

UNIVERSIDAD NACIONAL DE LA AMAZONÍA PERUANA



FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA



EXAMEN DE SUFICIENCIA PROFESIONAL

“EL MODELO DATA WAREHOUSE-OLAP (ONLINE ANALYTICAL PROCESSING)”

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS E INFORMÁTICA

Presentado por el Bachiller:

Paolo Héctor Sinti Cabrera

IQUITOS – PERÚ

2015



UNIVERSIDAD NACIONAL DE LA AMAZONIA PERUANA
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA

ACTA DE EXAMEN ORAL DE SUFICIENCIA PROFESIONAL

Siendo las 19:50 horas del día 28 de AGOSTO del 2015, en la Instalación del Auditorio de esta Facultad, se ha constituido el jurado examinador integrado por los siguientes miembros:

Presidente : Ing. Saúl Flores Nunta
Primer Miembro : Ing. Carlos Alberto García Cortegano
Segundo Miembro : Ing. Francisco Miguel Ruiz Hidalgo

Se procedió, al Acto Académico del Examen Oral de Suficiencia Profesional del Bachiller: **Paolo Héctor Sinti Cabrera**, quien sustentó el tema **"El Modelo Data Warehouse-OLAP (ONLINE ANALYTICAL PROCESSING)**, para optar el Título Profesional de Ingeniero de Sistema e Informática, de acuerdo a lo establecido en el Reglamento de Grados y Títulos y sustentado en la Ley N° 30220.

Posteriormente, al Acto de sustentación del informe final del bachiller se procedió al cálculo de Calificación y Condición Final, obteniéndose el siguiente resultado:

	Calificaciones	
	En número	En letras
Promedio de la Calificación Final de las Asignaturas.	<u>15.25</u>	<u>QUINCE 25/100</u>
Calificación de la Sustentación del Informe Final.	<u>12.80</u>	<u>DOCE 80/100</u>
Calificación Final	<u>14.03</u>	<u>CATORCE 3/100</u>

Se desprende que la Condición Final del Bachiller es (marcar el que corresponde):

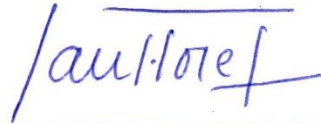
- Aprobado con excelencia (18 a 20 puntos).
 Aprobado por unanimidad (15 a 17.9 puntos).
 Aprobado por mayoría (12 a 14.9 puntos).
 Desaprobado (Menos de 12 puntos).

Siendo las 20:00 horas del mismo día, se da por concluido el acto académico, firmando en conformidad los miembros del Jurado Examinador.

Ing. Saúl Flores Nunta
Presidente

Ing. Carlos Alberto García Cortegano
Primer Miembro
Ing. Francisco Miguel Ruiz Hidalgo
Segundo Miembro

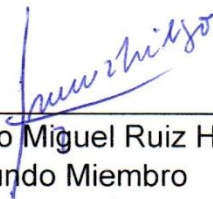
INFORME TÉCNICO DEL EXAMEN DE SUFICIENCIA PREVIA ACTUALIZACION ACADÉMICA APROBADO EN SUSTENTACIÓN PÚBLICA EL DIA **28 DE AGOSTO DEL 2015**, POR EL JURADO EXAMINADOR, DESIGNADO POR EL DECANO DE LA FACULTAD DE INGENIERIA DE SISTEMAS E INFORMÁTICA DE LA UNIVERSIDAD NACIONAL DE LA AMAZONIA PERUANA.



Ing. Saúl Flores Nunta
Presidente



Ing. Carlos Alberto García Cortegano
Primer Miembro



Ing. Francisco Miguel Ruiz Hidalgo
Segundo Miembro

RESUMEN.

En el presente trabajo, se sistematizan los conceptos inherentes al Modelo Data Warehouse, 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 los conceptos generales relacionados al Data Warehouse, Seguidamente, se introducirá a la definición de requerimientos y los procesos de negocio para modelar un Data Warehouse, y se expondrán sus aspectos más relevantes y significativos. Luego, se precisarán y detallarán todos los componentes que intervienen en la Integración de Datos, de manera organizada e intuitiva, atendiendo su interrelación. Posterior se describe el Diseño Dimensional para los procesos de Negocio.

Finalmente, se describirán algunos conceptos que deben tenerse en cuenta para la Minería de datos.

El principal objetivo de este trabajo práctico, es ayudar a comprender el complejo ambiente de Data Warehouse, 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 Warehouse, 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.

INDICE.

I.	JUSTIFICACION.....	8
II.	OBJETIVOS.....	9
III.	DESARROLLO DEL TEMA.....	10
1.	DATA WAREHOUSE.....	10
1.1.	INTRODUCCION.....	10
1.2.	HISTORIA.....	10
1.3.	DEFINICIÓN.....	11
1.4.	CARACTERISTICAS.....	12
1.4.1.	ORIENTADA AL NEGOCIO.....	12
1.4.2.	INTEGRADA.....	13
1.4.3.	VARIANTE EN EL TIEMPO.....	14
1.4.4.	NO VOLATIL.....	14
1.5.	CUALIDADES.....	15
1.6.	VENTAJAS.....	16
1.7.	DESVENTAJAS.....	16
1.8.	REDUNDANCIA.....	17
1.9.	ESTRUCTURA.....	17
1.10.	INTELIGENCIA DE NEGOCIOS.....	18
2.	EL MODELO DATA WARE HOUSE.....	19
2.1.	DATAWAREHOUSING.....	19
2.1.1.	SISTEMAS DATAWAREHOUSING.....	19
2.1.2.	DIFERENCIA ENTRE DATA WAREHOUSING Y DATAWAREHOUSE.....	21
2.2.	TECNOLOGIA OLAP.....	21
2.2.1.	BENEFICIOS.....	22
2.2.2.	MODELO DE DATOS.....	23
2.2.3.	APLICACIONES.....	23
3.	INTEGRACIÓN DE DATOS.....	24
3.1.	LOS PROCESOS EXTRACT, TRANSFORM AND LOAD (LOAD).....	24
3.1.1.	EXTRACCION.....	25
3.1.2.	TRANSFORMACION.....	25
3.1.2.1.	CODIFICACION.....	25
3.1.2.2.	MEDIDA DE ATRIBUTOS.....	25
3.1.2.3.	CONVENCIONES DE NOMBRAMIENTO.....	26
3.1.3.	CARGA.....	26
3.2.	LA LIMPIEZA DE DATOS.....	26

4.	DISEÑO DIMENSIONAL DEL PROCESO DEL NEGOCIO	27
4.1.	CONCEPTOS DE MODELADO DIMENSIONAL.....	27
4.2.	TABLAS DE DIMENSIONES.....	28
4.2.1.	DIMENSIÓN TIEMPO	29
4.2.2.	JERARQUIAS	29
4.2.3.	RELACION.....	30
4.2.4.	GRANULARIDAD.....	31
4.3.	TABLAS DE HECHOS	31
4.4.	CONCEPTOS ADICIONALES.....	33
4.4.1.	ESQUEMA EN ESTRELLA	33
4.4.2.	ESQUEMA COPO DE NIEVE	35
4.4.3.	ESQUEMA CONSTELACIÓN	36
5.	LA MINERIA DE DATOS.....	37
5.1.	PRINCIPALES MODELOS DE ANALISIS DE DATOS.....	38
5.1.1.	ANALISIS FACTORIAL	38
5.1.1.1.	TIPOS DE ANALISIS FACTORIAL	38
5.1.1.2.	APLICACIONES	39
5.1.2.	ANALISIS PREDICTIVO	39
5.1.2.1.	DEFINICION	39
5.1.2.2.	TIPOS	40
5.1.2.3.	APLICACIONES	41
5.1.3.	ANALISIS EXPLORATORIO DE DATOS.....	41
5.2.	EL MANEJO DE DATOS NO ESTRUCTURADOS	42
5.2.1.	DEFINICION DE DATOS NO ESTRUCTURADOS.....	42
5.2.2.	CARACTERISTICAS DE DATOS NO ESTRUCTURADOS.....	42
5.2.3.	TRATAMIENTO DE DATOS NO ESTRUCTURADOS.....	43
6.	CASO DE ÉXITO.	44
IV.	CONCLUSIONES.....	48
V.	REFERENCIAS BIBLIOGRAFICAS.....	49

INDICE DE TABLAS Y FIGURAS.

Fig. 1: Data WareHouse, Características.....	12
Fig. 2: Data WareHouse, variante en el tiempo.....	14
Fig. 3: Data WareHouse, No Volátil.....	15
Fig. 4: Data WareHouse, Estructura.....	17
Fig. 5: Modelo Data Warehousing.....	20
Tabla 1: Aplicaciones OLAP.....	23
Fig. 6: Esquema de Solución BI.....	24
Fig. 7: Tabla de Dimensiones.....	28
Fig. 8: Jerarquía de Geografía.....	30
Fig. 9: Tabla de Hechos.....	32
Fig. 10: Esquema en Estrella.....	33
Fig. 11: Esquema en Estrella.....	34
Fig. 12: Desnormalización.....	35
Fig. 13: Esquema Copo de Nieve.....	36
Fig. 14: Esquema Constelación.....	37

I. JUSTIFICACION.

Actualmente, en las actividades diarias de cualquier organización, se generan datos como producto secundario, que son el resultado de todas las transacciones operacionales que se realizan. Es muy común, que los mismos se almacenen y administren a través de sistemas transaccionales en bases de datos relacionales.

Pero, la idea central de este trabajo práctico, es que estos dejen de solo ser simples datos, para convertirse en información que enriquezca las decisiones de los ejecutivos.

Las organizaciones desean explotar y maximizar el valor de su información para lograr tener una mayor ventaja competitiva. Además, es un factor muy importante para las mismas, con el fin de incrementar sus ganancias, enfocarse en retener a sus clientes actuales, como así también conseguir nuevos.

Debido a lo expuesto anteriormente, sería ideal que las organizaciones tuviesen la posibilidad de segmentar y/o clasificar a sus clientes de alguna manera, en este caso, por su rentabilidad, para poder actuar en base a ello, confeccionando una estrategia que permita cumplir con este objetivo.

Los Data Warehouse implementados a través de la Inteligencia de Negocios permite realizar este tipo de segmentación, además, está orientada a encontrar información que no solo se encargue de responder a preguntas de lo que está sucediendo o ya sucedió, sino también, posibilita la construcción de modelos, mediante los cuales se podrán predecir eventos futuros.

II. OBJETIVOS

General

Conocer las definiciones y conceptos inherentes al Data Warehouse – OLAP, así como la importancia, ventajas y procesos involucrados para llevar a cabo un proyecto de inteligencia de negocios exitoso.

Específicos

- Identificar los conceptos y fundamentos que presenta el Data Warehouse-OLAP.
- Reconocer la importancia del modelado de un Data Warehouse y el impacto en los procesos del negocio.
- Identificar los procesos de Integración de Datos presentes en un Data Warehousing.
- Comprender los conceptos relacionado al proceso de Modelado Dimensional dentro de los procesos de Negocio.

III. DESARROLLO DEL TEMA

EL MODELO DATA WARE HOUSE-OLAP (ONLINE ANALYTICAL PROCESSING)

1. DATA WAREHOUSE

1.1. INTRODUCCION

Un Datawarehouse es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta. La creación de un datawarehouse representa en la mayoría de las ocasiones el primer paso, desde el punto de vista técnico, para implantar una solución completa y fiable de Business Intelligence.

Debido a que para llevar a cabo BI, es necesario gestionar datos guardados en diversos formatos, fuentes y tipos, para luego depurarlos e integrarlos, además de almacenarlos en un solo destino, depósito o base de datos que permita su posterior análisis y exploración, es imperativo y de vital importancia contar con una herramienta que satisfaga todas estas necesidades.

Esta herramienta es el Data Warehouse (DW), que básicamente se encarga de consolidar, integrar y centralizar los datos que la empresa genera en todos los ámbitos de una actividad de negocios (Compras, Ventas, Producción, etc), para luego ser almacenados mediante una estructura que permite el acceso y exploración de la información requerida con buena performance, facilitando posteriormente, una amplia gama de posibilidad de análisis multivariados, que permitirá la toma de decisiones estratégicas y tácticas.

La ventaja principal de este tipo de bases de datos radica en las estructuras en las que se almacena la información (modelos de tablas en estrella, en copo de nieve, cubos relacionales, etc.). Este tipo de persistencia de la información es homogénea y fiable, y permite la consulta y el tratamiento jerarquizado de la misma (siempre en un entorno diferente a los sistemas operacionales).

1.2. HISTORIA

El data warehouse es una evolución de los sistemas de bases de datos relacionales, es un proceso, no un producto. En 1988 los investigadores de IBM Barry Devlin y Paul Murphy inventaron el término warehouse de información y en 1991, W.H. "Bill" Inmon hizo las data warehouses prácticas cuando publicó una guía de cómo construir una data. Bill, cuyo verdadero nombre es William Harvey Inmon es llamado el "padre del Data Warehouse", por ser el responsable de la construcción, uso y mantenimiento del almacén de datos.

Ha escrito más de 46 libros y tiene más de 35 años de experiencia en tecnologías de manejo de bases de datos y diseño de data warehouse. En 1999 creó la Government Information Factory (Fábrica de Información Corporativa) sitio web para educar los profesionales y tomadores de decisiones sobre el almacenamiento de datos. Además fue el creador del Data Warehouse 2.0.

Como sus logros más recientes, desarrolló la tecnología para la inclusión de datos de texto no estructurados en el almacén de datos o data warehouse - el primero en el mundo llamado "ETL textual".

1.3. DEFINICIÓN

El DW posibilita la extracción de datos de sistemas operacionales y fuentes externas, permite la integración y homogenización de los datos de toda la empresa, provee información que ha sido transformada y sumariada, para que ayude en el proceso de toma de decisiones estratégicas y tácticas.

El DW, convertirá entonces los datos operacionales de la empresa en una herramienta competitiva, debido a que pondrá a disposición de los usuarios indicados la información pertinente, correcta e integrada, en el momento que se necesita.

Una de las definiciones más famosas sobre DW, es la de W. H. Inmon, quien define: "Un Data Warehouse es una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia".

Debido a que W. H. Inmon, es reconocido mundialmente como el padre del DW, la explicación de las características más sobresalientes de esta herramienta se basó en su definición.

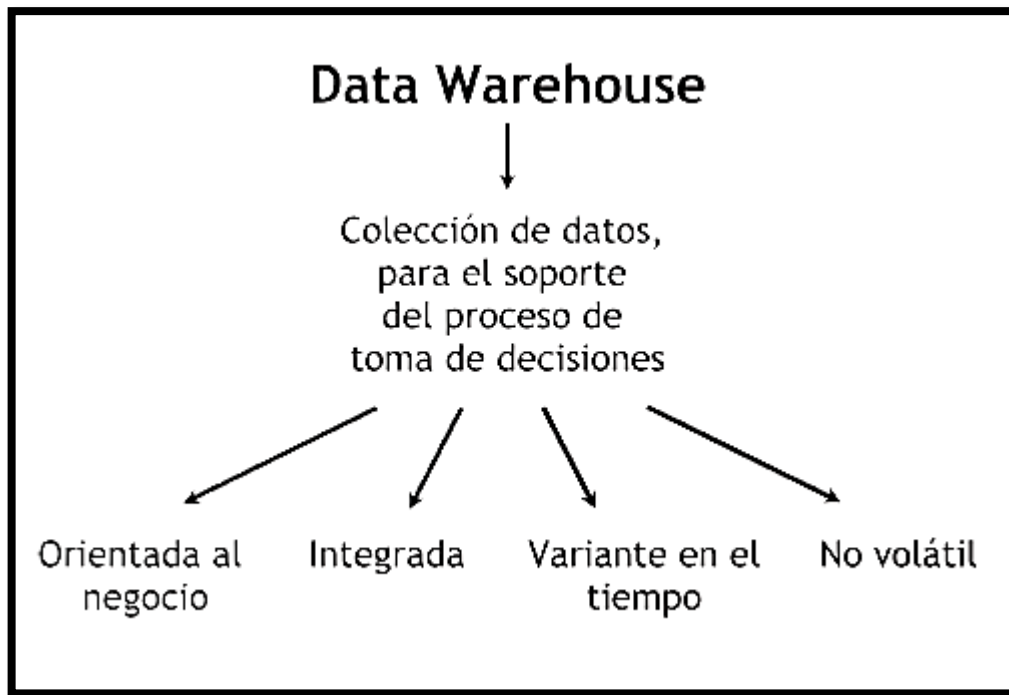


Fig.1: Data WareHouse, Características, Autor: Ing. Bernabéu, Ricardo Dario.

1.4. CARACTERISTICAS

1.4.1. 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.

A continuación, y con el fin de obtener una mejor comprensión de las diferencias existentes entre estos dos tipos de orientación, se realizará un análisis comparativo: Con respecto al nivel de detalle de los datos, el DW excluye la información que no será utilizada exclusivamente en el proceso de toma de decisiones; mientras que en los procesos orientados a las aplicaciones, se incluyen todos aquellos datos que son necesarios para satisfacer de manera inmediata los requerimientos funcionales de la actividad que soporten. Por ejemplo los datos comunes referidos al cliente, como su dirección de correo electrónico, fax, teléfono, D.N.I., código postal, etc, que son tan importantes de almacenar en cualquier sistema operacional, no son tenidos en cuenta en el depósito de datos por carecer de valor para la toma de decisiones, pero sí lo serán aquellos que indiquen el tipo de cliente, su clasificación, ubicación geográfica, sexo, edad, etc.

En lo que concierne a la interacción de la información, los datos operacionales mantienen una relación continua entre dos o más tablas, basadas en alguna regla comercial vigente; en cambio las relaciones encontradas en los datos residentes del DW son muchas, debido a que por lo general cada tabla del mismo estará conformada por la integración de varias tablas u otras fuentes del ambiente operacional, cada una con sus propias reglas de negocio inherentes.

El origen de este contraste es totalmente lógico, ya que el ambiente operacional se diseña alrededor de las aplicaciones u programas que necesite la organización para llevar a cabo sus actividades diarias y funciones específicas. Por ejemplo, una aplicación de una institución financiera manejará: préstamos, ahorros, tarjetas bancarias, cuentas, depósitos, etc. De esta manera, la base de datos combinará estos elementos en una estructura que se adapte a sus necesidades.

1.4.2. 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 Datos (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, aun 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.

1.4.3. 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:

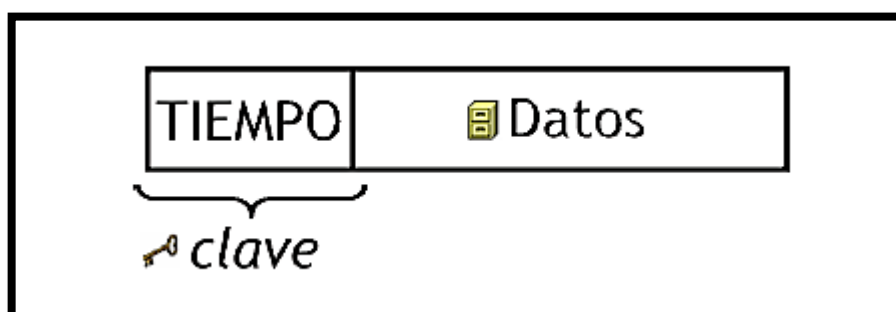


Fig. 2: Data WareHouse, variante en el tiempo, Autor: Ing. Bernabéu, Ricardo Dario.

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.

1.4.4. NO VOLATIL

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 depósito 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.

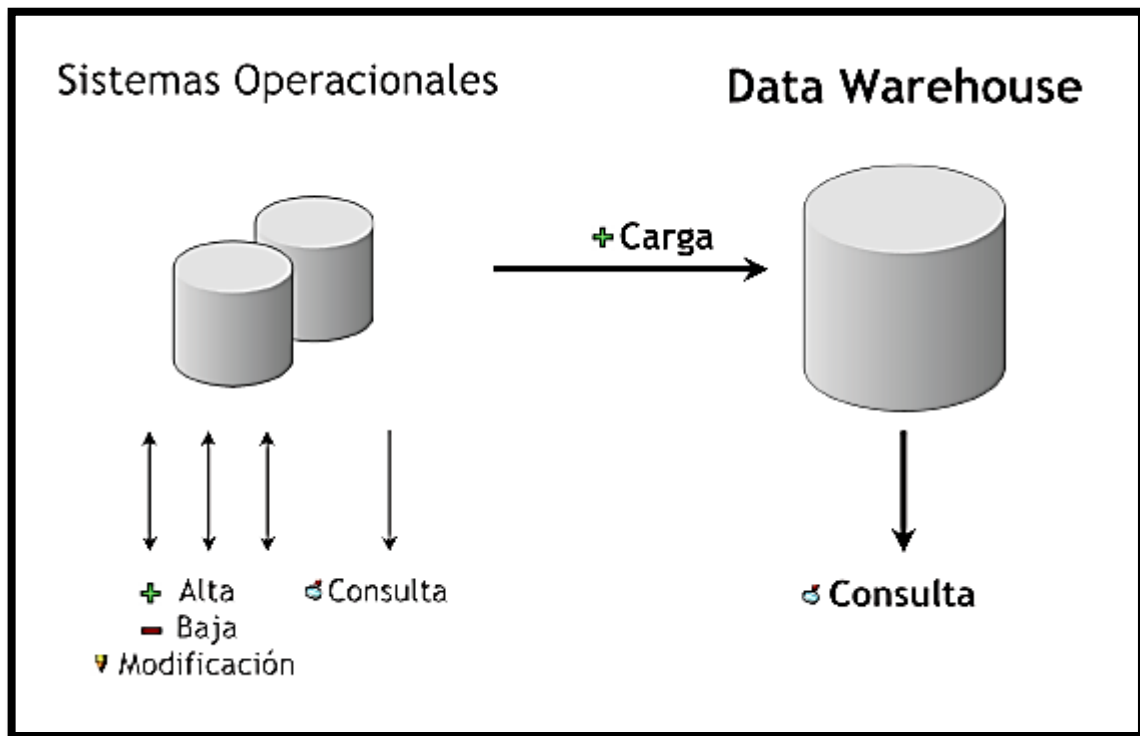


Fig.3: Data WareHouse, No Volátil, Autor: Ing Bernabéu, Ricardo Dario.

1.5. 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.

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.

1.6. 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.
- Aumento de la competitividad de los encargados de tomar decisiones.
- Permite la toma de decisiones estratégicas y tácticas.

1.7. 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.

1.8. REDUNDANCIA

Debido a que el DW recibe información histórica de diferentes fuentes, sencillamente se podría suponer que existe una repetición de datos masiva entre el ambiente DW y el operacional. Por supuesto, este razonamiento es superficial y erróneo, de hecho, hay una mínima redundancia de datos entre ambos ambientes.

Para entender claramente lo antes expuesto, se debe considerar lo siguiente:

- Los datos del ambiente operacional se filtran antes de pertenecer al DW. Existen muchos datos que nunca ingresarán, ya que no conforman información necesaria o suficientemente relevante para la toma de decisiones.
- El horizonte de tiempo es muy diferente entre los dos ambientes.
- El almacén de datos contiene un resumen de la información que no se encuentra en el ambiente operacional.

1.9. ESTRUCTURA

En la siguiente figura se puede apreciar mejor su respectiva estructura.

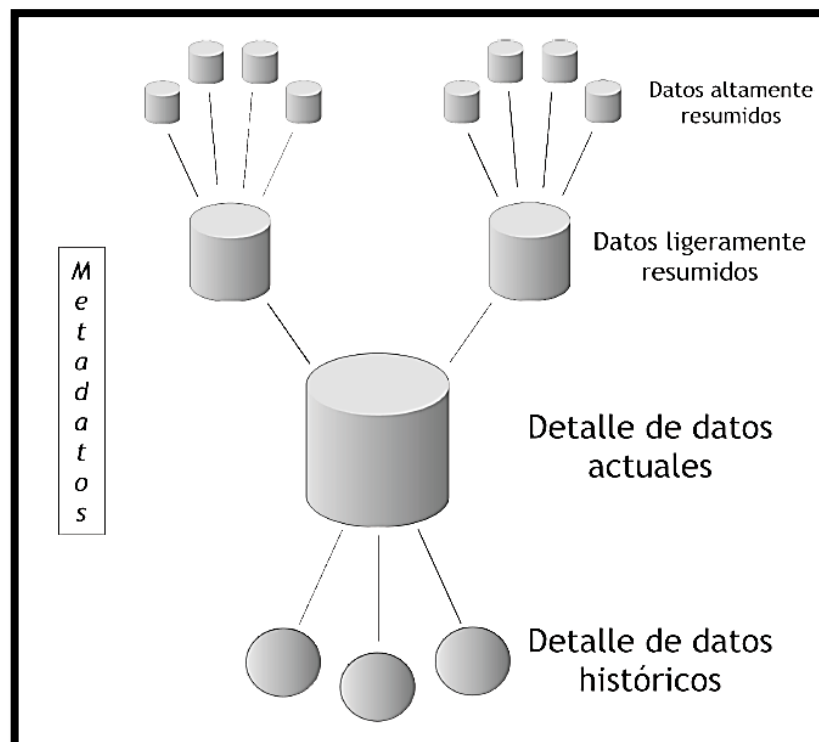


Fig. 4: Data WareHouse, Estructura, Autor: Ing. Bernabéu, Ricardo Dario.

Como se puede observar, los almacenes de datos están compuestos por diversos tipos de datos, que se organizan y dividen de acuerdo al nivel de detalle que posean.

A continuación se explicarán cada uno de estos tipos de datos:

Detalle de datos actuales: son aquellos que reflejan las ocurrencias más recientes.

Generalmente se almacenan en disco, aunque su administración sea costosa y compleja, con el fin de conseguir que el acceso a la información sea sencillo y veloz, ya que son bastante voluminosos. Su gran tamaño se debe a que los datos residentes poseen el más bajo nivel de granularidad, o sea, se almacenan a nivel de detalle.

Por ejemplo, aquí es donde se guardaría el detalle de una venta realizada en tal fecha.

Detalle de datos históricos: representan aquellos datos antiguos, que no son frecuentemente consultados. También se almacenan a nivel de detalle, normalmente sobre alguna forma de almacenamiento externa, ya que son muy pesados y en adición a esto, no son requeridos con mucha periodicidad. Este tipo de datos son consistentes con los de Detalle de datos actuales. Por ejemplo, en este nivel, al igual que en el anterior, se encontraría el detalle de una venta realizada en tal fecha, pero con la particularidad de que el día en que se registró la venta debe ser lo suficientemente antigua, para que se considere como histórica.

Datos ligeramente resumidos: son los que provienen desde un bajo nivel de detalle y sumarizan o agrupan los datos bajo algún criterio o condición de análisis. Habitualmente son almacenados en disco. Por ejemplo, en este caso se almacenaría la sumarización del detalle de las ventas realizadas en cada mes.

Datos altamente resumidos: son aquellos que compactan aún más a los datos ligeramente resumidos. Se guardan en disco y son muy fáciles de acceder. Por ejemplo, aquí se encontraría la sumarización de las ventas realizadas en cada año.

Metadatos: representan la información acerca de los datos. De muchas maneras se sitúa en una dimensión diferente al de otros datos del DW, ya que su contenido no es tomado directamente desde el ambiente operacional.

1.10. INTELIGENCIA DE NEGOCIOS

Es una arquitectura y colección de herramientas que buscan mejorar a las organizaciones, proporcionando vistas de aspectos de negocio a todos los empleados (estratégico, táctico, operacional) para que tomen mejores y más relevantes decisiones en menos tiempo y con la mayor información posible.

La inteligencia de negocios es la parte de la gestión empresarial encargada de la recogida, procesamiento y presentación de información relevante que facilite la toma de decisiones.

La inteligencia de negocios es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios.

La Inteligencia de Negocios se direcciona principalmente en Aplicaciones y Base de Datos de Soporte a la Toma de Decisiones, Personas correctas.

Desde un punto de vista más pragmático, y asociándolo directamente con las tecnologías de la información, podemos definir Business Intelligence como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la compañía) en información estructurada, para su explotación directa (reporting, análisis OLTP / OLAP, alertas...) o para su análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio.

2. EI MODELO DATA WARE HOUSE

2.1. DATAWAREHOUSING

2.1.1. SISTEMAS DATAWAREHOUSING

Los sistemas Data Warehouse o sistemas Data Warehousing surgen como un mecanismo de apoyo para la ayuda de toma de decisiones, en el que los datos de una organización se transforman en información estratégica, a la que además se puede acceder de manera sencilla y en el momento que se necesita. Con esta tecnología, los datos operacionales son una herramienta competitiva para las organizaciones. Se permite a los usuarios finales examinar los datos, realizar análisis y detectar tendencias, llevar a cabo el seguimiento de medidas críticas, producir informes con rapidez y detectar tendencias. De esta forma obtenemos una mayor ventaja competitiva en la organización, pudiéndonos anticipar a diversas situaciones.

Los sistemas que contienen datos operacionales (son los datos que se generan en las transacciones diarias de la organización) contienen información que es útil para los analistas de negocio. Por ejemplo, los analistas pueden usar esta información para ver qué productos se vendieron más en cierta población durante una época del año. Pero surgen varios problemas cuando los analistas de negocio intentan acceder directamente a estos datos:

Puede que los analistas no tengan el conocimiento suficiente para obtener los datos. Los datos operacionales pueden no estar en el mejor formato para ser usados con propósito de análisis.

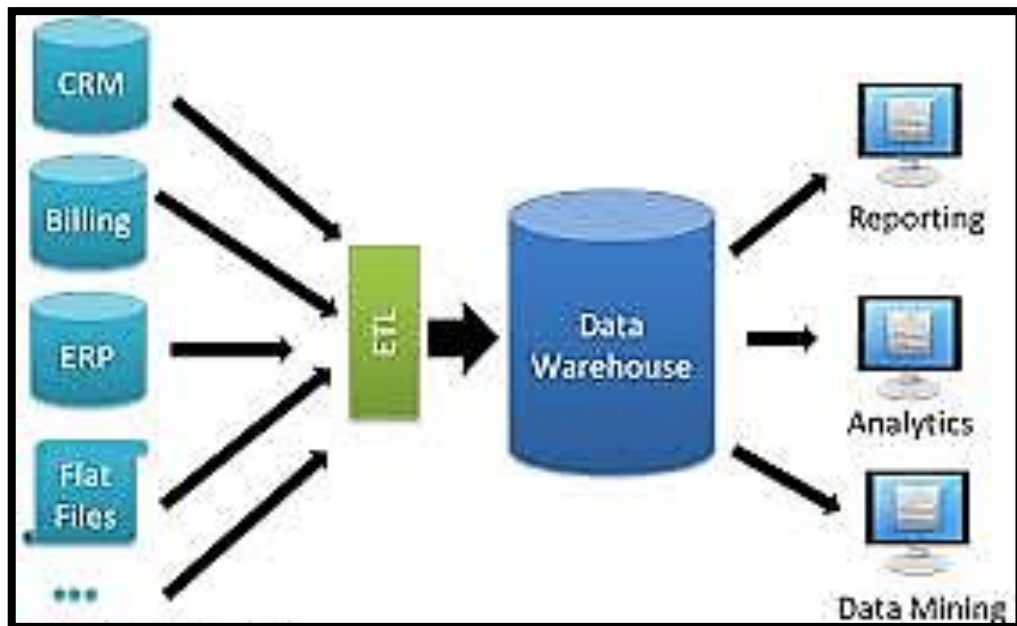


Fig. 5: Modelo Data Warehousing, Autor: Creación Propia.

La ausencia de una visión histórica hace difícil el análisis de los datos.

Un proceso de Data Warehousing soluciona estos problemas creando almacenes de datos informacionales. Los datos informacionales son datos que han sido extraídos de los datos operacionales y transformados para la toma de decisiones. Por ejemplo, limpiar los datos, realizar cálculos sobre éstos, separarlos de los datos operacionales...

Data Warehousing es el proceso de extraer y filtrar los datos de las operaciones comunes a la organización, procedentes de los distintos sistemas de información y/o sistemas externos, para transformarlos, integrarlos y almacenarlos en un depósito o almacén de datos (Data Warehouse) con el fin de acceder a ellos para dar soporte en el proceso de toma de decisiones de una organización.

El objetivo es convertir los datos operacionales en información relacionada y estructurada, homogénea, de mayor calidad y que se mantenga en el tiempo, es decir, los datos más recientes no sustituyen a los precedentes, pero tampoco se acumulan de cualquier manera, sino que se suelen mantener con un mayor nivel de detalle los datos actuales y de manera más agregada los datos anteriores.

Un punto fuerte del Data Warehousing es la meta-información. Cada dato está identificado por una descripción, un origen, historial o forma inicial y sucesiva. Este conjunto de datos sobre los datos es lo que se denomina como metadatos. Un metadato proporciona el contenido semántico necesario para que los datos puedan ser interpretados.

El Data Warehousing posibilita la extracción de datos de sistemas operacionales y fuentes externas, permite la integración y homogeneización de los datos de toda la

empresa, provee información que ha sido transformada y resumida, para que ayude en el proceso de toma de decisiones estratégicas y tácticas.

El Data Warehousing, convertirá entonces los datos operacionales de la empresa en una herramienta competitiva, debido a que pondrá a disposición de los usuarios indicados la información pertinente, correcta e integrada, en el momento que se necesita.

Pero para que el Data Warehousing pueda cumplir con sus objetivos, es necesario que la información que se extrae, transforma y consolida, sea almacenada de manera centralizada en una base de datos con estructura multidimensional denominada Data Warehouse (DW).

Una de las definiciones más famosas sobre DW, es la de William Harvey Inmon, quien define: "Un Data Warehouse es una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia".

Debido a que W. H. Inmon, es reconocido mundialmente como el padre del DW, la explicación de las características más sobresalientes de este concepto se basó en su definición.

2.1.2. DIFERENCIA ENTRE DATA WAREHOUSING Y DATA WAREHOUSE

Cuando queremos hacer referencia al proceso global en el que a partir de diferentes fuentes de datos (SGDB, ficheros planos, .csv, etc.) se crea y se mantiene un almacén central de datos y que puede ser consultado por herramientas con un propósito de análisis concreto y de ayuda a la toma de decisiones, se debe utilizar el término de Data Warehousing.

Para referirnos no al proceso en sí, sino al repositorio central de datos sobre el que se construye el sistema y que integra todos los datos de la organización desde el punto de vista del usuario y no de los procesos, nos estamos refiriendo a Data Warehouse.

2.2. TECNOLOGIA OLAP

El procesamiento analítico en línea OLAP (On Line Analytic Processing), es la componente más poderosa de los DW, ya que es el motor de consultas especializado de la bodega. 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 es que utilizan los 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, etc.
- Soporta análisis complejos de grandes volúmenes de datos.
- Complementa las actividades de otras herramientas que requieran procesamiento analítico en línea.
- Presenta al usuario una visión multidimensional de los datos (matricial) para cada tema de interés del negocio.
- Es transparente al tipo de tecnología que soporta el DW, ya sea ROLAP, MOLAP o HOLAP.
- Permite definir de forma flexible las dimensiones que se quieren analizar, sus restricciones, jerarquías y combinaciones.
- No tiene limitaciones con respecto al número máximo de dimensiones permitidas.

2.2.1. BENEFICIOS.

"Procesamiento analítico en línea", o OLAP proporciona un método para que las organizaciones acceder, ver y analizar datos corporativos con un alto rendimiento y flexibilidad. En las empresas el mundo globalizado de hoy en día se enfrenta a una mayor competencia y la expansión de sus operaciones a nuevos mercados. Por lo tanto, la velocidad a la que los ejecutivos de obtener información y tomar decisiones determina la competitividad de una empresa y su éxito a largo plazo. OLAP presenta información a los usuarios a través de un modelo de datos natural e intuitiva. A través de un estilo de navegación sencilla y la investigación, los usuarios finales pueden analizar rápidamente numerosos escenarios, generar informes "ad-hoc" y descubrir

tendencias y hechos relevantes, independientemente del tamaño, la complejidad, y la fuente de los datos corporativos. De hecho, la información puesto en bases de datos corporativas siempre ha sido más fácil que la eliminación de ellos. La información más grande y complejo almacenado, más difícil es para quitarlo. Tecnología OLAP elimina estas dificultades conducen a la información más cerca del usuario que lo necesita.

2.2.2. MODELO DE DATOS

En un modelo de datos OLAP, la información se organiza conceptualmente en cubos que almacenan valores o medidas cuantitativas. Las mediciones se identifican por dos o más descriptivos categorías dimensiones que forman la estructura de un cubo denominados. Una dimensión puede ser cualquier visión de negocio que tenga sentido para su análisis, como producto, departamento o tiempo. Este modelo de datos multidimensional simplifica para los usuarios el proceso de formulación de la investigación o "consultas" compleja, crear informes, realizar análisis comparativos, y muestra subconjuntos (rebanada) de interés. Por ejemplo, un cubo que contiene la información de ventas puede estar compuesta por las dimensiones de tiempo, área, producto, escenario de cliente (estimado o real) y medidas. Las medidas típicas que valor de las ventas, unidades vendidas, costos, márgenes, etc.

2.2.3. APLICACIONES.

Aplicación de OLAP es muy diversa y su uso es en diversas áreas de la empresa. Algunos tipos de aplicación donde se utiliza la tecnología son:

Finanzas	Análisis de L & P, G & P Informes, Económico, Análisis de Balances, Flujos de Efectivo, Cuentas por Cobrar,
Ventas	Análisis de ventas (por región, producto, proveedor, etc.), Pronóstico, Rentabilidad cliente / contrato, Análisis de Canal de Distribución, ...
Mercadeo	Análisis de precio / volumen, rentabilidad del producto, análisis de mercado, ...
Recursos Humanos	Análisis de Beneficios, Proyección Ganancias, Análisis "Recuento", ...
Manufactura	Gestión de Inventario, la cadena de suministro, planificación de la demanda, la materia prima Análisis de Costos, ...

Tabla 1: Aplicaciones OLAP, Autor: Creación Propia.

3. INTEGRACIÓN DE DATOS

3.1. LOS PROCESOS EXTRACT, TRANSFORM AND LOAD (LOAD)

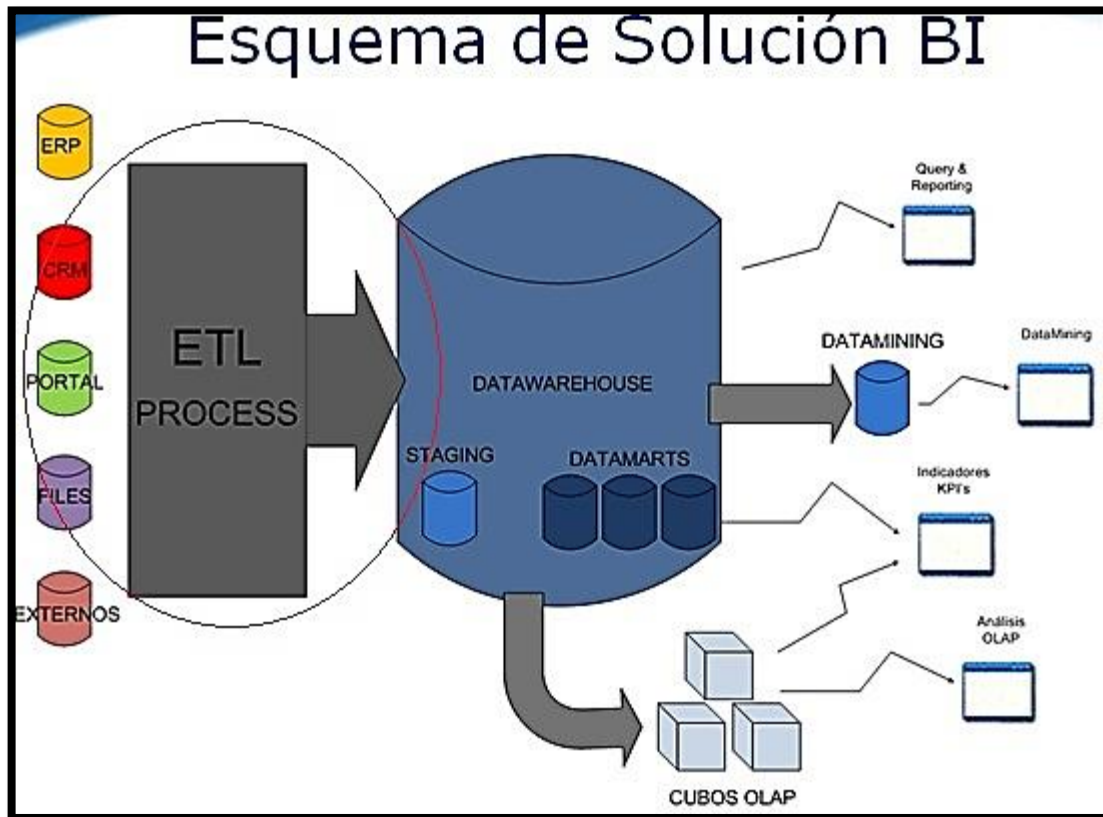


Fig.6: Esquema de Solución BI, Autor: Creación propia.

Para poder extraer los datos desde los OLTP, para luego manipularlos, integrarlos y transformarlos, para posteriormente cargar los resultados obtenidos en el DW, es necesario contar con algún sistema que se encargue de ello. Precisamente los ETL (Extracción, Transformación y Carga) son los que cumplirán con tal fin.

Tal y como sus siglas lo indican, los ETL, extraen datos de las diversas fuentes que se requieran, los transforman para resolver posibles problemas de inconsistencias entre los mismos y finalmente, después de haberlos depurado se procede a su carga en el depósito de datos.

3.1.1. EXTRACCION

Es aquí, en donde, basándose en las necesidades y requisitos del usuario, 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 propietario 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.

3.1.2. 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 DW. 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 DW estén integrados.

3.1.2.1. CODIFICACION

Una inconsistencia muy típica que se encuentra al intentar integrar varias fuentes de datos, es la de tener más de una sola forma de codificar un atributo en común. Por ejemplo, en el campo “género”, algunos diseñadores completan su valor con “0” y “1”, otros con “F” y “M”, otros con “Femenino” y “Masculino”, etc. Lo que se debe realizar en estos casos, es seleccionar o recodificar estos atributos, para que cuando la información llegue al DW, esté integrada de manera uniforme.

En la siguiente figura, se puede apreciar que de varias formas de codificar se escoge una, entonces cuando surge una codificación diferente a la seleccionada, se procede a su transformación.

3.1.2.2. 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 explicitadas 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.

3.1.2.3. CONVENCIONES DE NOMBRAMIENTO

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 el usuario sea más comprensible.

3.1.3. CARGA

Este proceso es el responsable de cargar la estructura de datos del DW con:

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.

3.2. LA 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:

- 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.
- Ignorarlos.

Un punto muy importante que se debe tener en cuenta al elegir alguna acción, es el de identificar el porqué 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.

4. DISEÑO DIMENSIONAL DEL PROCESO DEL NEGOCIO

4.1. CONCEPTOS DE MODELADO DIMENSIONAL

El modelado dimensional es una forma de acercar los datos a la manera en que estos serán convertidos en información útil para los usuarios del negocio. El objetivo final es que estos puedan encontrar de manera intuitiva y rápida la información que necesitan. La aplicación del modelo dimensional tiene lugar en la fase de diseño lógico, lo que permite la traducción del esquema resultante del diseño conceptual al plano lógico. El modelo dimensional se describe en el año 1996 por Ralph Kimball, como propuesta para el diseño de almacenes de datos (Data Warehouses), partiendo de la visión multidimensional que los usuarios tienen de los datos empresariales cuando se enfrentan a ellos con propósito de análisis (de análisis multidimensional –OLAP– en concreto).

El análisis multidimensional consiste en analizar los datos que hacen referencia a hechos, sean económicos o de otros tipos, desde la perspectiva de sus componentes o dimensiones (utilizando para ello algún tipo de métrica o medida de negocio). Este modelo tiene en cuenta que para el análisis multidimensional los datos se representan como si estuvieran en un espacio n-dimensional (cubo de datos), permitiendo su estudio en términos de hechos sujetos al análisis (facts, en inglés) y dimensiones que permiten diferentes puntos de vista por los que analizar esos hechos. La analogía del cubo (recuerde el cubo de Rubik) con la visión multidimensional es válida para comprender el concepto desde un punto de vista gráfico, pero sólo es válido para un modelo de tres dimensiones. Un modelo de más de tres dimensiones suele denominarse hipercubo, y ya resulta más difícil su representación gráfica.

El modelo dimensional distingue tres elementos básicos:

Hechos: es la representación en el data warehouse de los procesos de negocio de la organización. Por ejemplo: una venta puede identificarse como un proceso de negocio. Los hechos se podrán reconocer además porque siempre tienen asociada una fecha, y una vez registrados no se modifican ni se eliminan (para no perder la historia).

Métrica: son los indicadores de negocio de un proceso de negocio. Aquellos conceptos cuantificables que permiten medir nuestro proceso de negocio. Por ejemplo, en una venta tenemos el importe de la misma y la cantidad vendida. Existen métricas derivadas, como el precio unitario, que se obtiene al dividir el importe total por las unidades vendidas.

Dimensión: es la representación en el data warehouse de un punto de vista para los hechos de cierto proceso de negocio. Si regresamos al ejemplo de una venta, para la misma tenemos el cliente que ha comprado, la fecha en la que se ha realizado, el producto vendido,... Estos conceptos pueden ser considerados como vistas para este proceso de negocio. Puede ser interesante recuperar todas las compras realizadas por un cliente, o para un producto o familia de productos, o para un lapso determinado. Pero ¿qué hay de su representación? Igual que sucede en el modelo relacional, el modelo dimensional adopta el concepto de relación (tabla) como estructura básica del modelo. Pero a diferencia del modelo relacional, que no hace distinción entre relaciones, el modelo dimensional distingue entre relaciones de hecho (tablas de hecho) y relaciones de dimensión (tablas de dimensión).

4.2. TABLAS DE DIMENSIONES

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

Representan los ejes del cubo, y los aspectos de interés, mediante los cuales el usuario podrá filtrar y manipular la información almacenada en la tabla de hechos.

En la siguiente figura se pueden apreciar algunos ejemplos:

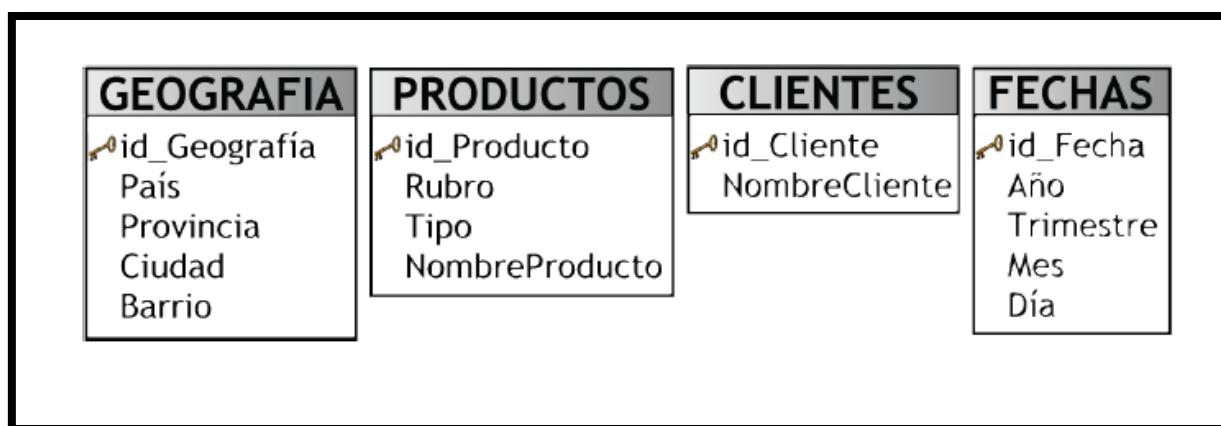


Fig. 7: Tabla de Dimensiones, Autor: Ing. Bernabéu, Ricardo Dario.

Como se puede observar, cada tabla posee un identificador único y al menos un atributo que describe los criterios de análisis relevantes de la organización, estos son por lo general de tipo texto. Usualmente la cantidad de tablas de dimensiones, aplicadas a un tema de interés en particular, varían entre tres y quince.

Así mismo, dentro de estas tablas pueden existir jerarquías¹¹ de datos, además, de acuerdo a las dimensiones del negocio, estará dada la granularidad que adoptará el modelo.

Los datos dentro de estas tablas, que proveen información del negocio o que describen alguna de sus características, son llamados datos de referencia. Entonces, se puede afirmar que una tabla de dimensión posee una clave primaria y uno o más datos de referencia.

4.2.1. DIMENSIÓN TIEMPO

En un DW, la dimensión Tiempo es obligatoria, y la definición de granularidad y jerarquía de la misma depende de la dinámica del negocio que se esté analizando, toda la información dentro de la bodega, como ya se ha explicado, posee su propio sello de tiempo que determina la ocurrencia y ubicación con elementos en iguales condiciones, representando de esta manera diferentes versiones de una misma situación.

Es importante tener en cuenta que el tiempo no es solo una secuencia cronológica representada de forma numérica, sino que posee fechas 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, 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. Si bien, el lenguaje SQL ofrece funciones del tipo DATE, en la dimensión Tiempo, se modelan y presentan atributos temporales que no pueden calcularse en SQL, lo cual le añade una ventaja más.

4.2.2. JERARQUIAS

Una jerarquía representa una relación¹² lógica entre dos o más atributos dentro de una misma dimensión.

Las jerarquías poseen las siguientes características:

Pueden existir varias en una misma dimensión.

Están compuestas por dos o más niveles.

Se tiene una relación “1-n” entre atributos consecutivos de un nivel superior y uno inferior.

La principal ventaja de manejar jerarquías, reside en poder analizar los datos desde su nivel más general al más detallado y viceversa.

La siguiente figura muestra un pequeño ejemplo de lo recién expuesto:

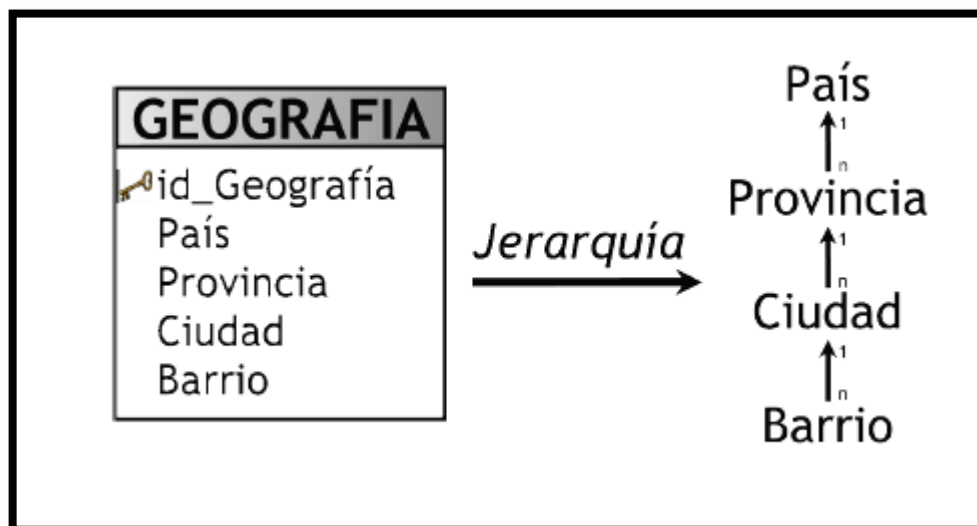


Fig. 8: Jerarquía de GEOGRAFIA, Autor: Ing. Bernabéu, Ricardo Dario.

En la figura anterior, se puede apreciar claramente cómo se construye una jerarquía. Se coloca el atributo más general en la cabecera y se comienza a disgregar los niveles hacia abajo, dependiendo siempre del rango del dato. En este caso, la figura se explica como sigue:

Un barrio pertenece solo a una ciudad. Una ciudad puede poseer uno o más barrios. Una ciudad pertenece solo a una provincia. Una provincia puede poseer una o más ciudades.

Una provincia pertenece solo a un país. Un país puede poseer una o más provincias.

4.2.3. RELACION

Una relación representa la forma en que dos atributos interactúan dentro de una jerarquía.

Existen básicamente dos tipos de relaciones:

Explícitas: son las más comunes y se pueden modelar a partir de atributos directos y están en línea continua de una jerarquía, por ejemplo, la figura anterior posee este tipo de relación, en donde un país posee una o más provincias y una provincia pertenece solo a un país.

Implícitas: son las que ocurren en la vida real, pero su relación no es de vista directa, por ejemplo, un país tiene uno o más ríos, pero un río pertenece a uno o más países.

4.2.4. 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 la empresa, pueden registrarse día a día, en cambio, los datos 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 de análisis, ya que los mismos podrán ser resumidos o sumariados.

4.3. TABLAS DE HECHOS

Las tablas de hechos contienen los hechos, medidas o indicadores que serán utilizados por los analistas de negocio para apoyar el proceso de toma de decisiones. 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 de los intervalos de tiempo de los mismos. 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:

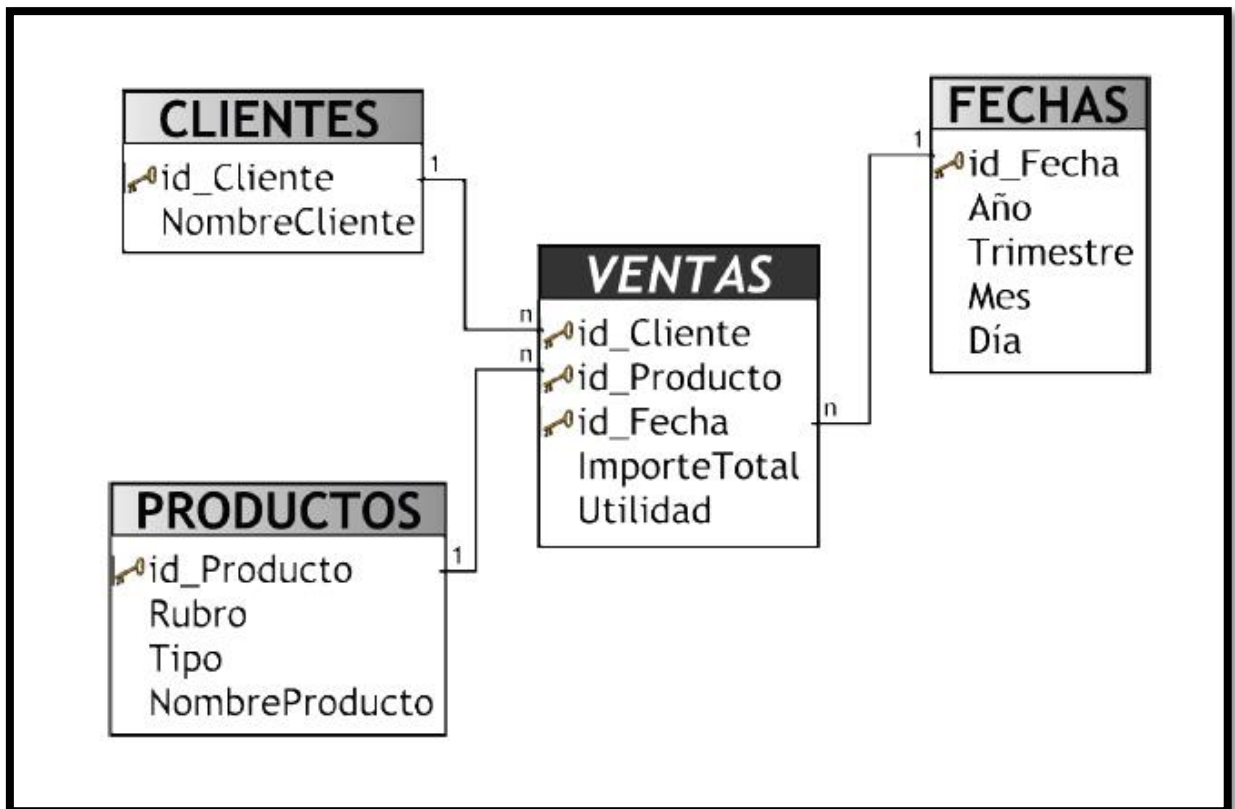


Fig. 9: Tabla de Hechos, Autor: Ing. Bernabéu, Ricardo Dario.

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”, “PRODUCTOS” y “FECHAS”, que están conectadas mediante sus claves primarias. Es por ello precisamente que la clave primaria de la tabla de hechos es la combinación de las claves primarias de sus dimensiones. Las medidas 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 todas aquellas sumalizaciones o acumulaciones preestablecidas que residen en una tabla de hechos para agilizar las consultas y permitir que los datos puedan ser accedidos por las diferentes dimensiones, y desde luego, explorados por ellas. Las sumalizaciones no están referidas solo a sumas, sino que también a promedios, mínimos, máximos, totales por sector, porcentajes, fórmulas predefinidas, etc, dependiendo de los requerimientos de información del negocio.

4.4. CONCEPTOS ADICIONALES

4.4.1. 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

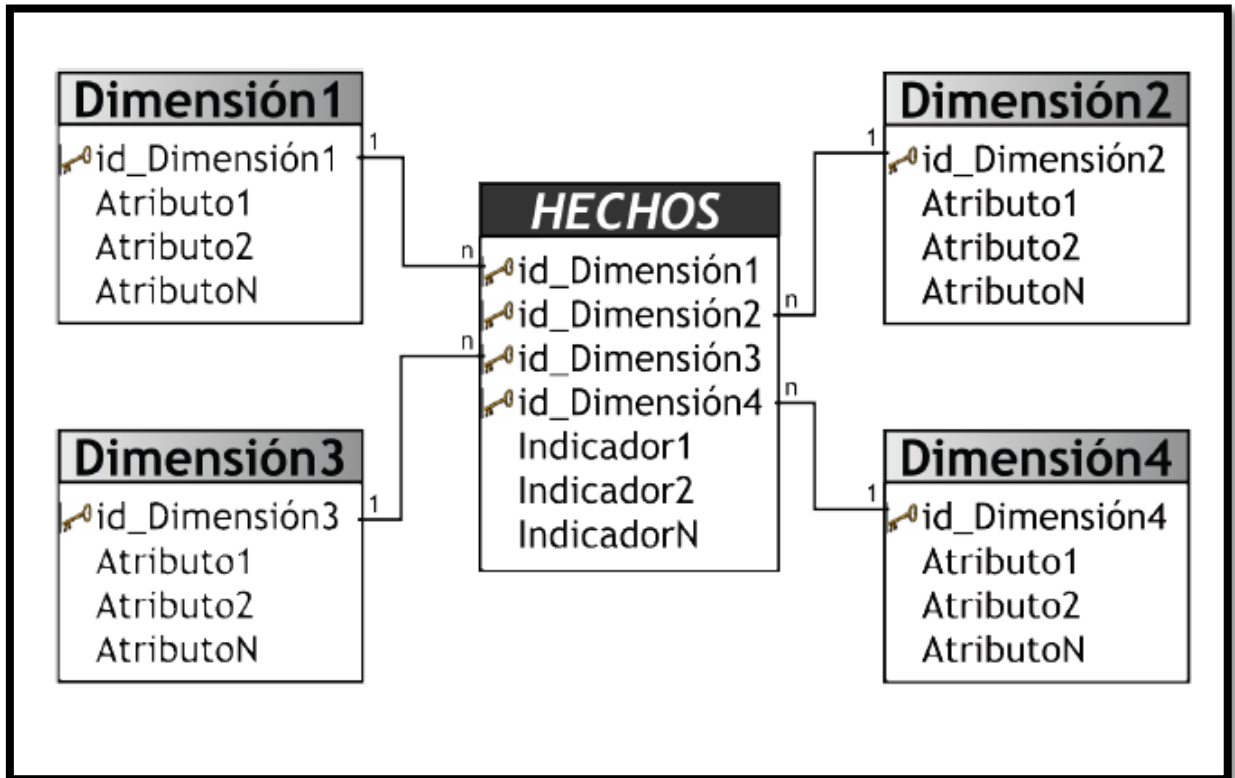


Fig. 10: Esquema en Estrella, Autor: Ing. Bernabéu, Ricardo Dario.

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.

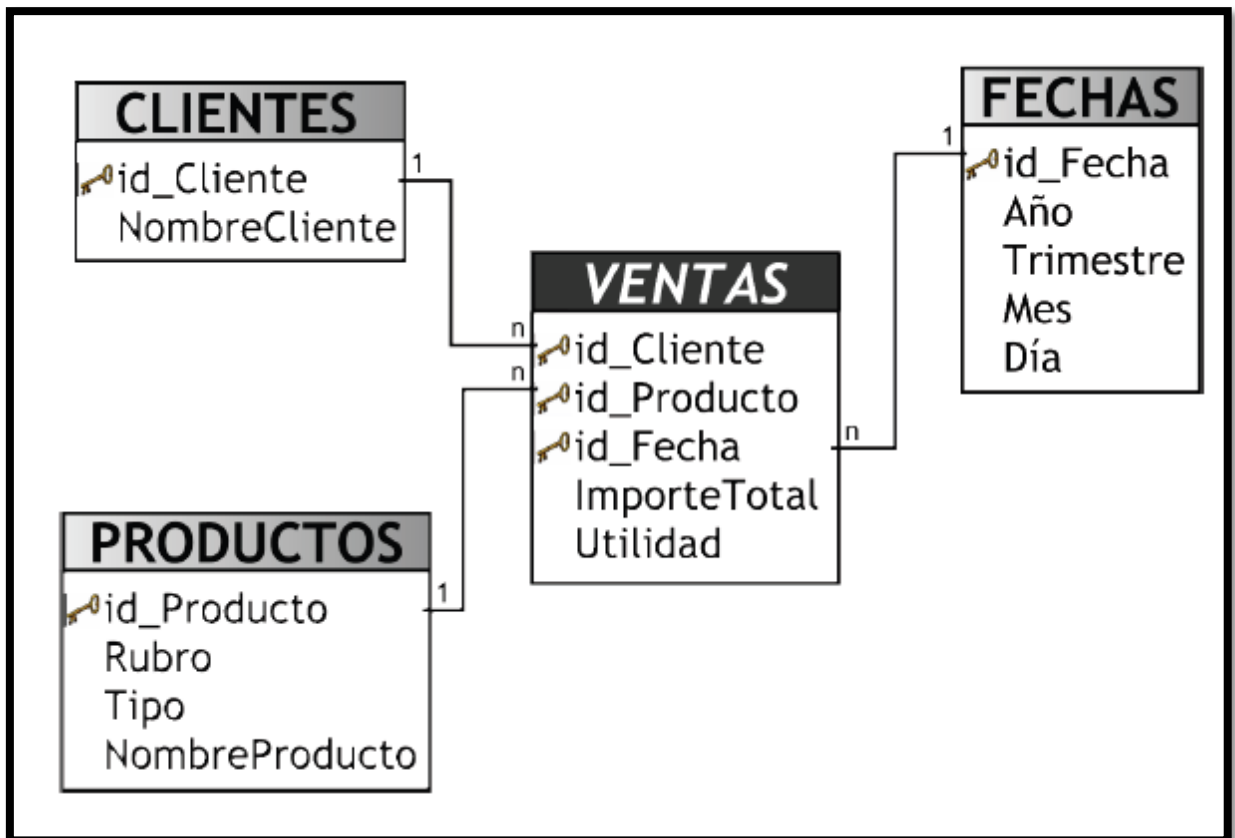


Fig. 11: Esquema en estrella 2, Ejemplo, Autor: Ing. Bernabéu, Ricardo Dario.

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 dimensión "PRODUCTOS" contiene los atributos "Rubro", "Tipo" y "NombreProducto". Si se normaliza esta tabla, se obtendrá el siguiente resultado:

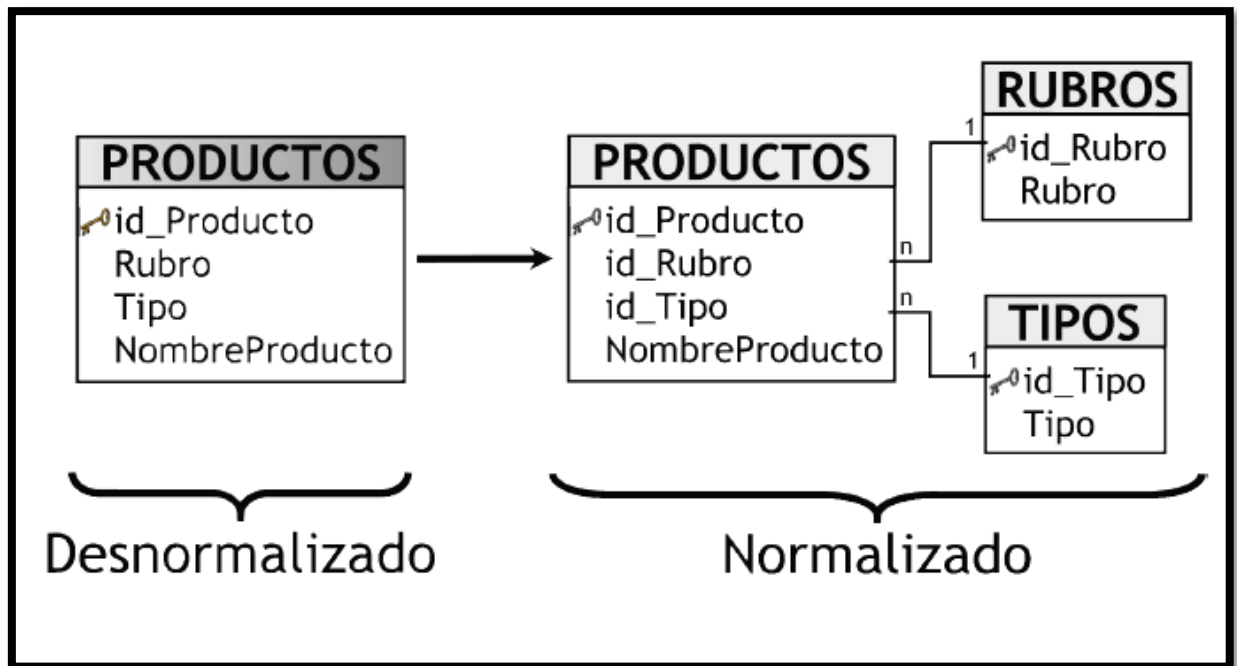


Fig. 12: Desnormalización, Autor: Ing. Bernabéu, Ricardo Dario.

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.

4.4.2. ESQUEMA COPO DE NIEVE

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

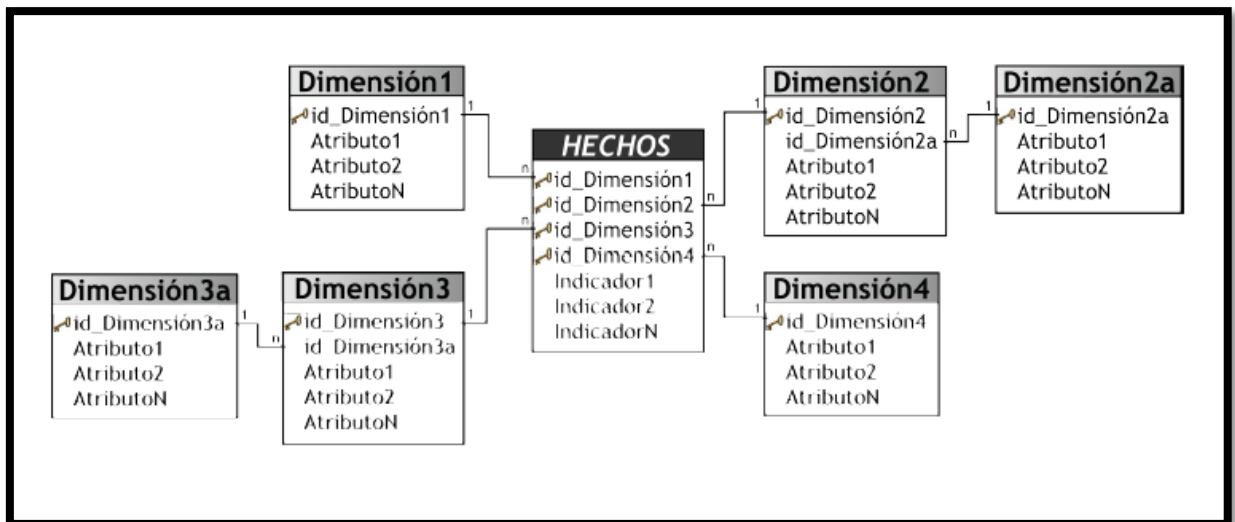


Fig. 13: Esquema Copo de Nieve, Autor: Ing. Bernabéu, Ricardo Dario.

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 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 dimensiones están normalizadas, por lo que requiere menos esfuerzo de diseño.

- Puede desarrollar clases de jerarquías fuera de las dimensiones, que permiten realizar análisis de lo general a lo detallado y viceversa.

4.4.3. ESQUEMA CONSTELACIÓN

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 sumalizaciones 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.

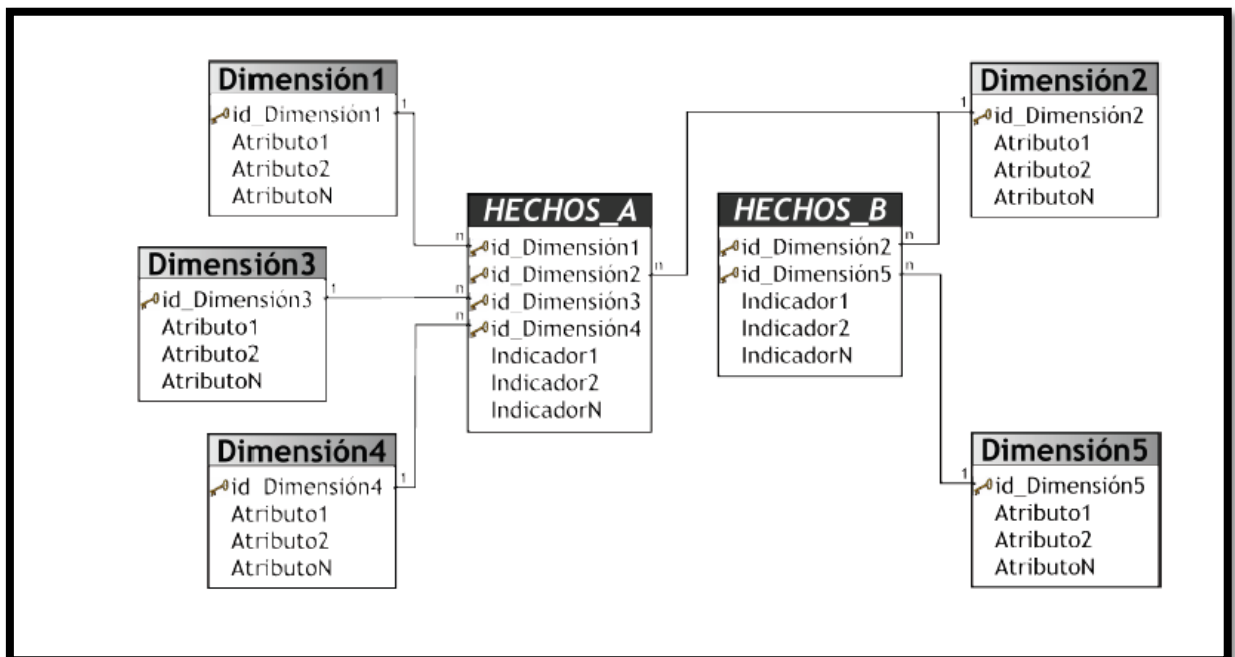


Fig. 14: Esquema Constelación, Autor: Ing. Bernabéu, Ricardo Dario.

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 aspectos claves del negocio con un mínimo esfuerzo adicional de diseño.

Contribuye a la reutilización de dimensiones, ya que una misma dimensión puede utilizarse para varias tablas de hechos.

No es soportado por todas las herramientas de consulta y análisis.

5. LA MINERIA DE DATOS

Data warehouse es una colección de datos históricos, que incluyen la copia de las transacciones de datos específicamente estructurados para la consulta y el análisis. Tal como lo indica su nombre, es el almacén de los datos estructurados.

Una vez que esa información está guardada y organizada en el data warehouse, se usa el data mining para explorarla y clasificarla, en busca de patrones para big data.

“Data mining es un conjunto de técnicas de extracción de datos, para detectar patrones de comportamiento a través de algoritmos matemáticos”, manifestó Licon.

Adicionalmente, sobre la minería de datos se puede ejecutar un conjunto de técnicas para realizar análisis predictivos y de tendencias; esto se conoce como analítica de datos (data analytics).

El **análisis de datos** es un proceso de inspeccionar, limpiar y transformar datos con el objetivo de resaltar información útil, lo que sugiere conclusiones, y apoyo a la toma de decisiones. El análisis de datos tiene múltiples facetas y enfoques, que abarca diversas técnicas en una variedad de nombres, en diferentes negocios, la ciencia, y los dominios de las ciencias sociales.

5.1. PRINCIPALES MODELOS DE ANALISIS DE DATOS

5.1.1. ANALISIS FACTORIAL

Es una técnica estadística de reducción de datos usada para explicar las correlaciones entre las variables observadas en términos de un número menor de variables no observadas llamadas factores. Las variables observadas se modelan como combinaciones lineales de factores más expresiones de error. El análisis factorial se originó en psicometría, y se usa en las ciencias del comportamiento tales como ciencias sociales, marketing, gestión de productos, investigación de operaciones y otras ciencias aplicadas que tratan con grandes cantidades de datos.

5.1.1.1. TIPOS DE ANALISIS FACTORIAL

El *análisis factorial exploratorio*, AFE, se usa para tratar de descubrir la estructura interna de un número relativamente grande de variables. La hipótesis *a priori* del investigador es que pueden existir una serie de factores asociados a grupos de variables. Las *cargas* de los distintos factores se utilizan para intuir la relación de éstos con las distintas variables. Es el tipo de análisis factorial más común.

El *análisis factorial confirmatorio*, AFC, trata de determinar si el número de factores obtenidos y sus *cargas* se corresponden con los que cabría esperar a la luz de una teoría previa acerca de los datos. La hipótesis *a priori* es que existen unos determinados factores preestablecidos y que cada uno de ellos está asociado con un determinado subconjunto de las variables. El análisis factorial confirmatorio entonces arroja un nivel de confianza para poder aceptar o rechazar dicha hipótesis.

5.1.1.2. APLICACIONES

El análisis factorial se utiliza para identificar factores que expliquen una variedad de resultados en diferentes pruebas. Por ejemplo, investigación en inteligencia halla que la gente que obtienen una nota alta en una prueba de habilidad verbal también se desempeña bien en pruebas que requieren habilidades verbales. Los investigadores explican esto mediante el uso de análisis factorial para aislar un factor a menudo llamado inteligencia cristalizada o inteligencia verbal, que representa el grado en el cual alguien es capaz de resolver problemas usando habilidades verbales.

Análisis factorial en psicología se asocia frecuentemente con la investigación sobre la inteligencia. Sin embargo, también se ha utilizado en un amplio rango de dominios, tales como personalidad, actitudes, creencias, etc. Está asociado a la psicometría, debido a que puede evaluar la validez de un instrumento estableciendo si el instrumento de verdad mide los factores postulados.

5.1.2. ANALISIS PREDICTIVO

El análisis predictivo agrupa una variedad de técnicas estadísticas de modelización, aprendizaje automático y minería de datos que analiza los datos actuales e históricos reales para hacer predicciones acerca del futuro o acontecimientos no conocidos.^{1 2} En el ámbito de los negocios los modelos predictivos extraen patrones de los datos históricos y transaccionales para identificar riesgos y oportunidades. Los modelos predictivos identifican relaciones entre diferentes factores que permiten valorar riesgos o probabilidades asociadas en base a un conjunto de condiciones, guiando así al decisor durante las operaciones de la organización.³

El efecto funcional que pretenden estas iniciativas técnicas es que el análisis predictivo provea una puntuación (probabilidad) para cada sujeto (cliente, empleado, paciente, producto, vehículo, componente, máquina y otra unidad en la organización. con el objeto de determinar, informar o influir procesos en la organización en el que participen un gran número de sujetos, tal y como ocurre en marketing, evaluación de riesgo de crédito, detección de fraudes, fabricación, salud y operaciones gubernamentales como el orden público.

5.1.2.1. DEFINICION

El análisis predictivo es un área de la minería del dato que pretende extraer conocimiento que le permita predecir tendencias y patrones de comportamiento. A menudo una circunstancia desconocida de interés se va a producir en el futuro pero el análisis predictivo se puede aplicar igualmente a lo desconocido tanto en el pasado, el presente o el futuro. Por ejemplo, identificar sospechosos después de haberse producido un crimen o un fraude con tarjeta de crédito. Lo fundamental del análisis

predictivo está en identificar relaciones entre las variables explicativas y las variables predictivas del pasado de forma que se pueda escalar a lo que está por ocurrir. Es importante advertir, en cualquier caso, que la fiabilidad y usabilidad de los resultados dependerán mucho del nivel de análisis del dato y la calidad de las hipótesis.

El análisis predictivo es a menudo conocido por predecir a un nivel de granularidad más elevado, por ejemplo, generando puntuaciones predictivas (probabilidades) para cada sujeto en la organización. Eso lo diferencia de la anticipación. Por ejemplo "la tecnología de análisis predictivo que aprende de la experiencia (dato) para predecir el comportamiento futuro de los individuos con el objetivo de tomar mejores decisiones"

5.1.2.2. TIPOS

Cuando se habla de análisis predictivo generalmente se quiere hablar de "modelos predictivos", datos de puntuaciones en base a modelos predictivos y previsiones. No obstante se está generalizando el uso del término para relacionarlo con disciplinas analíticas y está muy extendido su uso para la segmentación entre usuarios de negocio y decisores. Los propósitos y las técnicas estadísticas subyacentes en ambos casos varían.

Modelos predictivos

Los modelos predictivos son modelos de la relación entre el rendimiento específico de un sujeto en una muestra y uno o más atributos o características del mismo sujeto. El objetivo del modelo es evaluar la probabilidad de que un sujeto similar tenga el mismo rendimiento en una muestra diferente. Esta categoría engloba modelos en muchas áreas como el marketing, donde se buscan patrones de datos ocultos que respondan preguntas sobre el comportamiento de los clientes o modelos de detección de fraude.. Los modelos predictivos a menudo ejecutan cálculos durante las transacciones en curso, por ejemplo, para evaluar el riesgo o la oportunidad de un cliente o transacción en particular, de forma que aporte conocimiento a la hora de tomar una decisión. Gracias a los avances de ingeniería en el análisis de grandes volúmenes de datos estos modelos son capaces de simular el comportamiento humano frente a estímulos o situaciones específicas.

Modelos descriptivos

Los modelos descriptivos cuantifican las relaciones entre los datos de manera que es utilizada a menudo para clasificar clientes o contactos en grupos. A diferencia de los modelos predictivos que se centran en predecir el comportamiento de un cliente en particular (cómo ocurre con el riesgo de crédito), los modelos descriptivos identifican muy diferentes relaciones entre los clientes y los productos. Los modelos descriptivos no clasifican u ordenan a los clientes por su probabilidad de realizar una acción particular de la misma forma en la que lo hacen los modelos predictivos. Sin embargo,

los modelos descriptivos pueden ser utilizados por ejemplo para asignar categorías a los clientes según su preferencia en productos o su franja de edad.

5.1.2.3. APLICACIONES

El análisis predictivo puede ser aplicado en muchas circunstancias y las siguientes son sólo algunos ejemplos en los que el análisis predictivo ha demostrado tener un impacto especialmente positivo durante los últimos años:

CRM o Administración basada en la relación con los clientes, Asistencia sanitaria, Análisis de cobros, Venta cruzada, Fidelización del cliente, Mercadotecnia directa, Detección de fraude, Predicción de cartera, producto o economía, Gestión de riesgo.

5.1.3. ANALISIS EXPLORATORIO DE DATOS

El análisis exploratorio de datos definido por John W. Tukey (E.D.A.: *Exploratory data analysis*) es, básicamente, el tratamiento estadístico al que se someten las muestras recogidas durante un proceso de investigación en cualquier campo científico. Para mayor rapidez y precisión, todo el proceso suele realizarse por medios informáticos, con aplicaciones específicas para el tratamiento estadístico. Los E.D.A., no necesariamente, se llevan a cabo con una base de datos al uso, ni con una hoja de cálculo convencional; no obstante el programa SPSS y R (lenguaje de programación) son las aplicaciones más utilizadas, aunque no las únicas.

Por ejemplo, en el campo de la Arqueología el análisis técnico de una pieza puede ser simultáneo a la introducción de los datos, bien porque las fichas estén directamente informatizadas o, bien, porque se usen formularios en papel cuyos datos sean fáciles de introducir en el ordenador o computadora. Es posible, incluso, usar en la propia excavación, una serie de PDAs conectados en red inalámbrica instalada en el yacimiento arqueológico, que envíen numerosos datos de campo a una base de datos central que luego se usarán con fines diversos, entre ellos éste. Los pasos seguidos en el E. D. A. son básicamente dos:

- **Medición y descripción** de los datos tecnológicos —tipológicos— y dimensiones, por medio de la Estadística descriptiva. Aquí tenemos, por un lado, las medidas de tendencia central (promedios que, en una sola cifra, resumen todos los valores de una muestra: media, mediana y moda son las más habituales) y, por otro, las medidas de dispersión (que calculan hasta qué punto la muestra se agrupa o no en torno a esos promedios). Dentro de este apartado, se ha de procurar, además, calibrar la confianza de las muestras a través de tres estadímetros básicos: la desviación estándar de la muestra, la curtosis y la asimetría.

5.2. EL MANEJO DE DATOS NO ESTRUCTURADOS

La gestión de los **datos no estructurados** se ha convertido en uno de los principales retos a los que hacen frente las compañías en lo relativo a gestión de información y **Big Data**.

5.2.1. DEFINICION DE DATOS NO ESTRUCTURADOS.

Una posible definición de datos no estructurados, son aquellos datos no almacenados en una base de datos tradicional. La información no estructurada no puede ser almacenada en estructuras de datos relacionales predefinidas.

Se pueden establecer diferentes clasificaciones, vamos a considerar dos de ellas.

- **Datos no estructurados y semiestructurados.** Los datos semiestructurados serían aquellos datos que no residen de bases de datos relacionales, pero presentan una organización interna que facilita su tratamiento, tales como documentos XML y datos almacenados en bases de datos NoSQL.
- **Datos de tipo texto y no-texto.** Datos no estructurados de tipo texto podrían ser datos generados en las redes sociales, foros, e-mails, presentaciones Power Point o documentos Word, mientras que datos no-texto podrían ser ficheros de imágenes jpeg, ficheros de audio mp3 o ficheros de video tipo flash.

5.2.2. CARACTERISTICAS DE DATOS NO ESTRUCTURADOS

Las principales características de los datos no estructurados son las siguientes:

- **Volumen y crecimiento:** el volumen de datos y la tasa de crecimiento de los datos no estructurados es muy superior al de los datos estructurados. Por ejemplo, twitter genera 12 Terabytes de información cada día. De acuerdo con Gartner, la tasa anual de crecimiento de datos es del 40 a 60 por ciento, pero para los datos no estructurados en empresas, la tasa de crecimiento puede llegar al 80 por ciento (informe 2012).
- **Orígenes de datos:** El origen de los datos es muy diverso: datos generados en redes sociales, datos generados en foros, e-mails, datos extraídos de la web empleando técnicas de web semántica, documentos internos de la compañía (word, pdf, ppt).
- **Almacenamiento:** Debido a su estructura no podemos emplear arquitectura relacional, siendo necesario trabajar con herramientas 'Big Data', siendo crítico en estas arquitecturas los aspectos relacionados con la escalabilidad y paralelismo. Según el tipo de dato se impone el almacenamiento cloud. Monitorizar la frecuencia de uso y la detección de datos inactivos son aspectos críticos de cara a reducir costes de almacenamiento.

- **Terminología e idiomas:** La terminología es una cuestión crítica tratando datos no estructurados de tipo texto. Es habitual llamar a lo mismo de diferentes formas, de tal modo que es necesario una racionalización de la terminología. Otra cuestión es el idioma en el que se ha generado la información tratada.
- **Seguridad:** Hay que considerar que algunos datos no estructurados de tipo texto, pueden no ser seguros. Por otra parte el control de accesos a los mismos es complejo debido a cuestiones de confidencialidad y la difícil clasificación del dato.

5.2.3. TRATAMIENTO DE DATOS NO ESTRUCTURADOS

Las principales cuestiones a considerar en el tratamiento de información no estructurada son las siguientes:

- **Crear una plataforma escalable (infraestructura y procesos)** que permita tratar grandes cantidades de datos. Las tecnologías RDBMS son insuficientes para tratar información no estructurada. Es necesaria una capacidad de almacenamiento y una capacidad de proceso escalable. Teniendo en cuenta que el coste económico de mantener plataformas escalables, hay que considerar la opción cloud. Desde el punto de vista de los procesos, en ocasiones es interesante utilizar in-memory analytics.
- **Añadir información/estructura complementaria a los datos no estructurados.** Es importante añadir algún tipo de estructura a los datos no estructurados que ayude a su tratamiento. Por ejemplo, en una colección de tweets de redes sociales puede ser interesante añadir campos tales como el idioma, la localización geográfica para su posterior procesamiento. Esta estructura adicional que añadimos debe ser modelizada de cara a estar en constante evolución.
- **Crear conjuntos reducidos de datos que sean representativos.** Dado el volumen ingente de información, es importante trabajar con muestras de datos que sean estadísticamente representativos sobre los datos a analizar. Muchos análisis pueden llevarse a cabo con un grado de exactitud razonable, utilizando conjuntos de datos que son más pequeños en un orden de magnitud que la información en bruto.
- **Desarrollo de algoritmos.** Hay diferentes tipos de aproximación hacia la información no estructurada. Por ejemplo, para procesos de text mining, puede utilizarse natural language processing combinado con redes neuronales. Otras técnicas como redes bayesianas permiten descubrir patrones sobre múltiples dimensiones. Son importantes también las técnicas de visualización de datos.

- **Procesos de depuración/limpiado de datos.** Dado el ingente volumen de datos, se convierte en crítico la correcta gestión del histórico de datos. Detección de datos no usados o de frecuencia de consulta muy baja con objeto de limpiar información y liberar espacio.

Ejemplo sencillo tratamiento de datos no estructurados - redes sociales.

Dada la variada naturaleza de los datos no estructurados, hay infinidad de posibles procesos relacionados con ellos. A continuación mostramos un sencillo ejemplo de tratamiento de datos provenientes de redes sociales.

El objetivo de este análisis de datos es conocer la percepción que existe sobre el precio de determinado producto en twitter.

- **Extracción:** Utilizando una clase de java (ejemplo twitter4j) leemos el feed de Twitter disponible en <https://twitter.com/search/realtime>. Añadimos a los campos disponibles calificaciones del tipo: idioma, localización geográfica.
- **Transformación:** Filtramos todos aquellos tuits que contengan el nombre del producto. Refinamos el filtro introduciendo campos del tipo (“precio”) + (“barato”, “caro”, “económico”, etc...) , teniendo en cuenta el idioma en el que se generan los tuits. Valorar la opción en base al volumen de obtener una muestra representativa de los datos extraídos y filtrados.
- **Volcado a BBDD :** Insertamos en una tabla el registro del tuit con la calificación identificada (idioma, localización geográfica)
- **Informes:** Creamos informe que permita realizar análisis por tiempo y campos de calificación. Hay que considerar que este informe puede ser actualizado en tiempo real.

6. CASO DE ÉXITO.



Arcor se apoyo en MicroStrategy para crear un datawarehouse que brinde información de gestión a todas las áreas de la Compañía.

La principal productora de caramelos del mundo implementó soluciones de Business Intelligence de MicroStrategy para consolidar la información de los sistemas en todos los países donde tiene operación.

Arquitectura:

- Sistemas Fuente: J.D. Edwards, People Soft, Demantra, Dc Link, Siebel, sistemas legacy, etc
- ETL: Datastage
- Base de Datos: Teradata
- Plataforma BI: MicroStrategy

Aplicaciones:

- Áreas: Comercial de Mercado Interno y Comercio Exterior, Recursos Humanos, Compras, Marketing.
- Operaciones – Manufactura, Logística Industrial y Comercial.
- Administración Financiera, Sistemas

Arcor es una productora de alimentos de origen argentino, considerada una de las principales multilatinas de la región, con una facturación anual estimada en US\$ 3.500 millones hacia el cierre de 2012. Tras haber encarado un agresivo proceso de expansión internacional en la década del 90, la compañía se vio obligada a implementar soluciones de inteligencia de negocios para administrar con eficiencia la información comercial que los negocios de Consumo Masivo requerían. Las herramientas internas desarrolladas en Visual Basic, así como la extracción de información a través de Business Objects, cubos de Power Play y planillas de Excel no eran suficientes. Por esa razón, en el año 2003, la compañía implementó las primeras soluciones de **MicroStrategy**.

Pudo, así, crear el datawarehouse que, aún hoy, es el principal y único repositorio de datos al que acuden las distintas áreas de la compañía para realizar sus reportes de gestión.

Desde el año 2009 la compañía viene creciendo a ritmos que rondan el 20%. En ese año, sus ingresos fueron de US\$ 2.200 millones, mientras que en 2010 esa cifra trepó hasta los US\$ 2.600 millones. Un año más tarde, la facturación del grupo ascendió a US\$ 3.100 millones y, como se dijo, se prevé un 2012 con ingresos en torno a los US\$ 3.500 millones.

El buen desempeño global del grupo argentino con mayor cantidad de mercados abiertos en el mundo se observa también en el terreno de las exportaciones. En 2009 el grupo vendió al exterior US\$ 310 millones (US\$ 243 millones son sólo de la Argentina), que se convirtieron en US\$ 360 millones a finales de 2010. En 2011 las cosas continuaron mejorando en ese terreno, hasta los US\$ 380 millones en concepto de exportación y la previsión para el cierre de 2012 es que las ventas externas se ubiquen en los US\$ 470 millones, de los cuales US\$ 350 millones serán aportados por la Argentina.

Con semejante comportamiento de negocio, el nivel de información que se genera crece en la misma proporción. La necesidad de contar con datos precisos para tomar

decisiones estratégicas que permitan mantener esas cifras en alza sólo se logra cuando hay un soporte tecnológico que acompañe ese ritmo.

Todo en uno

El datawarehouse que tuvo inicio en los procesos comerciales de Consumo Masivo Argentina se fue extendiendo al resto de las áreas.

De esta manera se fue sumando información de ventas internas y externas de las distintas filiales, de recursos humanos de Argentina, Brasil y Chile, de Compras y lo vinculado con promociones y marketing, que se basan en encuestas para quioscos para integrar información como fotos, precios y demás aspectos de los distintos productos cuya explotación permite confeccionar indicadores de cobertura.

El área de manufactura también fue de la partida. El proyecto apuntó a determinar indicadores de producción y tiempos de cada línea. La herramienta de BI de **MicroStrategy** permitió construir una gran cantidad de tableros para poder tener datos fehacientes sobre cada uno de estos puntos.

En cuanto al uso interno, la implementación abordó la puesta en marcha de la mesa de ayuda centralizada del área de TI que, por caso, tomara en cuenta no sólo la cantidad de incidentes que se reportan diariamente sino, particularmente, los tiempos de resolución de cada uno de ellos y el cumplimiento de los objetivos. En definitiva, la evaluación integral de la eficiencia del sistema.

“El trabajo actual contempla la ejecución de 13 proyectos que traen la información de ventas de todos los países donde opera la compañía, en América, Asia y Europa. Los proyectos apuntan a consolidar la información de las ventas de los mercados internos y las del comercio exterior, además de la información de recursos humanos de Argentina, Brasil y Chile; información de compras; de promociones y marketing; de manufactura y sus indicadores de producción; y de uso interno, como la mesa de ayuda centralizada”, detalló Valeria Dorronsoro, líder de proyecto de Arcor.

Apuesta Tecnológica

Las inversiones en el crecimiento de la compañía son una constante en el Grupo Arcor. Los desembolsos previstos para el período 2011/2012 alcanzan los US\$ 300 millones, de los cuales la mitad se está destinando a la actualización tecnológica y ampliación de la capacidad productiva del Grupo en la Argentina, Brasil y México. El 50% restante involucra la inversión en dos plantas industriales, una de molienda húmeda en Arroyito, Córdoba, y otra de golosinas, chocolates y caramelos en Chile.

“Como parte de un proyecto de actualización integral de las plataformas tecnológicas de la compañía en el mundo, se unificaron los criterios de los sistemas viejos y nuevos, se tomó la información que provenía de ambos mundos y se consolidó en el datawarehouse”, añade Dorronsoro.



Junto a **MicroStrategy** se desarrollaron las interfaces de donde se obtenía la información para confeccionar los reportes y demás actividades que exigía cada área y/o filial de la compañía. *“Al ser el sistema de BI el único lugar donde se alojaban los datos de manera unificada, esto pasó a ser el input para alimentar otras herramientas, como las vinculadas con evaluación de costos, cálculos y pronósticos, entre muchas más. La plataforma, así, pasó a ser transaccional”*. El datawarehouse es la única fuente de información oficial, y desde la cual se extraen los datos para la confección de todos los indicadores de gestión. Una gran ventaja en una corporación donde las distintas filiales, y sus diversas unidades de negocios, pueden confeccionar sus propios reportes. Arcor cuenta con 40 plantas industriales, de las cuales 29 se ubican en la Argentina, cinco en Brasil, cuatro en Chile, una en Perú y una en México. El plan del Datawarehouse para este año incluye la incorporación de aquellos negocios del Grupo Arcor que aún no fueron implementados.

“Dentro del proyecto de actualización tecnológica, tenemos el objetivo para 2012 de agregar al datawarehouse los Negocios de Packaging (Papel y Cartón y Flexibles)”, indica Dorronsoro.

Beneficios

- Cambios de versiones ágiles y sencillas.
- Mayor eficiencia en la confección de reportes.
- Plataforma transaccional.
- Confección independiente de reportes.

IV. CONCLUSIONES

- El Data Warehouse – OLAP, es una herramienta de gran potencial que a partir de datos procesados en información se convierte en conocimiento para generar tomas de decisiones acertadas dentro de la organización, tanto a nivel estratégico, táctico y operativo, propiciando ventajas competitivas.
- Se ha identificado la gran relevancia de las características e importancia de un Modelo de Data Warehouse que se implementa dentro de un core de negocio y los impactos beneficiosos en gran medida que generan, como son de cubrir la necesidad de información en el tiempo preciso y de aportar a explotar y maximizar el valor de la información.
- Se ha conseguido identificar los procesos pilares de integración de datos como son la Extracción, Transformación y Carga existentes dentro de un sistema Data Warehousing.
- Se ha logrado comprender los componentes del Modelado Dimensional, que es una técnica de modelado muy diferente al modelo Entidad-Relación de base de datos, que de acuerdo a sus características son fáciles de consultar por las tecnologías OLAP.

V. REFERENCIAS BIBLIOGRAFICAS

[1] [Data Warehousing-Hefesto, 2007]

Autor. Bernabeu, Ricardo Dario Córdoba, **Título.** DATA WAREHOUSING: Investigación y Sistematización de Conceptos – HEFESTO: Metodología propia para la Construcción de un Data Warehouse - Ing. Bernabeu, Ricardo Dario, **Version.** 0.1, **Lugar de publicación.** Cordova, Argentina. **Año.** 2007.

[2] [Estrategia Competitiva, 2000]

Autor. Michael E. Porter. **Título.** ESTRATEGIA COMPETITIVA, Técnicas para el Análisis de los Sectores Industriales y de la Competencia. **Versión.** Vigésima séptima reimpresión. **Año.** 2000.

[3] [El Nuevo Directivo Racional, 1992]

Autor. Charles H. Kepner, Benjamin B. Tregoe, McGraw Hill **Título.** El Nuevo Directivo Racional, Análisis de problemas y toma de decisiones. **Año.** 1992.

[4] [Ingeniería de Software 2001]

Autor. Roger S. Pressman, **Título.** Ingeniería de Software, Un enfoque práctico. **Versión.** 5ta Edición **Año.** 2001.