HEFESTO

DATA WAREHOUSING: Investigación y Sistematización de Conceptos

HEFESTO: Metodología para la Construcción de un Data Warehouse

Ing. Bernabeu Ricardo Dario

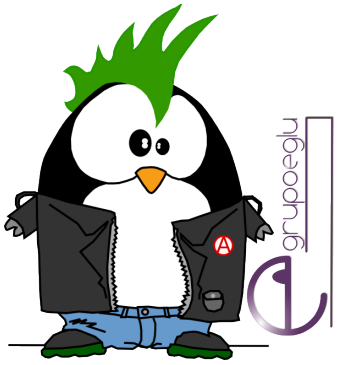
Córdoba, Argentina – Lunes 19 de Julio de 2010

Fecha	Versión	Autor/a	Detalle del cambio
Lunes 19 de Julio de 2010	2.0	Ing. Bernabeu Ricardo Dario	Actualización.
Lunes 31 de Agosto de 2009	1.2	Ing. Fernandez Carlos	Sección: Area de Datos.
Martes 21 de Abril de 2009	1.1	Ing. Bernabeu Ricardo Dario	Actualización.
Sábado 17 de Enero de 2009	1.0	Ing. Bernabeu Ricardo Dario	Actualización.
Miércoles 07 de Noviembre de 2007	0.1	Ing. Bernabeu Ricardo Dario	Versión Inicial.

...si supiese qué es lo que estoy haciendo,
no lo llamaría INVESTIGACIÓN...

Albert Einstein

Contacto:



- Blogs:
  - Personal: HEFESTO [http://tgx-hefesto.blogspot.com].
  - Dataprix: Dataprix [http://www.dataprix.com/blogs/bernabeu-dario].
- Mail: darioSistemas@gmail.com (poner en asunto [HEFESTO]).
- Red Social:
  - LinkedIn [http://www.linkedin.com/in/bernabeudario].
  - XING [https://www.xing.com/profile/Dario_Bernabeu].
  - Open Business Intelligence [http://www.redopenbi.com/profile/BernabeuRDario].
- Soluciones OSBI:
  - eGlu BI [http://www.eglubi.com.ar].
  - Mail: dbernabeu@grupoeglu.com.ar

A partir de la versión 2.0 de esta publicación, se han dejado de lado todos los términos que tienden a ”masculinizar” el lenguaje y en su lugar se ha optado por otra forma de expresión que es inclusiva para todos los géneros.

Por ejemplo, en vez de escribir ”los usuarios”, se utiliza ”l@s usuari@s”.

Índice general

Parte I

DATA WAREHOUSING: Investigación y Sistematización de Conceptos

RESUMEN

En esta primera parte de la publicación, 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 cuales 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.

Capítulo 1

BUSINESS INTELLIGENCE

1.1. Introducción

Actualmente, en las actividades diarias de cualquier organización, se generan datos como producto secundario, que son el resultado de todas las transacciones 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 esta publicación, es que estos dejen de solo ser simples datos, para convertirse en información que enriquezca las decisiones de l@s usuari@s.

Precisamente, la inteligencia de negocios (Business Intelligence - BI), permite que el proceso de toma de decisiones esté fundamentado sobre un amplio conocimiento de sí mismo y del entorno, minimizando de esta manera el riesgo y la incertidumbre.

Además, propicia que las organizaciones puedan traducir sus objetivos en indicadores de estudio, y que estos puedan ser analizados desde diferentes perspectivas, con el fin de encontrar información que no solo se encargue de responder a preguntas de lo que está sucediendo o ya sucedió, sino también, que posibilite la construcción de modelos, mediante los cuales se podrán predecir eventos futuros.

Cuando se nombra el término inteligencia, se refiere a la aplicación combinada de información, habilidad, experiencia y razonamientos, para resolver un problema de negocio.

Cabe destacar, que la aplicación de soluciones BI no es solo para grandes-medianas empresas, sino para quien desee tomar decisiones a través del análisis de sus datos. Es por ello que las soluciones BI no solo se enfocarán a resolver temas relacionados a: aumentar la rentabilidad, disminuir costos y obtener la famosa ventaja competitiva.

De acuerdo a lo planteado anteriormente se presentarán dos grandes ejemplos de la aplicación de BI, una en una empresa de ventas de productos, la otra en una biblioteca vecinal:

1. Empresa de venta de productos: en este caso la aplicación de BI podrá resolver las siguientes preguntas.
    - ¿Quiénes son l@s mejores client@s?.
    - ¿Cómo minimizar costos y maximizar las prestaciones?.

- ¿Cuál será el pronóstico de ventas del próximo mes?.

2. Biblioteca vecinal: en este caso la aplicación de BI podrá resolver las siguientes preguntas.
   - ¿Cuál es la temática más consultada?.
   - ¿Qué días hay mayor concurrencia, y por qué?.
   - ¿Qué libros deben ser adquiridos?.

1.2. Definición

Se puede describir BI, como un concepto que integra por un lado el almacenamiento y por el otro el procesamiento de grandes cantidades de datos, con el principal objetivo de transformarlos en conocimiento y en decisiones en tiempo real, a través de un sencillo análisis y exploración.

La definición antes expuesta puede representarse a través de la siguiente fórmula:

Datos + Análisis = **Conocimiento**

Este conocimiento debe ser oportuno, relevante, útil y debe estar adaptado al contexto de la organización.

Existe una frase muy popular acerca de BI, que dice: “Inteligencia de Negocios es el proceso de convertir datos en conocimiento y el conocimiento en acción, para la toma de decisiones”.

BI hace hincapié en los procesos de recolectar y utilizar efectivamente la información, con el fin de mejorar la forma de operar de una organización, brindando a sus usuari@s, el acceso a la información clave que necesitan para llevar a cabo sus tareas habituales y más precisamente, para poder tomar decisiones oportunas basadas en datos correctos y certeros.

Al contar con la información exacta y en tiempo real, es posible, aparte de lo ya mencionado, identificar y corregir situaciones antes de que se conviertan en problemas y en potenciales pérdidas de control de la empresa, pudiendo conseguir nuevas oportunidades o readaptarse frente a la ocurrencia de sucesos inesperados.

La Inteligencia de Negocios tiene sus raíces en los Sistemas de Información Ejecutiva[1] (Executive Information Systems – EIS) y en los Sistemas para la Toma de Decisiones[2] (Decision Support Systems – DSS), pero ha evolucionado y se ha transformado en todo un conjunto de tecnologías capaces de satisfacer a una gran gama de usuari@s junto a sus necesidades específicas en cuanto al análisis de información.

1.3. Proceso de BI

A fin de comprender cómo una organización puede crear inteligencia de sus datos, para, como ya se ha mencionado, proveer a l@s usuari@s finales oportuna y acertadamente acceso a esta información, se describirá a continuación el proceso de BI. El mismo

[1] Ver sección 3.6.5, en la página 70.

[2] Los DSS son una clase especial de sistemas de información cuyo objetivo es analizar datos de diferentes procedencias y brindar soporte para la toma de decisiones.

esta dividido en cinco fases, las cuales serán explicadas teniendo como referencia el siguiente gráfico, que sintetiza todo el proceso:

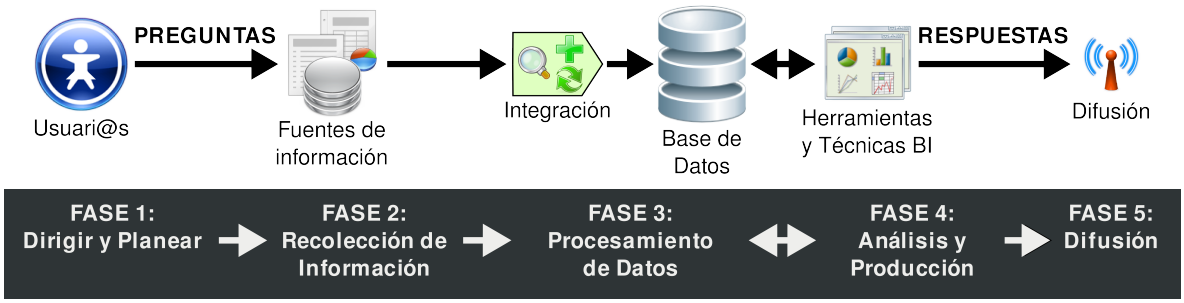


Figura 1.1: Fases del proceso BI.

- FASE 1: Dirigir y Planear. En esta fase inicial es donde se deberán recolectar los requerimientos de información específicos de l@s diferentes usuari@s, así como entender sus diversas necesidades, para que luego en conjunto con ell@s se generen las preguntas que les ayudarán a alcanzar sus objetivos.
- FASE 2: Recolección de Información. Es aquí en donde se realiza el proceso de extraer desde las diferentes fuentes de información de la empresa, tanto internas como externas, los datos que serán necesarios para encontrar las respuestas a las preguntas planteadas en el paso anterior.
- FASE 3: Procesamiento de Datos. En esta fase es donde se integran y cargan los datos en crudo en un formato utilizable para el análisis. Esta actividad puede realizarse mediante la creación de una nueva base de datos, agregando datos a una base de datos ya existente o bien consolidando la información.
- FASE 4: Análisis y Producción. Ahora, se procederá a trabajar sobre los datos extraídos e integrados, utilizando herramientas y técnicas propias de la tecnología BI, para crear inteligencia. Como resultado final de esta fase se obtendrán las respuestas a las preguntas, mediante la creación de reportes, indicadores de rendimiento, cuadros de mando, gráficos estadísticos, etc.
- FASE 5: Difusión. Finalmente, se les entregará a l@s usuari@s que lo requieran las herramientas necesarias, que les permitirán explorar los datos de manera sencilla e intuitiva.

1.4. Beneficios

Entre los beneficios más importantes que BI proporciona a las organizaciones, vale la pena destacar los siguientes:

- Reduce el tiempo mínimo que se requiere para recoger toda la información relevante de un tema en particular, ya que la misma se encontrará integrada en una fuente única de fácil acceso.
- Automatiza la asimilación de la información, debido a que la extracción y carga de los datos necesarios se realizará a través de procesos predefinidos.

- Proporciona herramientas de análisis para establecer comparaciones y tomar decisiones.
- Cierra el círculo que hace pasar de la decisión a la acción.
- Permite a l@s usuari@s no depender de reportes o informes programados, porque los mismos serán generados de manera dinámica.
- Posibilita la formulación y respuesta de preguntas que son claves para el desempeño de la organización.
- Permite acceder y analizar directamente los indicadores de éxito.
- Se pueden identificar cuáles son los factores que inciden en el buen o mal funcionamiento de la organización.
- Se podrán detectar situaciones fuera de lo normal.
- Permitirá predecir el comportamiento futuro con un alto porcentaje de certeza, basado en el entendimiento del pasado.
- L@s usuari@s podrán consultar y analizar los datos de manera sencilla e intuitiva.

Capítulo 2

DATA WAREHOUSING & DATA WAREHOUSE

2.1. Introducción

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 o base de datos que permita su posterior análisis y exploración, es imperativo y de vital importancia contar con un proceso que satisfaga todas estas necesidades. Este proceso se denomina Data Warehousing.

El Data Warehousing (DWH), es el encargado de extraer, transformar, consolidar, integrar y centralizar los datos que una organización genera en todos los ámbitos de su actividad diaria (compras, ventas, producción, etc) y/o información externa relacionada. Permitiendo de esta manera el acceso y exploración de la información requerida, a través de una amplia gama de posibilidades de análisis multivariables, con el objetivo final de dar soporte al proceso de toma de desiciones estratégico y táctico.

2.2. Definición

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 sumarizada, 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 l@s usuari@s indicad@s 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.

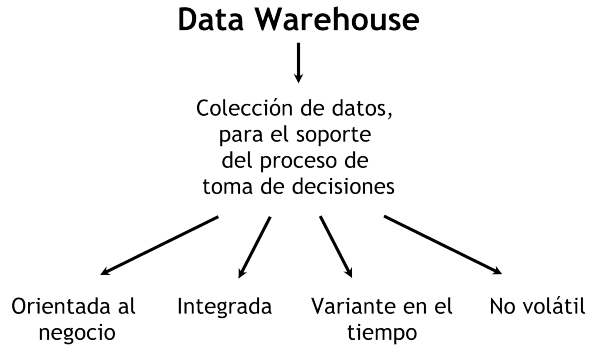


Figura 2.1: Data Warehouse, características.

Cabe aclarar que los términos almacén de datos y depósito de datos, son análogos a DW, y se utilizarán de aquí en adelante para referirse al mismo.

2.3. Características

2.3.1. 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 organización. 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 a l@s client@s, 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, 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 empresa minorista manejará: stock, lista de precios, cuentas corrientes, pagos diferidos, impuestos, retenciones, ventas, notas de crédito, compras, etc. De esta manera, la base de datos combinará estos elementos en una estructura que se adapte a sus necesidades.

En contraposición, siguiendo con el ejemplo anterior, en una empresa minorista el ambiente DW se organizará alrededor de entidades de alto nivel tales como: clientes, productos, rubros, proveedores, vendedores, zonas, etc. Que son precisamente aquellos sujetos mediante los cuales se desea analizar la información. Esto se debe a que el depósito de datos se diseña para realizar consultas e investigaciones sobre las actividades de la organización y no para soportar los procesos que se realizan en ella.

En síntesis, la ventaja de contar con procesos orientados a la aplicación, esta fundamentada en la alta accesibilidad de los datos, lo que implica un elevado desempeño y velocidad en la ejecución de consultas, ya que las mismas están predeterminadas; mientras que en el DW para satisfacer esta ventaja se requiere que la información este desnormalizada, es decir, con redundancia[1] y que la misma esté dimensionada, para evitar tener que recorrer toda la base de datos cuando se necesite realizar algún análisis determinado, sino que simplemente la consulta sea enfocada por variables de análisis que permitan localizar los datos de manera rápida y eficaz, para poder de esta manera satisfacer una alta demanda de complejos exámenes en un mínimo tiempo de respuesta.

2.3.2. Integrada

La integración implica que todos los datos de diversas fuentes que son producidos por distintos departamentos, secciones y aplicaciones, tanto internos como externos, deben ser consolidados en una instancia antes de ser agregados al DW, y deben por lo tanto ser analizados para asegurar su calidad y limpieza, entre otras cosas. A este proceso se lo conoce como Integración de Datos, y cuenta con diversas técnicas y subprocesos para llevar a cabo sus tareas. Una de estas técnicas son los procesos ETL: Extracción, Transformación y Carga de Datos[2] (Extraction, Transformation and Load).

Si bien el proceso ETL es 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 DW.

La integración de datos, resuelve diferentes 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.

La causa de dichos problemas, se debe principalmente a que a través de los años l@s diseñador@s y programador@s no se han basado en ningún estándar concreto para

[1] Ver sección 2.7, en la página 16.
[2] Ver sección 3.3, en la página 21.

definir nombres de variables, tipos de datos, etc., ya sea por carecer de ellos o por no creer que sean necesarios. Por lo cual, cada uno por su parte ha dejado en cada aplicación, módulo, tabla, etc., su propio estilo personalizado, confluyendo de esta manera en la creación de modelos muy inconsistentes e incompatibles entre sí.

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

2.3.3. 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:

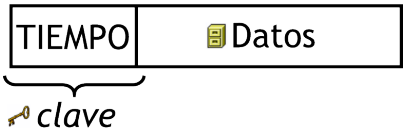


Figura 2.2: Data Warehouse, variante en el tiempo.

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

Es importante tener en cuenta la granularidad[3] de los datos, así como también la intensidad de cambio natural del comportamiento de los fenómenos de la actividad que se desarrolle, para evitar crecimientos incontrolables y desbordamientos de la base de datos.

El intervalo de tiempo y periodicidad de los datos debe definirse de acuerdo a la necesidad y requisitos de l@s usuari@s.

[3] Ver sección 3.4.4.5, en la página 37.

Es elemental aclarar, que el almacenamiento de datos históricos, es lo que permite al DW desarrollar pronósticos y análisis de tendencias y patrones, a partir de una base estadística de información.

2.3.4. No volátil

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

La actualización, o sea, insertar, eliminar y modificar, se hace de forma muy habitual en el ambiente operacional sobre una base, registro por registro, en cambio en el 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 concurrencia y recuperación.

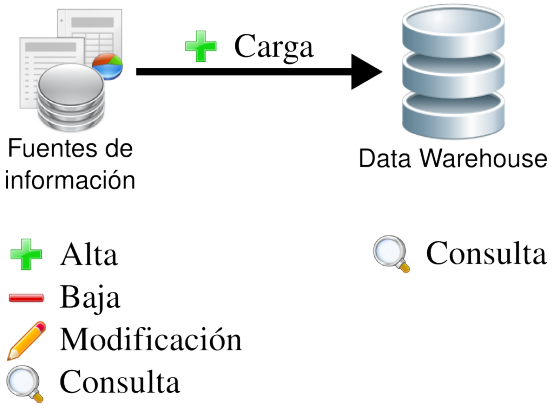


Figura 2.3: Data Warehouse, no volátil.

2.4. 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 y áreas, en un solo lugar centralizado. Es por esta razón que el depósito puede ser soportado y mantenido sobre diversos medios de almacenamiento.

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

Organiza y almacena los datos que se necesitan para realizar consultas y procesos analíticos, con el propósito de responder a preguntas complejas y brindarles a l@s usuari@s finales la posibilidad de que mediante una interface amigable, intuitiva y fácil de utilizar, puedan tomar decisiones sobre los datos sin tener que poseer demasiados conocimientos informáticos. El DW permite un acceso más directo, es decir, la información

gira en torno al negocio, y es por ello que también l@s usuari@s pueden sentirse cómod@s al explorar los datos y encontrar relaciones complejas entre los mismos.

Cabe aclarar que el Data Warehousing no se compone solo de datos, ni tampoco solo se trata de un depósito de datos aislado. El Data Warehousing hace referencia a 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 que son empleadas, 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 Datos[4] (SGBD).

  Todas las cualidades expuestas anteriormente, son imposibles de saldar en un típico ambiente operacional, y esto es una de las razones de ser del Data Warehousing.

2.5. Ventajas

A continuación se enumerarán algunas de las ventajas más sobresalientes que trae aparejada la implementación de un Data Warehousing 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 (internas y/o externas) 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 l@s usuari@s necesitan, en el momento adecuado y en el formato apropiado.

[4] Ver sección 4.3, en la página 75.

- Logra un impacto positivo sobre los procesos de toma de decisiones. Cuando l@s usuari@s tienen acceso a una mejor calidad de información, la empresa puede lograr por sí misma: aprovechar el enorme valor potencial de sus recursos de información y transformarlo en valor verdadero; eliminar los retardos de los procesos que resultan de información incorrecta, inconsistente y/o inexistente; integrar y optimizar procesos a través del uso compartido e integrado de las fuentes de información; permitir a l@s usuari@s adquirir mayor confianza acerca de sus propias decisiones y de las del resto, y lograr así, un mayor entendimiento de los impactos ocasionados.

- Aumento de la eficiencia de l@s encargad@s de tomar decisiones.

- L@s usuari@s pueden acceder directamente a la información en línea, lo que contribuye a su capacidad para operar con mayor efectividad en las tareas rutinarias o no. Además, pueden tener a su disposición una gran cantidad de valiosa información multidimensional, presentada coherentemente como fuente única, confiable y disponible en sus estaciones de trabajo. Así mismo, l@s usuari@s tienen la facilidad de contar con herramientas que les son familiares para manipular y evaluar la información obtenida en el DW, tales como: hojas de cálculo, procesadores de texto, software de análisis de datos, software de análisis estadístico, reportes, tableros, etc.

- Permite la toma de decisiones estratégicas y tácticas.

2.6. Desventajas

A continuación se enumerarán algunas de las desventajas más comunes que se pueden presentar en la implementación de un Data Warehousing:

- 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 l@s usuari@s.

- Existe resistencia al cambio por parte de l@s usuari@s.

- 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 tod@s l@s usuari@s 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 usuari@ podrá tener acceso a ellos.

- Infravaloración de los recursos necesarios para la captura, carga y almacenamiento de los datos.

- Infravaloración del esfuerzo necesario para su diseño y creación.

- Incremento continuo de los requerimientos de l@s usuari@s.

- Subestimación de las capacidades que puede brindar la correcta utilización del DWH y de las herramientas de BI en general.

2.7. 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.

- Los datos experimentan una considerable transformación, antes de ser cargados al DW. La mayor parte de los datos se alteran significativamente al ser seleccionados, consolidados y movidos al depósito.

En vista de estos factores, se puede afirmar que, la redundancia encontrada al cotejar los datos de ambos ambientes es mínima, ya que generalmente resulta en un porcentaje menor del 1 %.

2.8. Estructura

Los DW estructuran los datos de manera muy particular y existen diferentes niveles de esquematización y detalle que los delimitan.

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

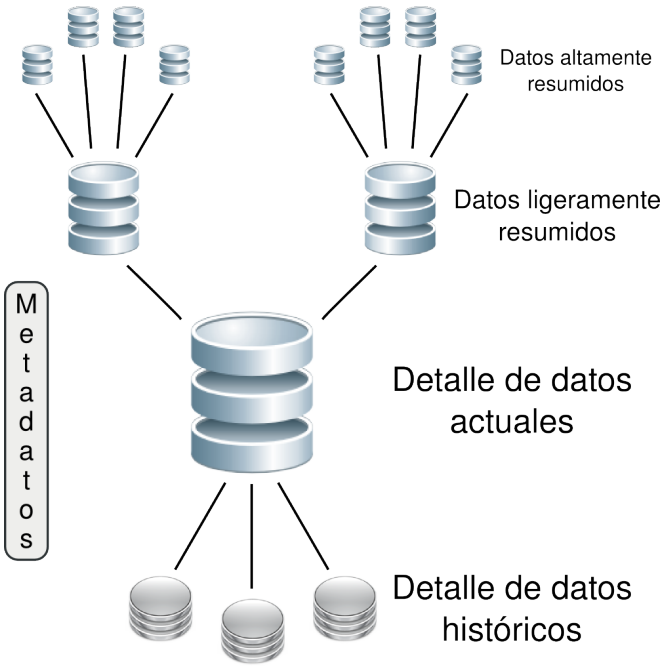


Figura 2.4: Data Warehouse, estructura.

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 o granularidad 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[5]: 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.

Estos diferentes niveles de detalle o granularidad, se obtienen a través de tablas de hechos agregadas y/o preagregadas[6].

2.9. Flujo de Datos

El DW posee un flujo de datos estándar y generalizado, el cual puede apreciarse mejor en la siguiente figura.

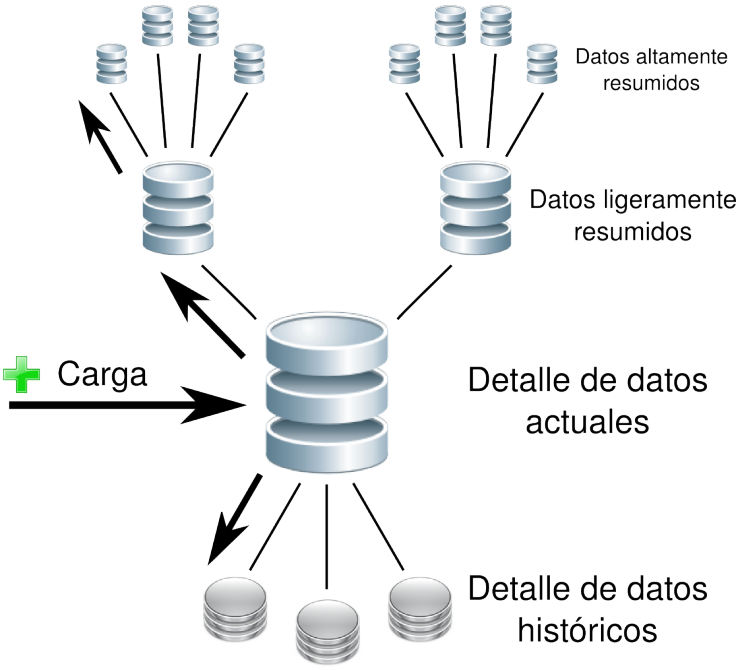


Figura 2.5: Data Warehouse, flujo de datos.

Cuando la información ingresa al depósito de datos se almacena a nivel de Detalle de datos actuales. Los datos permanecerán allí hasta que ocurra alguno de los tres eventos siguientes:

- Sean borrados del depósito de datos.
- Sean resumidos, ya sea a nivel de Datos ligeramente resumidos o a nivel de Datos altamente resumidos.
- Sean archivados a nivel de Detalle de datos históricos.

[5]Ver sección 3.4.9, en la página 49.
[6]Ver sección 3.4.3.1, en la página 32.

Capítulo 3

ARQUITECTURA DEL DATA WAREHOUSING

3.1. Introducción

En este punto y teniendo en cuenta que ya se han detallado claramente las características generales del Data Warehousing, se definirán y describirán todos los componentes que intervienen en su arquitectura o ambiente.

A través del siguiente gráfico se explicitará la estructura del Data Warehousing:

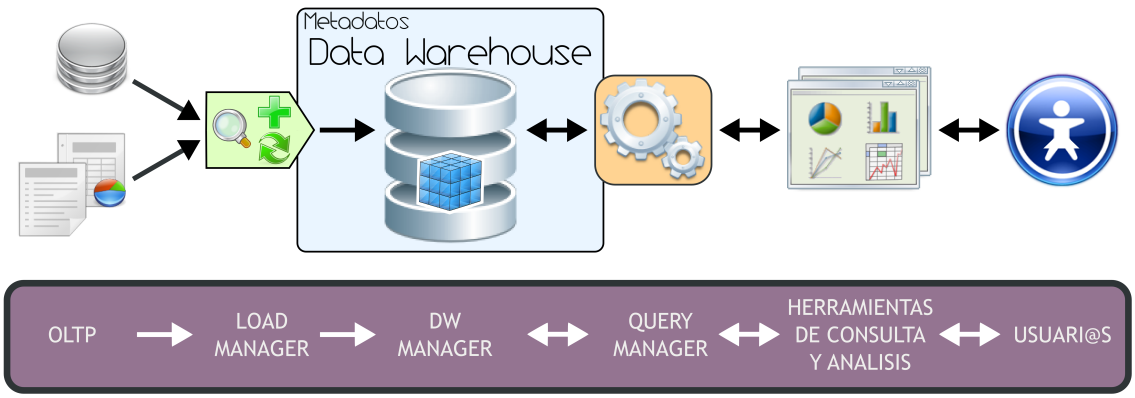


Figura 3.1: Data Warehousing, arquitectura.

Tal y como se puede apreciar, el ambiente esta formado por diversos elementos que interactúan entre sí y que cumplen una función específica dentro del sistema. Por ello es que al abordar la exposición de cada elemento se lo hará en forma ordenada y teniendo en cuenta su relación con las demás partes.

Básicamente, la forma de operar del esquema superior se resume de la siguiente manera:

- Los datos son extraídos desde aplicaciones, bases de datos, archivos, etc. Esta información generalmente reside en diferentes tipos de sistemas, orígenes y arqui-

tecturas y tienen formatos muy variados.

- Los datos son integrados, transformados y limpiados, para luego ser cargados en el DW.
- Principalmente, la información del DW se estructura en cubos multidimensionales, ya que estos preparan esta información para responder a consultas dinámicas con una buena performance. Pero también pueden utilizarse otros tipos de estructuras de datos para representar la información del DW, como por ejemplo Business Models.
- L@s usuari@s acceden a los cubos multidimensionales, Business Models (u otro tipo de estructura de datos) del DW utilizando diversas herramientas de consulta, exploración, análisis, reportes, etc.

A continuación se detallará cada uno de los componentes de la arquitectura del Data Warehousing, teniendo como referencia siempre el gráfico antes expuesto, pero resaltando el tema que se tratará.

3.2. OLTP

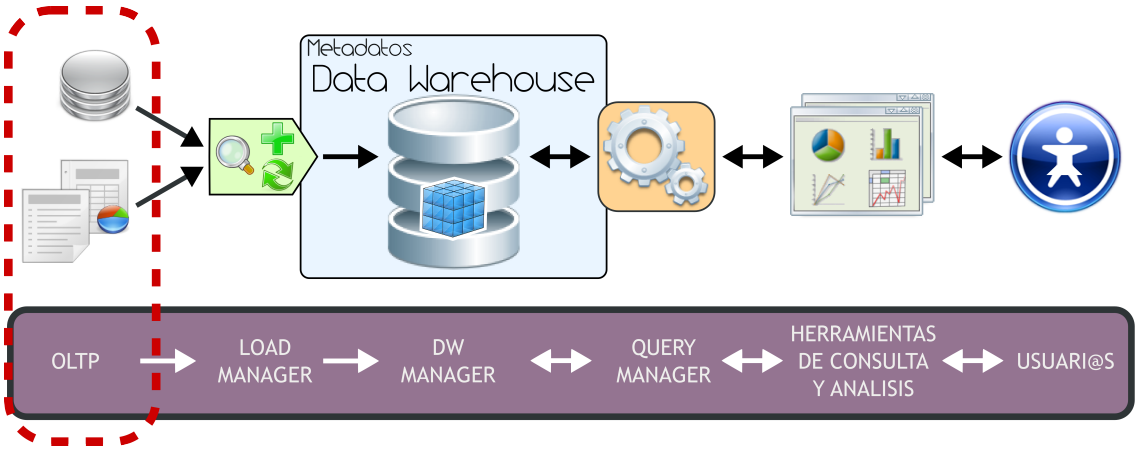


Figura 3.2: OLTP.

OLTP (On Line Transaction Processing), representa toda aquella información transaccional que genera la empresa en su accionar diario, además, de las fuentes externas con las que puede llegar a disponer.

Como ya se ha mencionado, estas fuentes de información, son de características muy disímiles entre sí, en formato, procedencia, función, etc.

Entre los OLTP más habituales que pueden existir en cualquier organización se encuentran:

- Archivos de textos.
- Hipertextos.

- Hojas de cálculos.
- Informes semanales, mensuales, anuales, etc.
- Bases de datos transaccionales.

3.3. Load Manager

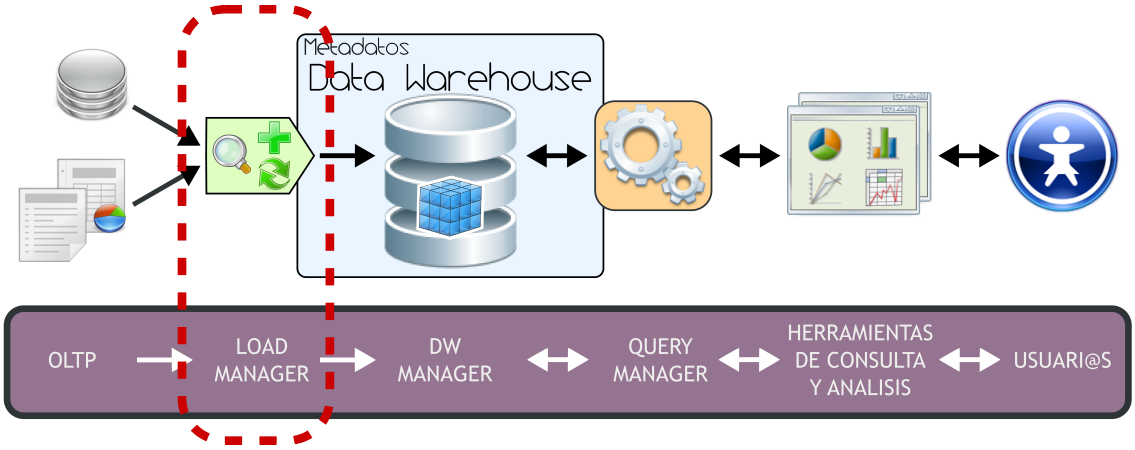


Figura 3.3: Load Manager.

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, la Integración de Datos en 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 DW, 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 DW.

Como se mencionó anteriormente cuando se trataron las características del DW[1], 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 DW.

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.

[1]Ver sección 2.3.2, en la página 11.

3.3.1. Extracción

Es aquí, en donde, basándose en las necesidades y requisitos de l@s usuari@s, 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 DW.

3.3.2. Transformación

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.

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.

3.3.2.1. Codificación

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”, algun@s diseñador@s 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 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.

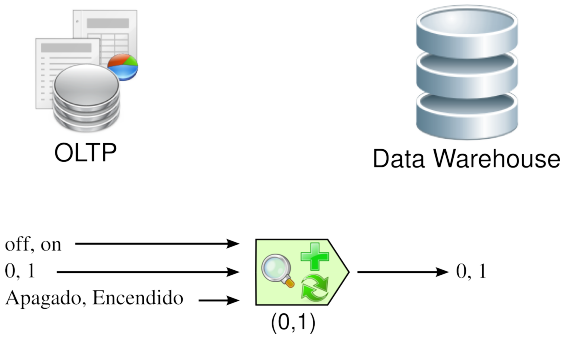


Figura 3.4: Transformación: codificación.

3.3.2.2. 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 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.

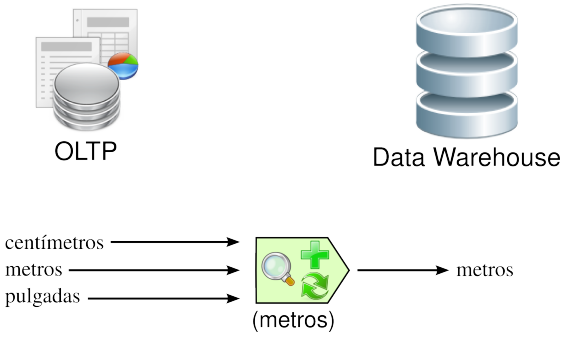


Figura 3.5: Transformación: medida de atributos.

3.3.2.3. 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 l@s usuari@s sea más comprensible.

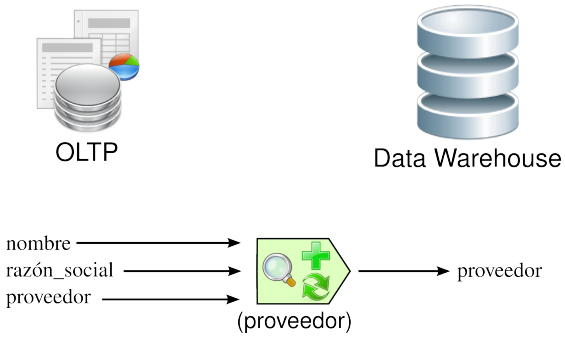


Figura 3.6: Transformación: convenciones de nombramiento.

3.3.2.4. Fuentes múltiples

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

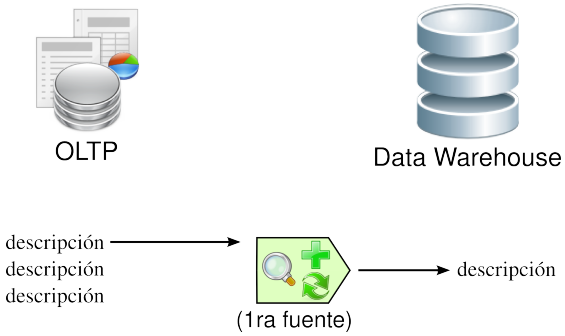


Figura 3.7: Transformación: fuentes múltiples.

3.3.2.5. 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. Se puede dar que en algunos casos, los valores faltantes sean inexistentes, ya que por ejemplo, l@s nuev@ asociad@s o client@s, no poseerán consumo medio del último año.

3.3.3. Carga

1. 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 DW. 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 DW y los requerimientos de l@s usuari@s. 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 DW. 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 DW 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.

2. Por otra parte, el proceso de Carga tiene la tarea de mantener la estructura del DW, y trata temas relacionados con:
   - Relaciones muchos a muchos[2].
   - Claves Subrogadas[3].
   - Dimensiones Lentamente Cambiantes[4].
   - Dimensiones Degeneradas[5].

3.3.4. Proceso ETL

A continuación, se explicará en síntesis el accionar del proceso ETL, y cuál es la relación existente entre sus diversas funciones. En la siguiente figura se puede apreciar mejor lo antes descrito:

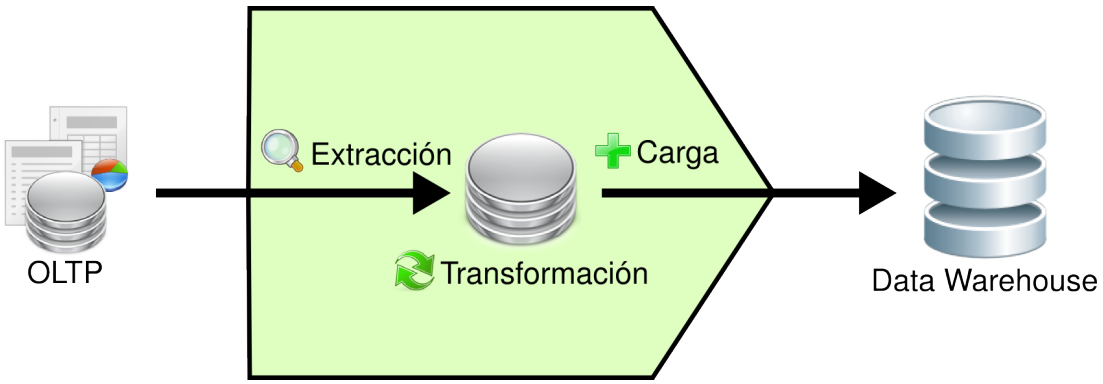


Figura 3.8: Proceso ETL.

Los pasos que se siguen son:

[2]Ver sección 6.12, en la página 123.
[3]Ver sección 6.13, en la página 124.
[4]Ver sección 6.14, en la página 125.
[5]Ver sección 6.15, en la página 129.

- Se extraen los datos relevantes desde los OLTP y se depositan en un almacenamiento intermedio.
- Se integran y transforman los datos, para evitar inconsistencias.
- Se cargan los datos desde el almacenamiento intermedio hasta el DW.

3.4. Data Warehouse Manager

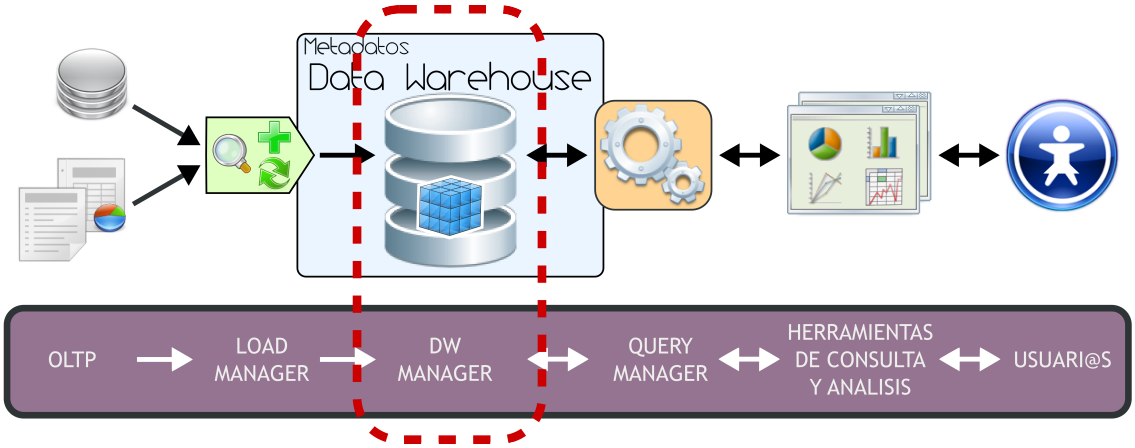


Figura 3.9: Data Warehouse Manager.

El DW Manager presenta las siguientes características y funciones principales:

- Se constituye típicamente al combinar un SGBD con software y aplicaciones dedicadas.
- Almacena los datos de forma multidimensional[6], es decir, a través de tablas de hechos[7] y tablas de dimensiones[8].
- Gestiona las diferentes estructuras de datos que se construyan o describan sobre el DW, como Cubos Multidimensionales[9], Business Models[10], etc.
- Gestiona y mantiene metadatos.

Además, el DW Manager se encarga de:

- Transformar e integrar los datos fuentes y del almacenamiento intermedio en un modelo adecuado para la toma de decisiones.
- Realizar todas las funciones de definición y manipulación del depósito de datos, para poder soportar todos los procesos de gestión del mismo.

[6] Ver sección 3.4.1, en la página 28.
[7] Ver sección 3.4.3, en la página 30.
[8] Ver sección 3.4.2, en la página 28.
[9] Ver sección 3.4.4, en la página 33.
[10] Ver sección 4.5, en la página 76.

- Ejecutar y definir las políticas de particionamiento[11]. El objetivo de realizar esto, es conseguir una mayor eficiencia y performance en las consultas al no tener que manejar todo el grueso de los datos. Esta política debe aplicarse sobre la tabla de hechos que, como se explicará más adelante, es en la que se almacena toda la información que será analizada.
- Realizar copias de resguardo incrementales o totales de los datos del DW.

3.4.1. Base de datos multidimensional

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[12], Business Model[13], 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[14] (Star Scheme).
- Esquema copo de nieve[15] (Snowflake Scheme).
- Esquema constelación[16] o copo de estrellas (Starflake Scheme).

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

- Relacional – ROLAP[17].
- Multidimensional – MOLAP[18].
- Híbrido – HOLAP[19].

3.4.2. Tablas de Dimensiones

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

Representan los aspectos de interés, mediante los cuales l@s usuari@s podrán filtrar y manipular la información almacenada en la tabla de hechos.

En la siguiente figura se pueden apreciar algunos ejemplos:

[11] Ver sección 4.4, en la página 76.
[12] Ver sección 3.4.4, en la página 33.
[13] Ver sección 4.5, en la página 76.
[14] Ver sección 3.4.5.1, en la página 37.
[15] Ver sección 3.4.5.2, en la página 39.
[16] Ver sección 3.4.5.3, en la página 40.
[17] Ver sección 3.4.7.1, en la página 42.
[18] Ver sección 3.4.7.2, en la página 43.
[19] Ver sección 3.4.7.3, en la página 44.