poblarlo con datos, utilizando técnicas de limpieza y calidad de datos, procesos ETL, etc.; luego se definirán las reglas y políticas para su respectiva actualización, así como también los procesos que la llevarán a cabo.

La realización de estas tareas pueden contener una lógica realmente compleja en algunos casos. Afortunadamente, en la actualidad existen muchos softwares que se pueden emplear a tal fin, y que nos facilitarán el trabajo.

Se debe evitar que el Data Warehouse sea cargado con valores faltantes o anómalos, así como también se deben establecer condiciones y restricciones para asegurar que solo se utilicen los datos de interés.

2.5.4.2. ACTUALIZACION DE LOS DATOS

Cuando se haya cargado en su totalidad el Data Warehouse, se deben establecer sus políticas y estrategias de actualización o refresco de datos.

Una vez realizado esto, se tendrán que llevar a cabo las siguientes acciones:

- Especificar las tareas de limpieza de datos, calidad de datos, procesos ETL, etc., que deberán realizarse para actualizar los datos del DW.
- Especificar de forma general y detallada las acciones que deberá realizar cada software.

2.6. DISEÑO FISICO DE UN DATA WAREHOUSE

El diseño físico es la creación de la base de datos con sentencias SQL. Durante el proceso de diseño físico, convertir los datos recogidos durante la fase de diseño lógico en una descripción de la estructura de base de datos física.

Decisiones de diseño físicas son impulsadas principalmente por los aspectos de rendimiento de consulta y mantenimiento de bases de datos.

Diseño físico define un modelo para su almacén de datos que consta de entidades, atributos y relaciones.

Las entidades están unidos entre sí mediante relaciones. Los atributos se utilizan para describir las entidades. El identificador único (UID) distingue entre una instancia de una entidad y otra.

Durante el proceso de diseño físico, habrá que traducir los esquemas previstos en las estructuras de bases de datos reales. En este momento, usted tiene que asignar:

- Las entidades a las tablas.
- Relaciones con restricciones de clave externa.
- Atributos de columnas.
- identificadores únicos primarios a restricciones de clave principal.
- Los identificadores únicos a las restricciones de claves únicas.

Una vez que haya convertido su diseño lógico a uno físico, tendrá que crear algunas o todas de las siguientes estructuras:

- Tablespaces.
- Tables and Partitioned Tables.
- Views. Integrity Constraints.
- Dimensions.

Algunas de estas estructuras requieren espacio en disco. Otros existen sólo en el diccionario de datos. Además, las siguientes estructuras se pueden crear para mejorar el rendimiento:

- Indexes and Partitioned Indexes.
- Materialized Views.

2.6.1. TABLESPACES

Un **Tablespaces** se compone de uno o más archivos de datos, que son estructuras físicas dentro del sistema operativo que esté utilizando. Un archivo de datos se asocia con un único

Tablespaces. Desde una perspectiva de diseño, los **Tablespaces** son contenedores para las estructuras de diseño físico.

Los espacios de tabla también deben representar unidades de negocio lógicas si es posible. Debido a un **Tablespaces** es la granularidad más gruesa para backup y recuperación o el mecanismo **tablespaces** transportables, el diseño de negocio lógico afecta a las operaciones de la disponibilidad y mantenimiento. Ahora puede utilizar los archivos de datos ultra grandes, una mejora significativa en grandes bases de datos.

2.6.2. TABLES AND PARTITIONED TABLES

Las **tables** son la unidad básica de almacenamiento de datos. Ellos son el recipiente para la cantidad que se espera de los datos en bruto en su almacén de datos. El uso de **partitioned tables** en lugar de los sin particiones aborda el problema fundamental de apoyo a los volúmenes de datos muy grandes por lo que le permite dividir en partes más pequeñas y más manejables. El criterio principal de diseño para la partición es de gestión, aunque también verá los beneficios de rendimiento en la mayoría de los casos debido a la poda partición o procesamiento paralelo inteligente. Por ejemplo, puede elegir una estrategia de partición basado en una fecha de la transacción de venta y una granularidad mensual. Si usted tiene cuatro años de valor de los datos, puede borrar los datos de un mes, ya que se convierte en más de cuatro años con una sola, sentencia DDL rápido y cargar nuevos datos mientras que sólo afecta a 1/48 de la tabla completa. Preguntas de negocio con respecto al último trimestre sólo afectarán a los tres meses, lo que equivale a tres particiones, o 3 / 48ths del volumen total. Partición de tablas de gran tamaño mejora el rendimiento, ya que cada pieza particionado es más manejable. Típicamente, particionar basado en fechas de las transacciones en un almacén de datos. Por ejemplo, cada mes, el valor de un mes de datos se puede asignar su propia partición.

2.6.3. VIEWS

Una view es una presentación a la medida de los datos contenidos en una o varias tablas u otras vistas. Una vista toma la salida de una consulta y la trata como una tabla. Las vistas no requieren ningún espacio en la base de datos.

2.6.4. INTEGRITY CONSTRAINTS

Las restricciones de integridad se utilizan para hacer cumplir las reglas de negocio asociadas a su base de datos y para evitar tener información no válida en las tablas. Las restricciones de integridad en el Data Warehouse difieren de las limitaciones en los entornos OLTP. En entornos OLTP, evitan principalmente la inserción de datos no válidos en un registro, que no es un gran problema en entornos de Data Warehouse porque la precisión ya ha sido garantizada. En entornos de Data Warehouse, las restricciones sólo se utilizan para reescritura de consultas. NO restricciones NULL son particularmente comunes en los Data Warehouse. Bajo algunas circunstancias específicas, las limitaciones necesitan espacio en la base de datos.

2.6.5. INDEXES AND PARTITIONED INDEXES

Los índices son estructuras opcionales asociados con tablas o clusters. Además de los índices B-tree clásico, los índices de mapa de bits son muy comunes en los entornos de Data Warehouse. Índices de mapa de bits se optimizan las estructuras de índices para las operaciones de ajuste orientadas. Además, son necesarios para algunos métodos optimizados de acceso a datos tales como transformaciones estrellas. Los índices son como si fueran tablas en los que las tablas puede particionarse, aunque la estrategia de partición no depende de la estructura de la tabla. Particionar índices hace que sea más fácil de administrar el Data Warehouse durante la actualización y mejora el rendimiento de la consulta.

2.6.6. MATERIALIZED VIEWS

Vistas materializadas son resultados de la consulta que se han almacenado con antelación para los cálculos de larga ejecución no son necesarios cuando realmente ejecuta las sentencias SQL. Desde un punto de diseño de vista físico, vistas materializadas se asemejan a **Tables and Partitioned Tables** y se comportan como los índices en que se utilizan de forma transparente y mejoran el rendimiento.

3. INTEGRACIÓN DE DATOS PARA UN DATA WAREHOUSE

Para poder extraer los datos desde los OLTP, para luego manipularlos, integrarlos y transformarlos, para posteriormente cargar los resultados obtenidos en el Data Warehouse, es necesario contar con algún sistema que se encargue de ello. Precisamente, la Integración de Datos es quien cumplirá con tal fin.

La Integración de Datos agrupa una serie de técnicas y subprocesos que se encargan de llevar a cabo todas las tareas relacionadas con la extracción, manipulación, control, integración, depuración de datos, carga y actualización del Data Warehouse, etc. Es decir, todas las tareas que se realizarán desde que se toman los datos de los diferentes OLTP hasta que se cargan en el Data Warehouse.

Si bien los procesos ETL (Extracción, Transformación y Carga) son solo una de las muchas técnicas de la Integración de Datos, el resto de estas técnicas puede agruparse muy bien en sus diferentes etapas. Es decir, en el proceso de Extracción tendremos un grupo de técnicas enfocadas por ejemplo en tomar solo los datos indicados y mantenerlos en un almacenamiento intermedio; en el proceso de Transformación por ejemplo estarán aquellas técnicas que analizarán los datos para verificar que sean correctos y válidos; en el proceso de Carga de Datos se agruparán por ejemplo técnicas propias de la carga y actualización del Data Warehouse.

A continuación, se detallará cada una de estas etapas, se expondrá cuál es el proceso que llevan a cabo los ETL y se enumerarán cuáles son sus principales tareas.

3.1. EXTRACION

Es aquí, en donde, basándose en las necesidades y requisitos de los usuarios, se exploran las diversas fuentes OLTP que se tengan a disposición, y se extrae la información que se considere relevante al caso.

Si los datos operacionales residen en un SGBD Relacional, el proceso de extracción se puede reducir a, por ejemplo, consultas en SQL o rutinas programadas. En cambio, si se encuentran en un sistema no convencional o fuentes externas, ya sean textuales, hipertextuales, hojas de cálculos, etc, la obtención de los mismos puede ser un tanto más dificultoso, debido a que, por ejemplo, se tendrán que realizar cambios de formato y/o volcado de información a partir de alguna herramienta específica.

Una vez que los datos son seleccionados y extraídos, se guardan en un almacenamiento intermedio, lo cual permite, entre otras ventajas:

- Manipular los datos sin interrumpir ni paralizar los OLTP, ni tampoco el DW.
- No depender de la disponibilidad de los OLTP.
- Almacenar y gestionar los metadatos que se generarán en los procesos ETL.
- Facilitar la integración de las diversas fuentes, internas y externas.

El almacenamiento intermedio constituye en la mayoría de los casos una base de datos en donde la información puede ser almacenada por ejemplo en tablas auxiliares, tablas temporales, etc. Los datos de estas tablas serán los que finalmente (luego de su correspondiente transformación) poblarán el Data Warehouse.

3.2. TRANSFORMACION

Esta función es la encargada de convertir aquellos datos inconsistentes en un conjunto de datos compatibles y congruentes, para que puedan ser cargados en el Data Warehouse. Estas acciones se llevan a cabo, debido a que pueden existir diferentes fuentes de información, y es vital conciliar un formato y forma única, definiendo estándares, para que todos los datos que ingresarán al Data Warehouse estén integrados.

Los casos más comunes en los que se deberá realizar integración, son los siguientes:

- Codificación.
- Medida de atributos.
- Convenciones de nombramiento.
- Fuentes múltiples.

Además de lo antes mencionado, esta función se encarga de realizar, entre otros, los procesos de Limpieza de Datos (Data Cleansing) y Calidad de Datos.

3.2.1. CODIFICACION

Una inconsistencia muy típica que se encuentra al intentar integrar varias fuentes de datos, es la de contar con más de una forma de codificar un atributo en común. Por ejemplo, en el campo “estado”, algunos diseñadores completan su valor con “0” y “1”, otros con “Apagado” y “Encendido”, otros con “off” y “on”, etc. Lo que se debe realizar en estos casos, es seleccionar o recodificar estos atributos, para que cuando la información llegue al Data Warehouse, esté integrada de manera uniforme.

3.2.2. MEDIDA DE ATRIBUTOS

Los tipos de unidades de medidas utilizados para representar los atributos de una entidad, varían considerablemente entre sí, a través de los diferentes OLTP. Por ejemplo, al registrar la longitud de un producto determinado, de acuerdo a la aplicación que se emplee para tal fin, las unidades de medidas pueden ser presentadas en centímetros, metros, pulgadas, etc.En esta ocasión, se deberán estandarizar las unidades de medidas de los atributos, para que todas las fuentes de datos expresen sus valores de igual manera. Los algoritmos que resuelven estas inconsistencias son generalmente los más complejos.

3.2.3. CONVENCIONES DE NOMBRAMIENTOS

Usualmente, un mismo atributo es nombrado de diversas maneras en los diferentes OLTP. Por ejemplo, al referirse al

nombre del proveedor, puede hacerse como “nombre”, “razón_social”, “proveedor”, etc. Aquí, se debe utilizar la convención de nombramiento que para los usuarios sea más comprensible.

3.2.4. FUENTES MULTIPLES

Un mismo elemento puede derivarse desde varias fuentes. En este caso, se debe elegir aquella fuente que se considere más fiable y apropiada

3.2.5. LIMPIEZA DE DATOS

Su objetivo principal es el de realizar distintos tipos de acciones contra el mayor número de datos erróneos, inconsistentes e irrelevantes.

Las acciones más típicas que se pueden llevar a cabo al encontrarse con Datos Anómalos (Outliers) son:

- Ignorarlos.
- Eliminar la columna.
- Filtrar la columna.
- Filtrar la fila errónea, ya que a veces su origen, se debe a casos especiales.
- Reemplazar el valor.
- Discretizar los valores de las columnas. Por ejemplo de 1 a 2, poner “bajo”; de 3 a 7, “óptimo”; de 8 a 10, “alto”. Para que los outliers caigan en “bajo” o en “alto” sin mayores problemas.

Las acciones que suelen efectuarse contra Datos Faltantes (Missing Values) son:

- Ignorarlos.
- Eliminar la columna.
- Filtrar la columna.
- Filtrar la fila errónea, ya que a veces su origen, se debe a casos especiales.
- Reemplazar el valor.
- Esperar hasta que los datos faltantes estén disponibles.

Un punto muy importante que se debe tener en cuenta al elegir alguna acción, es el de identificar el por qué de la anomalía, para luego actuar en consecuencia, con el fin de evitar que se repitan, agregándole de esta manera más valor a los datos de la organización.

3.3. CARGA

Esta función se encarga, por un lado de realizar las tareas relacionadas con:

- Carga Inicial (Initial Load).
- Actualización o mantenimiento periódico (siempre teniendo en cuenta un intervalo de tiempo predefinido para tal operación).

La carga inicial, se refiere precisamente a la primera carga de datos que se le realizará al Data Warehose. Por lo general, esta tarea consume un tiempo bastante considerable, ya que se deben insertar registros que han sido generados aproximadamente, y en casos ideales, durante más de cinco años.

Los mantenimientos periódicos mueven pequeños volúmenes de datos, y su frecuencia está dada en función del gránulo del Data Warehouse y los requerimientos de los usuarios. El objetivo de esta tarea es añadir al depósito aquellos datos nuevos que se fueron generando desde el último refresco.

Antes de realizar una nueva actualización, es necesario identificar si se han producido cambios en las fuentes originales de los datos recogidos, desde la fecha del último mantenimiento, a fin de no atentar contra la consistencia del Data Warehouse. Para efectuar esta operación, se pueden realizar las siguientes acciones:

- Cotejar las instancias de los OLTP involucrados.
- Utilizar disparadores en los OLTP.
- Recurrir a Marcas de Tiempo (Time Stamp), en los registros de los OLTP.
- Comparar los datos existentes en los dos ambientes (OLTP y DW).
- Hacer uso de técnicas mixtas.

Si este control consume demasiado tiempo y esfuerzo, o simplemente no puede llevarse a cabo por algún motivo en particular, existe la posibilidad de cargar el Data Warehouse desde cero, este proceso se denomina Carga Total (Full Load).

Ingresarán al DW, para su carga y/o actualización:

- Aquellos datos que han sido transformados y que residen en el almacenamiento intermedio.
- Aquellos datos de los OLTP que tienen correspondencia directa con el depósito de datos.

Se debe tener en cuenta, que los datos antes de moverse al almacén de datos, deben ser analizados con el propósito de asegurar su calidad, ya que este es un factor clave, que no debe dejarse de lado.

Por otra parte, el proceso de Carga tiene la tarea de mantener la estructura del Data Warehouse, y trata temas relacionados con:

- Relaciones muchos a muchos.
- Claves Subrogadas.
- Dimensiones Lentamente Cambiantes
- Dimensiones Degeneradas.

4. CONSULTAS AL DATA WAREHOUSE

Las herramientas de consulta y análisis son sistemas que permiten a los usuarios realizar la exploración de datos del Data Warehouse. Básicamente constituyen el nexo entre el depósito de datos y los usuarios.

Utilizan la metadata de las estructuras de datos que han sido creadas previamente (cubos multidimensionales, Business Models, etc.) para trasladar a través de consultas SQL los requerimientos de los usuarios, para luego, devolver el resultado obtenido.

Estas herramientas también pueden emplear simples conexiones a bases de datos (JNDI, JDBC, ODBC), para obtener la información deseada.

A través de una interfaz gráfica y una serie de pasos, los usuarios generan consultas que son enviadas desde la herramienta de consulta y análisis, este a su vez realiza la extracción de información al Data Warehouse y devuelve los resultados obtenidos a la herramienta que se los solicitó. Luego, estos resultados son expuestos ante los usuarios en formatos que le son familiares.

El mismo, se lleva a cabo a través de seis pasos sucesivos:

1. Los usuarios seleccionan o establecen que datos desean obtener del Data Warehouse, mediante las interfaces de la herramienta que utilice.
2. La herramienta recibe el pedido de los usuarios, construye la consulta (utilizando la metadata) y la envía al Query Manager.
3. El Query Manager ejecuta la consulta sobre la estructura de datos con la que se esté trabajando (cubo multidimensional, Business Model, etc.).
4. El Query Manager obtiene los resultados de la consulta.
5. El Query Manager envía los datos a la herramienta de consulta y análisis.
6. La herramienta presentan a los usuarios la información requerida.

Una de las principales ventajas de utilizar estas herramientas, es que los usuarios no se tienen que preocupar por conocer cuáles son las características y funcionalidades de las estructuras de datos utilizadas, ni por saber emplear el lenguaje SQL, solo se deben enfocar en el análisis.

Las herramientas de consulta y análisis, en general, comparten las siguientes características:

- **Accesibilidad a la información:** permiten el acceso a la información a través de las diferentes estructuras de datos de forma transparente a los usuarios finales, para que estos solo se enfoquen en el análisis y no en el origen y procedencia de los datos.

- **Apoyo en la toma de decisiones:** permiten la exploración de los datos, a fin de seleccionar, filtrar y personalizar los mismos, para la obtención de información oportuna, relevante y útil, para apoyar el proceso de toma de decisiones.
- **Orientación los usuarios finales:** permiten a través de entornos amigables e intuitivos, que los usuarios puedan realizar análisis y consultas, sin poseer conocimientos técnicos. Si bien lo realmente importante son los datos mismos, que estos puedan ser interpretados y analizados por los usuarios dependerá en gran medida de cómo se presenten y dispongan.

Existen diferentes tipos de herramientas de consulta y análisis, y de acuerdo a la necesidad, tipos de usuarios y requerimientos de información, se deberán seleccionar las más propicias al caso. Entre ellas se destacan las siguientes:

- Reportes y Consultas.
- OLAP.
- Dashboards.
- Data Mining.
- EIS.

4.1. REPORTES Y CONSULTAS

Se han desarrollado muchas herramientas para la producción de consultas y reportes, que ofrecen a los usuarios, a través de pantallas gráficas intuitivas, la posibilidad de generar informes avanzados y detallados del tema de interés que se este analizando. Los usuarios solo deben seguir una serie de simples pasos, como por ejemplo seleccionar opciones de un menú, presionar tal o cual botón para especificar los elementos de datos, sus condiciones, criterios de agrupación y demás atributos que se consideren significativos.

Actualmente las herramientas de generación de reportes y consultas cuentan con muchas prestaciones, las cuales permiten dar variadas formas y formatos a la presentación de la información. Entre las opciones más comunes se encuentran las siguientes:

- Parametrización de los datos devueltos.
- Selección de formatos de salida (planilla de cálculo, HTML, PDF, etc.).
- Inclusión de gráficos de tortas, barras, etc. Utilización de plantillas de formatos de fondos. Inclusión de imágenes. Formatos tipográficos. Links a otros reportes.

4.2. OLAP

El procesamiento analítico en línea OLAP (On Line Analytic Processing), es el componente más poderoso del Data Warehousing, ya que es el motor de consultas especializado del depósito de datos.

Las herramientas OLAP, son una tecnología de software para análisis en línea, administración y ejecución de consultas, que permiten inferir información del comportamiento del negocio.

Su principal objetivo es el de brindar rápidas respuestas a complejas preguntas, para interpretar la situación del negocio y tomar decisiones. Cabe destacar que lo que es realmente interesante en OLAP, no es la ejecución de simples consultas tradicionales, sino la posibilidad de utilizar operadores tales como drill-up, drill-down, etc, para explotar profundamente la información.

Además, a través de este tipo de herramientas, se puede analizar el negocio desde diferentes escenarios históricos, y proyectar como se ha venido comportando y evolucionando en un ambiente multidimensional, o sea, mediante la combinación de diferentes perspectivas, temas de interés o dimensiones. Esto permite deducir tendencias, por medio del descubrimiento de relaciones entre las perspectivas que a simple vista no se podrían encontrar sencillamente.

Las herramientas OLAP requieren que los datos estén organizados dentro del depósito en forma multidimensional, por lo cual se utilizan cubos multidimensionales.

Además de las características ya descritas, se pueden enumerar las siguientes:

- Permite recolectar y organizar la información analítica necesaria para los usuarios y disponer de ella en diversos formatos, tales como tablas, gráficos, reportes, tableros de control, etc.
- Soporta análisis complejos de grandes volúmenes de datos.
- Complementa las actividades de otras herramientas que requieran procesamiento analítico en línea.
- Presenta a los usuarios una visión multidimensional de los datos (matricial) para cada tema de interés del negocio.
- Es transparente al tipo de tecnología que soporta el Data Warehouse, ya sea ROLAP, MOLAP u HOLAP.
- No tiene limitaciones con respecto al número máximo de dimensiones permitidas.
- Permite a los usuarios, analizar la información basándose en más criterios que un análisis de forma tradicional.
- Al contar con muestras grandes, se pueden explorar mejor los datos en busca de respuestas.
- Permiten realizar agregaciones y combinaciones de los datos de maneras complejas y específicas, con el fin de realizar análisis más estratégicos.

4.3. DASHBOARDS

Los Dashboards se pueden entender como una colección de reportes, consultas y análisis interactivos que hacen referencia a un tema en particular y que están relacionados entre sí.

Existen diversas maneras de diseñar un Dashboard, cada una de las cuales tiene sus objetivos particulares, pero a modo de síntesis se expondrán algunas características generales que suelen poseer:

- Presentan la información altamente resumida.
- Se componen de consultas, reportes, análisis interactivos, gráficos (de torta, barras, etc), semáforos, indicadores causa-efecto, etc.
- Permiten evaluar la situación de la empresa con un solo golpe de vista.
- Poseen un formato de diseño visual muy llamativo.

4.4. DATA MINING

Esta herramienta constituye una poderosa tecnología con un gran potencial que ayuda y brinda soporte a los usuarios, con el fin de permitirles analizar y extraer conocimientos ocultos y predecibles a partir de los datos almacenados en un Data Warehouse o en un OLTP. Claro que es deseable que la fuente de información sea un Data Warehouse, por todas las ventajas que aporta.

La integración con el depósito de datos facilita que las decisiones operacionales sean implementadas directamente y monitorizadas.

Implementar Data Mining permitirá analizar factores de influencia en determinados procesos, predecir o estimar variables o comportamientos futuros, segmentar o agrupar ítems similares, además de obtener secuencias de eventos que provocan comportamientos específicos.

Una de las principales ventajas del Data Mining es que, como recién se ha hecho mención, permite inferir comportamientos, modelos, relaciones y estimaciones de los datos, para poder desarrollar predicciones sobre los mismos, sin la necesidad de contar con patrones o reglas preestablecidas, permitiendo tomar decisiones proactivas y basadas en un conocimiento acabado de la información.

Además brinda la posibilidad de dar respuesta a preguntas complicadas sobre los temas de interés, como por ejemplo ¿Qué está pasando?, ¿Por qué? y ¿Qué pasaría sí?, estos cuestionamientos aplicados a una empresa podrían ser: ¿Cuál de los productos de tal marca y clase serán más vendidos en la zona norte en el próximo semestre? y ¿por qué? Además se podrán ver los resultados en forma de reportes tabulares, matriciales, gráficos, tableros, etc.

Entonces, se puede definir Data Mining como una técnica para descubrir patrones y relaciones entre abundantes cantidades de datos, que a simple vista o que mediante otros tipos de análisis no se pueden deducir, ya que tradicionalmente consumiría demasiado tiempo o estaría fuera de las expectativas.

Los sistemas Data Mining se desarrollan bajo lenguajes de última generación basados en Inteligencia Artificial y utilizan métodos matemáticos tales como:

- Redes Neuronales.
- Sistemas Expertos.
- Programación Genética.
- Árboles de Decisión.

4.5. EIS

EIS (Executive Information System) proporciona medios sencillos para consultar, analizar y acceder a la información de estado del negocio. Además, pone a disposición facilidades para que los usuarios puedan conseguir los datos buscados rápidamente, empleando el menor tiempo posible para comprender el uso de la herramienta.

Usualmente, EIS se utiliza para analizar los indicadores de performance y desempeño del negocio o área de interés, a través de la presentación de vistas con datos simplificados, altamente consolidados, mayormente estáticos y preferentemente gráficos.

El concepto principal de esta herramienta, se basa en el simple hecho de que los ejecutivos no poseen tiempo, ni las habilidades necesarias para analizar grandes cantidades de datos.

5. APLICACIONES DE DATA WAREHOUSE

5.1. MARKETING

La aplicación de tecnologías de Data Warehouse supone un nuevo enfoque de Marketing, haciendo uso del Marketing de Base de Datos. En efecto, un sistema de Marketing Warehouse implica un marketing científico, analítico y experto, basado en el conocimiento exhaustivo de clientes, productos, canales y mercado.

Este conocimiento se deriva de la disposición de toda la información necesaria, tanto interna como externa, en un entorno de Data Warehouse, persiguiendo con toda esta información, la optimización de las variables controladas del Marketing Mix y el soporte a la predicción de las variables no controlables (mediante técnicas de Data Mining). Basándose en el conocimiento exhaustivo de los clientes se consigue un tratamiento personalizado de los mismos tanto en el día a día (atención comercial) como en acciones de promoción específicas.

Las áreas en las que se puede aplicar las tecnologías de Data Warehouse a Marketing son, entre otras:

- Investigación Comercial
- Segmentación de mercados
- Identificación de necesidades no cubiertas y generación de nuevos productos, o modificación de productos existentes

- Fijación de precios y descuentos
- Definición de la estrategia de canales de comercialización y distribución
- Definición de la estrategia de promoción y atención al cliente
- Relación con el cliente:
- Programación, realización y seguimiento de acciones comerciales
- Lanzamiento de nuevos productos
- Campañas de venta cruzada, vinculación, fidelización, etc.
- Apoyo al canal de venta con información cualificada

5.2. ANÁLISIS DE RIESGO FINANCIERO

El Data Warehouse aplicado al análisis de riesgos financieros ofrece capacidades avanzadas de desarrollo de aplicaciones para dar soporte a las diversas actividades de gestión de riesgos. Es posible desarrollar cualquier herramienta utilizando las funciones que incorpora la plataforma, gracias a la potencionalidad estadística aplicada al riesgo de crédito.

Así se puede usar para llevar a cabo las siguientes funcionalidades:

- **Para la gestión de la posición:**

  Determinación de la posición, Cálculo de sensibilidades, Análisis what/if, Simulaciones, Monitorización riesgos contra límites, etc.

- **Para la medición del riesgo:**
  Soporte metodología RiskMetrics (Metodología registrada de J.P. Morgan / Reuters), Simulación de escenarios históricos, Modelos de covarianzas, Simulación de Montecarlo, Modelos de valoración, Calibración modelos valoración, Análisis de rentabilidad, Establecimiento y seguimiento. de límites, Desarrollo/modificación modelos, Stress testing, etc.

El uso del Data Warehouse ofrece una gran flexibilidad para creación o modificación de modelos propios de valoración y medición de riesgos, tanto motivados por cambios en la regulación, como en avances en la modelización de estos instrumentos financieros.

Ello por cuanto se puede almacenar y poner a disposición información histórica de mercado y el uso de técnicas de Data Mining nos simplifica la implantación de cualquier método estadístico. Los métodos de previsión, se pueden realizar usando series históricas, (GARCH, ARIMA, etc.)

5.3. DATA WAREHOUSE PARA LA PRESTACION DEL SERVICIOS PUBLICO DE INFORMACION ESTADISTICAS

Este proyecto consiste en aplicar las tecnologías de base de datos y Datawarehousing en el desarrollo de un almacén integrado de datos definitivos con información estadística obtenida de los programas de censos nacionales, encuestas y registros administrativos para la elaboración de productos, la toma de decisiones y la planeación facilitando que el personal del instituto pueda atender con mayor oportunidad los

requerimientos de información de los usuarios del INEGI en el marco del Sistema Nacional de Información Estadística y Geográfica.

6. DATA WAREHOUSE EN TIEMPO REAL

Con el software de almacenamiento de datos, una de las limitaciones más comunes es la ventana de tiempo disponible para el procesamiento de extracción por lotes sobre los sistemas de origen. Generalmente el proceso de extracción que consume muchos recursos debe realizarse fuera de las horas laborables y restringe el acceso a los sistemas de origen críticos.

Un software de integración de datos en tiempo real y de bajo impacto puede liberar a sus sistemas de esas ventanas de lotes. Cuando el componente de extracción utiliza un método no invasivo –como la lectura de los registros de transacciones de la base de datos para capturar solo los datos modificados– no generará una carga sobre los sistemas de origen. Por lo tanto, la extracción de datos puede realizarse en cualquier momento del día y durante todo el día, mientras los usuarios están en línea.

Cuando la extracción se produce en tiempo real, los datos pueden aportar un valor excepcional para el negocio, aunque no modifica la forma en que se ajustan los elementos en el proceso de recolección de datos para dar soporte a esa naturaleza de datos en tiempo real. Y aún más, los datos tienen que estar efectivamente protegidos, y es difícil aplicar técnicas de recuperación de desastres y de respaldo sobre datos que están en constante movimiento.

Pero la misma tecnología que puede permitir la integración de datos en tiempo real para los almacenes de datos, también puede ser utilizada para proteger aún más los datos. Después de todo, la tecnología que mueve datos en tiempo real también interactúa con los datos en tiempo real, creando un punto de entrada para las tecnologías de protección de datos. Sin embargo, la velocidad y la eficiencia de los datos en movimiento puede verse afectada por la latencia introducida durante el proceso de protección.

Eso significa que una de las primeras consideraciones a tener en cuenta cuando se desplazan a un régimen de recolección de datos activo que se integra con un almacén de datos, debe ser el flujo de los datos a través de sistemas de TI y la latencia que se puede introducir. En otras palabras, la integración de datos en tiempo real requiere de una comprensión de los datos en movimiento y de los componentes que mejoran o impiden ese movimiento.

Obviamente, las empresas quieren proteger sus datos. Sin embargo, a medida que crece la demanda por el volumen de datos, la tecnología de almacenamiento se convierte en un activo crítico sobre el que se apoya la continuidad del negocio. Y a medida que los análisis de datos en tiempo real se convierten en parte de un proceso de línea de negocio, también cae dentro del ámbito de la continuidad. El enfoque más básico para proporcionar seguridad y continuidad de los datos, es la replicación de hardware o software que mantiene automáticamente una copia secundaria de los datos críticos. No

son desconocidos los métodos de respaldo en las instalaciones y que se basan en software de código abierto.

Las empresas están invirtiendo en cinco áreas críticas relacionadas con la gestión de los datos: recuperación de desastres, alta disponibilidad, copia de seguridad, rendimiento de procesamiento de datos y migración hacia bases de datos más avanzadas. Esto prepara el escenario para que las TI desarrollen tecnologías avanzadas, como la integración de datos en tiempo real y sus elementos de infraestructura asociados. Además, esas inversiones estratégicas pueden brindar los recursos presupuestarios para acelerar la adopción de tecnologías en tiempo real, al tiempo que mejoran el rendimiento de la inversión y justifican el modelo de negocio propuesto para un proyecto de integración de datos en tiempo real.

No obstante, es fundamental asignar esas áreas de inversión en elementos en especie de un sistema de integración de datos en tiempo real, y eso trae aparejada una comprensión profunda de los componentes que conforman ese sistema y cómo tales componentes son impulsados por las necesidades de los datos de la organización. Entre ellos se incluyen los siguientes:

- Volumen de datos (tamaño de los datos y cantidad de actualizaciones)
- Frecuencia del movimiento de datos
- Requisitos de transformación
- Lapsos de interrupción y continuidad del negocio

Son esos elementos los que impulsarán qué productos se elegirán para construir una infraestructura completa para la integración de datos en tiempo real. Pero la expresión en tiempo real nos lleva a un significado un tanto diferente al incorporar las tecnologías de adquisición de datos. Algunas tecnologías se centran en el concepto de “momento adecuado” para la inteligencia empresarial (BI). La expresión se refiere a las diversas necesidades de los usuarios finales para acceder a la inteligencia, y eso significa que tales necesidades cambian en diferentes casos de uso.

Sin embargo, para un almacenamiento de datos en funcionamiento, la tecnología no debe basarse en un paradigma del momento adecuado. La tecnología debería ofrecer verdaderas capacidades de tiempo real y luego dejar que el usuario de negocios elija el momento adecuado para acceder a los datos. Sin embargo, algunas empresas pueden hallar valor en la ideología del momento adecuado de BI, lo cual plantea la pregunta: ¿Cuándo una organización debería utilizar integración de datos en tiempo real?

En el mundo real, las empresas utilizan arquitecturas mixtas de TI de múltiples proveedores (a menudo un legado de la historia de la empresa). Cuando elija una tecnología de integración de datos en tiempo real, busque una que fácilmente pueda reunir información de una variedad de bases de datos y plataformas de aplicaciones. Esta es la clave más importante para el éxito.

La plataforma de integración es la base para los datos en tiempo real, y la compatibilidad de productos cruzados es uno de sus principales habitantes. Pero encontrar una plataforma que combine estos elementos y que admita el procesamiento en tiempo real sin traer dificultades será todo un desafío.

CONCLUSIONES

Al llegar al final del proyecto, podemos decir que el desarrollo de un Sistema de DW es complejo. Abarca gran cantidad de personas, usuarios finales del DW, usuarios de los sistemas fuente, desarrolladores, etc. También comprende diferentes componentes, sistemas fuente, herramientas de extracción, herramientas de consulta, redes, aplicaciones de usuario final, bases de datos, etc. Un Sistema de DW debe combinar todos éstos elementos y brindar un producto final consistente, amigable y confiable que a su vez esté preparado para enfrentar los continuos cambios que surjan debido a su tan variada estructura. Un trabajo no trivial.

El desarrollo de un sistema de DW es diferente al desarrollo de un sistema operacional. A grandes rasgos, uno apoya al negocio transaccional y el otro al decisional. El área de DW es un terreno nuevo en el cual queda mucho por experimentar. No existe aún, una metodología universalmente reconocida para construir un sistema de esta índole. Es por esto que construímos una estrategia de trabajo que sirva de guía durante el desarrollo. La misma, fue elaborada por nosotros a lo largo del proyecto y actualizada a medida que la íbamos experimentando. Consideramos la estrategia un aporte interesante al proyecto que puede ser de utilidad a futuros desarrollos.

No disponer de herramientas de apoyo específicas para la construcción de un sistema de DW como ser, herramientas de extracción transformación e integración de datos, detectores de reglas en los datos, reconocedores de patterns, etc., agregó complejidad al desarrollo, afectó los tiempos del mismo y privó de experimentar nuevas tecnologías existentes en la actualidad. De todas formas se desarrolló un sistema lo más completo y confiable posible.

REFERENCIAS Web

www.intellicomp.cl /datawh.htm

www.gcc.com.mx/soluciones/dwhouse.htm

www.bftsystems.com/sol_dataware.htm

www.wnet.es/Productos/infsagent1.htm

www.compag.es/soluciones/dwareymart/Secciones/dwsodwxx.htm

www.mtginc.com/spanish/whatwedo/technical/dws.html1

www.sun.es/success/warehouse/bankinter

www.194.174.14.5/Espanol/es_data_warehouse.htm

www.fciencias.unam.mx/revista/soluciones/30s/No34/dataware.htm1

www.business.carleton.ca/~aramirez/Espol/Modulo1/index.html1

www.datawarehouse.com

www.digital.com/alphaserver/solutions/dataware/dataware.htm1