Universidad Central "Martha Abreu" de Las Villas

Facultad "Matemática-Física-Computación"

Licenciatura en Ciencias de la Computación

Trabajo de Diploma

Título: Almacén de Datos para la Toma de Decisiones en Postgrado versión 1.

Autor: Merly Arrizabalaga Martínez.

Tutor: Dr. Rosendo Moreno Rodríguez.

Seminario: Base Datos.

Santa Clara, junio 2011

"Año 53 de la Revolución"

Dictamen.

El que suscribe, _____
_____, hago constar que el trabajo titulado _____
_____
_____ fue realizado en la Universidad Central "Marta Abreu" de Las Villas como parte de la culminación de los estudios de la especialidad de _____, autorizando a que el mismo sea utilizado por la institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos ni publicado sin la autorización de la Universidad.

_____
Firma del autor

Los abajo firmantes, certificamos que el presente trabajo ha sido realizado según acuerdos de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

_____ _____
Firma del tutor Firma del jefe del Laboratorio

_____
Fecha

Pensamiento

Dedicatoria

A mi mamá por estar siempre apoyándome
Por ser la mejor mamá del mundo.

A mi esposo Jorge Antonio por ser la persona más especial de mi vida y por haber confiado en mi todos estos años.

A mi hermana Marla.

Agradecimientos

Resumen

El presente trabajo consiste en el diseño e implementación de un almacén de datos que partiendo de la información histórica asociada a la actividad de postgrado, permita el almacenamiento de la misma en formato dimensional, basado en Software Libre y compatible con los requerimientos del proyecto SIGENU. Además el trabajo refiere al desarrollo del módulo ETL correspondiente que permita transformar y almacenar la información histórica del SGP. Este almacén garantizará mayor disponibilidad y calidad de la información y la realización de análisis dinámicos de la misma, facilitando así la toma de decisiones estratégicas de la Dirección de Postgrados de la UCLV.

Abstract

This work involves designing and implementing a data warehouse based on historical information associated with the postgraduate activity, allows the storage of the same dimensional format based on Free Software and supports SIGENU project requirements. Further work concerns the development of the module that allows for ETL process and store the historical information of the GSP. This store will ensure greater availability and quality of information and performing dynamic analysis of the same, thus facilitating the strategic decisions of the Department of Graduate Studies of the UCLV.

Tabla de contenido

Tabla de Figura

Introducción

Antecedentes

Desde hace varios años, la Universidad Central “Marta Abreu” de las Villas participa en el proyecto SIGENU con la responsabilidad de la implantación de un Sistema de Control de Postgrado (SPG) compatible con la plataforma SIGENU y los nuevos lineamientos de nuestro país en el uso de Software Libre. Para incrementar su eficiencia, este sistema debe proporcionar a los usuarios información analítica y estratégica, esto se logra al aprovechar la información que a diario es almacenada en sus bases de datos operativas. Al intentar utilizar esta información de las bases de datos operativas para tomar decisiones, se presentan varios problemas: existe demasiada información almacenada ya con carácter histórico, muy genérica de la cual no se pueden sacar conclusiones fácilmente. La información muchas veces es irrelevante para el área interesada en mejorar sus decisiones, y la organización termina por desaprovechar todos estos datos, perdiendo un proceso de aprendizaje de sus propios logros e información.

La educación postgraduada es uno de los objetivos esenciales de la Educación Superior. Se conoce comúnmente como el cuarto nivel. Se ocupa desde la superación profesional que cada graduado universitario debe recibir a partir de su titulación universitaria, para incrementar sus aptitudes científico-técnicas, sociales y políticas; hasta la especialización académica a través de programas avanzados de maestrías, especialidades y doctorados. Hasta los momentos actuales se ha cumplido con una labor de educación de postgrado intensiva y extensiva, incluyendo programas académicos denominados de “amplio acceso”; pero las nuevas tendencias político-sociales y económicas del país permiten afirmar que se debe restringir en algunos casos y de seleccionar en otros el acceso y los resultados de este nivel de educación, sobre todo hacerlos más acordes a las verdaderas necesidades del país. Para tomar decisiones adecuadas es necesario hacer análisis de tendencias, buscar información

no revelada en los sistemas operacionales, resumir datos hasta ciertos niveles por diversos criterios y tener la posibilidad de brindar informes diversos pero muy concretos tanto de forma textual como gráfica, pero siempre partiendo de criterios ad hoc, no pre-analizados.

El Sistema de Control de Postgrado (SPG) es de gran utilidad, ya que mediante el mismo se controla todo lo que está vinculado con las acciones de postgrado, pero las versiones anteriores en este momento no satisfacen ni los requerimientos técnicos del Ministerio de Educación Superior (MES) para este tipo de sistemas, ni otras derivadas de actualizaciones al Reglamento de la Educación Postgraduada de la República de Cuba No 132-2004 y sus últimas adecuaciones hechas en la Resolución No. 166/09. En estos momentos y como parte de otras tesis de diploma se acomete el desarrollo de un nuevo sistema de información para el control de esa actividad, denominado Sistema de Postgrado versión 5.4. No obstante, dado que ese sistema tiende a almacenar datos y brindar información operativa sobre la actividad (que incluye planificación de programas y ediciones de postgrado, control de estudiantes, sus matrículas y sus resultados, control de profesores, etc.) y está regido por el análisis sistémico desarrollado por expertos de la Dirección de Postgrado de la UCLV, no incluye posibilidades de desarrollo de análisis para la toma de decisiones que tomen en consideración otros aspectos no revelados. De aquí la importancia de desarrollar un módulo independiente que denominamos Almacén de Datos para la Toma de Decisiones en Postgrado.

Planteamiento del problema

Se desea crear un almacén de datos que partiendo de la información histórica asociada al Sistema de Gestión de Postgrado, permita el almacenamiento de la misma en formato dimensional y a partir de allí posibilite generar informes ad hoc que brinden tendencias del aprovechamiento académico de graduados universitarios, instituciones de la producción y los servicios, etc., y posibilite valorar la eficiencia y productividad asociada a las actividades de postgrado. Este debe garantizar mayor disponibilidad y calidad de la información, así como la realización de análisis dinámicos de la misma, facilitando de esta forma la toma de decisiones.

Objetivos General

Diseñar e implementar un almacén de datos y los módulos ETL correspondientes que permita transformar y almacenar la información histórica del SGP, en un modelo

dimensional, utilizando Software Libre, de manera que sea compatible con los requerimientos del proyecto SIGENU.

Objetivos específicos

1. Realizar un estudio detallado de las necesidades de análisis dinámicos asociados a la actividad de postgrado, desde el punto de vista de la toma de decisiones a nivel universitario.
2. Diseñar una estructura dimensional para un almacén de datos asociado con el postgrado, e implementar la misma en un SGBD de software libre.
3. Desarrollar las transformaciones de datos necesarias para cargar los datos históricos desde la base de datos operacional en el almacén de datos.

Pregunta de investigación

1. ¿Cuáles datos son necesarios guardar para conformar un registro histórico relevante de la actividad de Postgrado?
2. ¿Qué tipo de formato dimensional será más adecuado para el almacén de datos asociado a postgrado?
3. ¿Cómo implementar un almacén de datos utilizando software libre que permita guardar y gestionar el modelo de registro histórico del SGP?
4. ¿Cómo implementar los programas que posibiliten la carga, limpieza y transformación de los datos desde el sistema operacional al almacén de datos?

Justificación de la investigación

Desde hace más de 20 años en la Universidad Central de Las Villas se ha desarrollado y explotado un Sistema para el control de Postgrado a nivel de áreas que desarrollan esta actividad (Facultades y Centros de Investigación o de Estudios). En estos momentos la última versión desarrollada por el tutor de este trabajo sobre el SGBD Access, adolece de la actualización que permita cumplir con todo lo establecido en el actual Reglamento de Postgrado de la República de Cuba. Por ejemplo, no se obtiene la interrelación entre Maestrías y Diplomados que lo conforman, además de que faltan los reportes correspondientes a las adecuaciones que desde el 2004 (fecha de la última actualización de la versión 4.3) a la fecha se han realizado en las estadísticas nacionales oficiales.

El Ministerio de Educación Superior, ha encomendado a varios centros el desarrollo de varios sistemas de control de las actividades de las Universidades, bajo el proyecto SIGENU. Actualmente existe el Sistema de Control Docente o Académico

(control de pregrado), aplicado en diferentes Universidades del país. El MES dio a la UCLV la responsabilidad hace cuatro años del desarrollo del Sistema de control de Postgrado, dentro del marco del proyecto SIGENU.

Esta tarea la acomete oficialmente el Departamento de Producción de Software (DPS) de la UCLV, pero por diversos motivos han dado al traste con la implementación adecuada del mismo, el que debe hacerse sobre un conjunto de tecnologías de Software Libre especificadas en las disposiciones del MES.

Un módulo para la gestión del registro histórico de postgrado (RHP), para un almacenamiento de la historia de los cursos terminados, posibilita enfrentar los constantes y necesarios cambios en la gestión de los cursos en marcha, cambios que ahora afectarían solo a un reducido número de datos, sin tener que cambiar toda una historia del postgrado en Cuba, mientras el RHP almacenara los datos con interés histórico, datos más estables, que no han mantenido su formato por muchos años. Además utilizar un RPH, permitirá a otros centros del MES que estén utilizando otros Sistemas de Gestión de Postgrado, almacenar los cursos cerrados y así estar integrados a un sistema de gestión nacional, de donde se sacaran reportes históricos y certificaciones de los postgrados terminados.

Por todo lo anteriormente planteado se hizo imprescindible emprender el desarrollo de un módulo para el registro histórico del SGP.

Viabilidad

La experiencia del tutor en el trabajo de control de Postgrado durante más de 20 años lo avala como un experto fiable para obtener un modelo correcto de la gestión de Postgrado. La disponibilidad de una plataforma basada en software libre, para la carga y transformación de datos, compatible con las tecnologías utilizadas en la plataforma de desarrollo, facilitan el desarrollo y prueba de las transformaciones necesarias para crear y cargar el nuevo almacén de datos.

También se cuenta con bibliografía actualizada para estudiar las herramientas de programación propuestas y la teoría de bases de datos y sistemas de información y con el equipamiento mínimo necesario para desarrollar el sistema.

Tipo de Investigación

Esta investigación es de tipo exploratoria.

Hipótesis de investigación

La experiencia acumulada durante el desarrollo de las versiones anteriores de SPG, permiten identificar un conjunto de datos invariantes que, más allá de su formato, son

de interés estadístico para el Sistema de Control Postgrado. Estas entidades de carácter histórico se pueden modelar utilizando un almacén de datos que se puede implementar utilizando herramientas de Software Libre y pueden ser utilizados de forma integrada con la Bases de Datos Relacionales PostgreSQL del SGP en la transformación y carga de su registro de datos históricos.

Organización de la Tesis

La tesis está organizada en tres capítulos que se refieren a los siguientes aspectos:

Capítulo 1: Marco Teórico referente a Almacenes de Datos, Modelado Dimensional, tecnologías de ETL (carga, limpieza y transformación), explotación de un almacén de datos, desarrollo de técnicas estadísticas avanzadas de uso del almacén.

Capítulo 2: Descripción amplia de necesidades de información adicionales en la actividad de postgrado y diseño del modelo dimensional (estrella o constelación) del almacén propuesto para la actividad de postgrado.

Capítulo 3: Descripción de la implementación de la estructura del almacén de datos sobre un SGBD de software libre y las transformaciones necesarias para realizar el proceso de ETL

Capítulo I. Marco Teórico.

Introducción

En este capítulo se muestran los conceptos básicos sobre almacenes de datos así como los procesos que intervienen en la creación y uso de un AD. En él, se hace una introducción al modelado multidimensional y sus características principales.

Para llegar a lo que es un almacén de datos o sea su concepto es necesario primeramente saber que es un Sistema de Información y dentro de cuál de las clasificaciones de este entran los almacenes de datos.

1.1- Sistema de Información

Existen diversas definiciones de lo que es un Sistema de Información una de la más completa es:

"Un conjunto formal de procesos que, operando sobre una colección de datos estructurada según las necesidades de la empresa, recopilan, elaboran y distribuyen la información (o parte de ella) necesaria para las operaciones de dicha empresa y para las actividades de dirección y control correspondientes (decisiones) para desempeñar su actividad de acuerdo a su estrategia de negocio".(Gloria Ponjuan, 2004)

Los SI se pueden clasificar desde un punto de vista empresarial, la primera clasificación se basa en la jerarquía de una organización y se llamó el modelo de la pirámide (Figura 1). Según la función a la que vayan destinados o el tipo de usuario final del mismo, los SI más extendidos en la actualidad según (Laudon, Jane, & Kenneth, 2006).

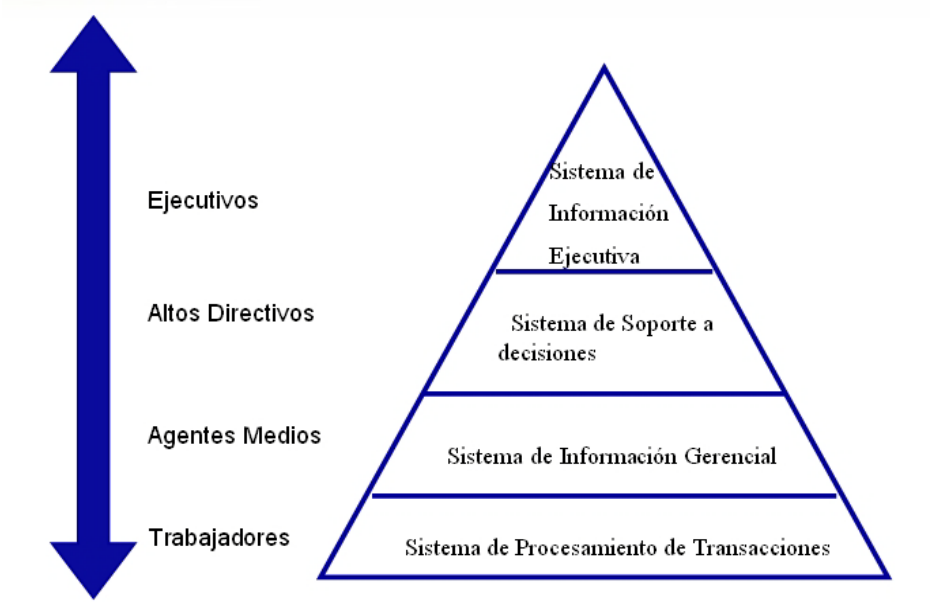

Figura 1: Modelo de la Pirámide

1. **Sistema de procesamiento de transacciones** (**TPS**) – también son llamados Sistema de procesamiento de transacciones en línea (**OLTP**). Gestiona la información referente a las transacciones producidas en una empresa u organización. Son aquellos que procesan y ejecutan tareas cotidianas de la organización, necesarias para su gestión. Son altamente estructurados y predefinidos. Característicos de los años 50 y 60 e imprescindibles en los 90. Ejemplos de **OLTP**: Sistema de reservación de líneas aéreas, Sistemas de nómina y contabilidad e inventario.

2. **Sistemas de información gerencial** (**MIS**).- Orientados a solucionar problemas empresariales en general. Son los sistemas informativos que le sirven al nivel de dirección para funciones de planeamiento, control y toma de decisiones. Típicamente estos sistemas son casi exclusivos para eventos internos de la organización. Ejemplos: Sistemas para objetivos de planificación propio del máximo nivel de dirección.

3. **Sistemas de soporte a decisiones** (**DSS**).- Herramienta para realizar el análisis de las diferentes variables de negocio con la finalidad de apoyar el proceso de toma de decisiones. Ejemplo: Sistemas de análisis de factibilidad de una inversión cooperada.

4. **Sistemas de información ejecutiva** (**EIS**).- Herramienta orientada a usuarios de nivel gerencial, que permite monitorizar el estado de las variables de un área o unidad de la empresa a partir de información interna y externa a la misma.

A continuación se realizara un breve panorama en los sistemas de información **DSS** debido a que es aceptado que estos sistemas son la base de un A*lmacén de Datos.*

1.1.1-Sistemas de Soporte a Decisiones

Debido a que hay muchos enfoques para la toma de decisiones y debido a la amplia gama de ámbitos en los cuales se toman las decisiones, el concepto de **sistema de apoyo a las decisiones** (**DSS** por sus siglas en inglés *Decisionsupportsystem*) es muy amplio. Un **DSS** puede adoptar muchas formas diferentes. Algunas de las definiciones de **DSS** son:

Un **DSS**, en términos muy generales, es "*un sistema basado en computador que ayuda en el proceso de toma de decisiones*" (Finlay, 1994) *y otros*).

En términos bastante más específicos, un **DSS** es *"un sistema de información basado en un computador interactivo, flexible y adaptable, especialmente desarrollado para apoyar la solución de un problema de gestión no estructurado para mejorar la toma de decisiones. Utiliza datos, proporciona una interfaz amigable y permite la toma de decisiones en el propio análisis de la situación"* (Turban, 1995).

En general, podemos decir que un **DSS** asiste a los usuarios en el proceso de análisis de datos en una organización con el propósito de producir información que les permita tomar mejores decisiones. Los analistas que utilizan el **DSS** están más interesados en identificar tendencias que en buscar algún registro individual en forma aislada (Harinarayan, Rajaraman, Ullman, 1996). Con ese propósito, los datos de las diferentes transacciones se almacenan y consolidan en una base de datos central denominada Almacén de Datos (AD); los analistas utilizan esas estructuras de datos para extraer información de sus negocios que les permita tomar mejores decisiones (Gupta, Harinarayan, Rajaraman, Ullman, 1997).
Basándose en el esquema de datos fuente y en los requisitos de información de la organización, el objetivo del diseñador de un **DSS** es sintetizar esos datos para reducirlos a un formato que le permita, al usuario de la aplicación, utilizarlos en el análisis del comportamiento de la empresa.

Las características de los ADs hacen que las estrategias de diseño para las bases de datos operacionales generalmente no sean aplicables para el diseño de ADs (Kimball, 1996) y (Inmon, 2005).

Estos sistemas no sólo tienen un enfoque diferente al de los operacionales, sino que, por lo general, tienen un almacenamiento diferente. Son estos sistemas sobre los que se basa la tecnología de Almacén de Datos.

1.2 - Introducción a los Almacenes de Datos

Los datos que son de interés para una organización se encuentran, frecuentemente, dispersos en múltiples fuentes de información. Para que un usuario pueda acceder a esas fuentes de un modo integrado, hace falta construir un sistema que integre (física o lógicamente) los datos de esas fuentes. Sin esta integración sería necesario consultar independientemente cada una de las fuentes y luego combinar las respuestas obtenidas.

Una solución a este problema de la integración de fuentes de datos, la constituyen los sistemas de almacenes de datos.

Los almacenes de datos integran información procedente de múltiples fuentes de datos independientes en una única base de datos, funcionando como un repositorio de información histórica que puede ser consultado directamente por los analistas de la organización. El analista usa el almacén para detectar tendencias y anomalías dentro de las actividades del negocio, conocer el estado actual de áreas de interés de la organización, y tomar decisiones de futuro.

Usualmente el almacén de datos está separado de los otros sistemas y aplicaciones de la organización, en una base de datos propia. Una razón obvia para ello reside en el hecho de que en las organizaciones es frecuente que existan diferentes repositorios de datos, soportados por tecnologías diferentes, lo que dificulta el acceso integrado a la información. Otras razones son (Porter, Romer, 1994):

− Los distintos objetivos de uso: los sistemas operacionales están orientados a los procesos (procesamiento de transacciones) mientras que los sistemas de almacenes de datos están orientados al análisis de los datos.

− Las distintas características de mantenimiento: los sistemas operacionales cambian constantemente, mientras que los sistemas de almacenes de datos no son volátiles, sólo crecen incrementalmente.

En la Figura 2, se muestra la estructura de un sistema de AD, en ella se observa como el almacén integra datos procedentes de distintas fuentes de datos, tanto internas como externas a la organización. Esta integración es el resultado de un proceso (Extracción, Transformación y Carga (ETL)), en el que los datos son preparados de la forma más adecuada para facilitar el análisis.
Los usuarios del almacén realizan este análisis por medio de herramientas OLAP (On Line AnalitycalProcesing) y de Minería de Datos (Data Mining), herramientas que permiten explorar los datos almacenados y sacar conocimiento a partir de ellos.

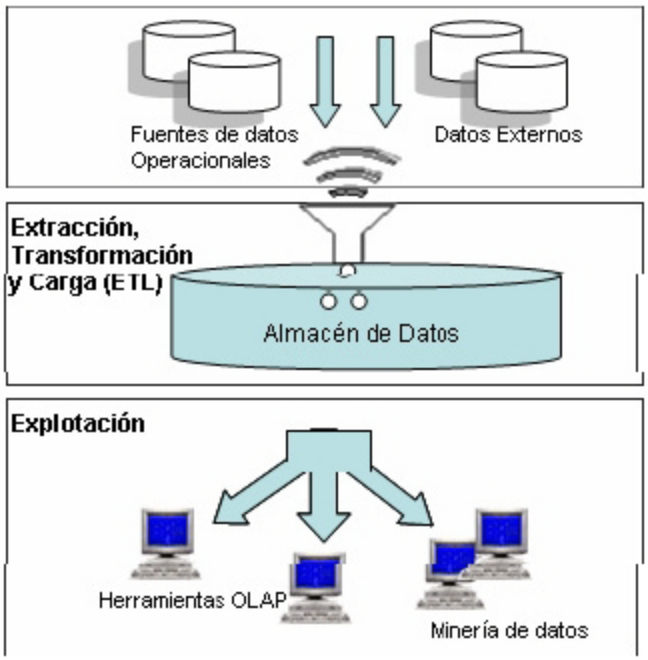

Figura 2. Sistema de Almacenes de Datos

La definición de AD, más extendida, es la propuesta por Bill Inmon:

"Un AD es una colección de datos orientados al dominio, integrados, no volátiles y variables en el tiempo, organizados para dar apoyo al proceso de toma de decisiones"(Inmon, 2005).

Esta definición, incluye el objetivo (ayuda a la toma de decisiones) y las principales características (orientados al dominio, integrados, no volátiles y variables en el tiempo). A continuación, se explican con detalle cada una de estas características:

Orientado al dominio significa que el almacén de datos está enfocado a los datos relacionados con un área de actividad del negocio. Ésta es la diferencia de enfoque con respecto a un sistema operacional que se diseña para dar soporte a los procesos básicos y cotidianos de la organización. La Tabla 1 muestra las diferencias entre los dos tipos de orientaciones en distintos contextos.

Operacional	Almacén de datos
Matrícula	Estudiantes
Prestamos	Clientes
Nóminas	Empleados
Tarjetas crédito	Clientes
Inventario	Productos

Tabla 1. Dos tipos de orientación.

Integrada significa que los datos, independientemente de las fuentes de las que proceden, son almacenados en un único repositorio, unificando su formato (integración de formato) y unificando su significado (integración semántica). La integración es un problema para muchas empresas, particularmente cuando existen muchos tipos de tecnología en uso. Por ello el proceso de integración exige costosos y largos procesos de limpieza, estandarización y agregación (resumen) de los datos.

Variante en el tiempo significa que los datos son asociados con un punto en el tiempo: diario, mensual, bimestral, trimestral, semestral, año fiscal, periodo de pago,

etc. El almacén contiene grandes volúmenes de información histórica sobre las actividades de la organización y va variando en el tiempo, recibiendo periódicamente nuevos datos.

No volátil significa que los datos no cambian (no son actualizados) una vez que se añaden al almacén de datos. Cualquiera que use el almacén de datos tiene la seguridad de que la misma consulta producirá siempre los mismos resultados.

Se puede caracterizar un **AD** haciendo un contraste de cómo los datos de un negocio son almacenados en un **AD**, difieren de los datos operacionales usados por las aplicaciones de producción o gestión.
Estos contrastes se muestran en la tabla 2:

Base de dato Operacionales	Almacén de Datos
Datos Operacionales	Dato del negocio para información
Orientado a la aplicación	Orientado al Sujeto
Actual	Actual+histórico
Detallada	Detallada+más resumida
Cambia continuamente	Estable

Tabla 2. Diferentes tipos de información.

1.2.1. Procesos que intervienen en la construcción y uso de un almacén de datos.

Para comprender mejor qué es un sistema de AD, es interesante considerar los procesos que intervienen en su construcción y uso. A continuación se describe cada uno de ellos:

Fuentes: Representan los diferentes sistemas operacionales que suministran los datos al almacén de datos.

Extracción: Es el proceso que extrae datos de las fuentes operacionales para enviarlos al almacén de datos (selección de datos). Debe realizarse una selección de

registros y campos los sistemas operacionales, ya que no todos los datos de las fuentes son relevantes para el almacén de datos. Ejemplo: la Figura 3 ilustra una selección de datos de la fuente operacional; se han seleccionado dos campos del registro (categoría e importe) y sólo interesan los registros que en categoría contengan como valor 1, 2 o 3 y que la fecha sea 30-09-2004.

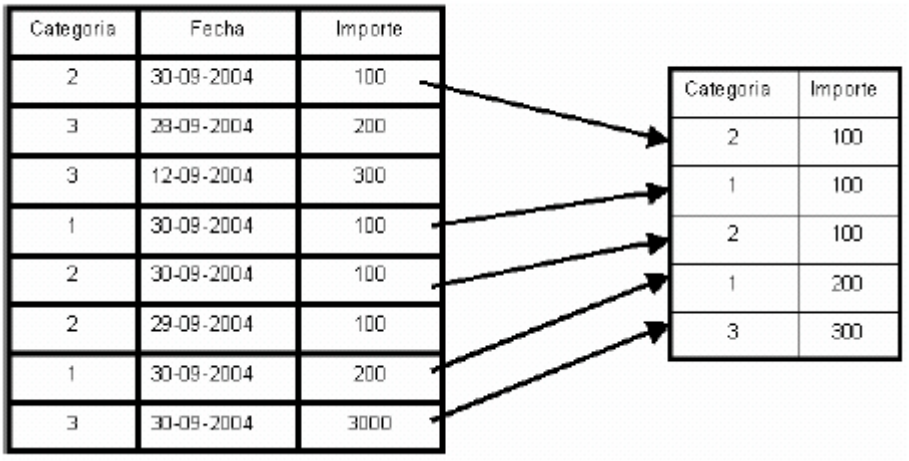

Figura 3. Extracción de datos.

Transformación: Es el proceso que prepara los datos de la manera adecuada, para ser incorporados al almacén de datos. El proceso de transformación se compone de las siguientes actividades: limpieza de datos, integración de formato, integración semántica, conversión de estructuras internas, integración de datos, resumen o agregación de datos.

- Limpieza de datos. Es necesaria porque los datos, procedentes de distintas fuentes, pueden contener errores y anomalías: inconsistencias, valores perdidos, violaciones de restricciones de integridad, etc. Para esta actividad, se pueden emplear herramientas de limpieza y/o herramientas de inspección de datos. Se puede hablar de dos tipos de limpieza:

    - Limpieza moderada: Consiste en detectar anomalías comunes; porejemplo, identificar que Av. y Avenida representan la misma información o bien eliminar puntos, comas, comillas y otros signos de puntuación.

- Limpieza intensa: Consiste en que el usuario proponga reglas para realizar la limpieza de datos; por ejemplo, el señor Juan López tiene una serie de concesiones con distintas direcciones, ¿debe considerarse como un mismo cliente o no?, esta decisión determina como se estructurará y se almacenará la información. Para trabajos de limpieza intensos, es conveniente utilizar herramientas que se han desarrollado para esas tareas. Existen dos grandes competidores: Enterprise/Integrator de Apertus Technologies y la herramienta Integrity Data Reengineering de Vality.

➢ Integración de formato: Una situación frecuente en las organizaciones consiste en que cada una de las fuentes operacionales está soportada por tecnología de diferente fabricante. Esto puede provocar que el mismo atributo sea de un determinado tipo en uno de los sistemas y de otro tipo en otro sistema. Por ejemplo, los números de cuenta bancaria o los números de teléfono pueden ser almacenados como un tipo alfanumérico en un sistema y como un tipo numérico en otro. La fecha puede ser almacenada en muchos formatos, “ddmmyy”, ”ddmmyyyy”, “yymmdd”, “yyyymmdd”, “ddmonyyyy”. Algunos sistemas almacenan los datos en miles de segundos, otros sistemas usan un entero que es el número de segundos a partir de un instante de tiempo, por ejemplo 1 de enero de 1900. Los atributos que almacenan datos de tipo dinero también son un problema, algunos sistemas almacenan el dinero como un valor entero y esperan que la aplicación inserte el punto decimal, otros tienen el punto decimal ya incorporado. En sistemas diferentes, datos de diferente tamaño son usados para valores de tipo alfanumérico como el nombre, la dirección, la descripción del producto, etc.

➢ Integración semántica: La integración semántica hace referencia al significado de los datos. Ya que la información de un almacén de datos proviene de diferentes sistemas operacionales y éstos son usados por diferentes secciones de la organización, en cada sección se puede dar un significado distinto a los datos provocando confusión al analista. Es imprescindible, por lo tanto, que cada dato que se inserte en el almacén de datos tenga un significado preciso, que sea conocido por todos los usuarios. Para este fin, un almacén de

datos debe disponer de un diccionario de datos que describa cada dato registrado en el almacén. Este diccionario es parte del almacén y es de hecho un repositorio de datos que describe los datos. El concepto de "datos acerca de datos" es referido usualmente como metadatos.

➢ Conversión de estructuras internas: Frecuentemente, los datos son estructurados de forma distinta, cuando pasan de un sistema operacional a un sistema de almacén de datos. En la Tabla 3, se muestra un ejemplo

Sistema operacional	Almacén de datos
01 Registro de piezas 02 Clave 02 U/M 02 Descripción 02 Cantidad 02 Ultima entrada 02 Código de ensamblaje 02 Cantidad mínima ...	01 Registro de piezas 02 Clave 02 Descripción 02 U/M 02 Código de ensamblaje 02 Cantidad ...

Tabla 3. Conversión de estructuras.

➢ Resumen o agregación de datos: En los sistemas de almacenes de datos, es común realizar resúmenes de registros, aplicando funciones de agregación, cuando éstos pasan de los sistemas operacionales al almacén de datos. Ejemplo: en la Figura 4 se agregan los registros por categoría.

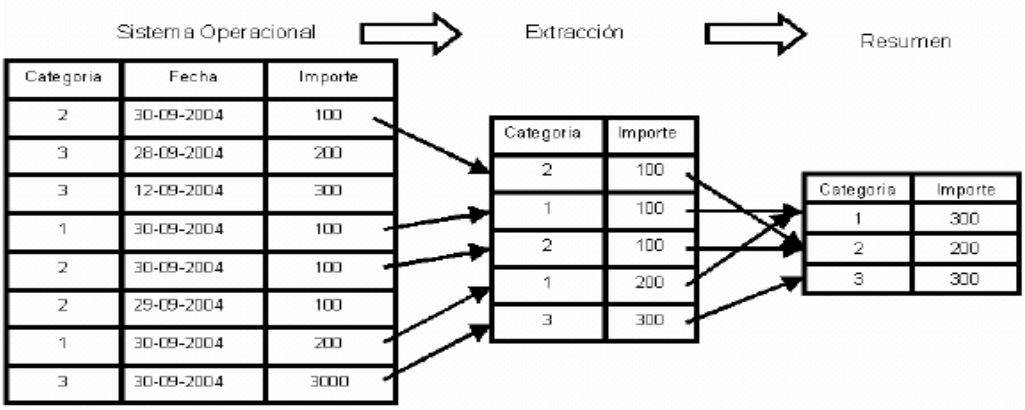

Figura 4. Agregación de datos.

- Integración de datos: Es frecuente que los datos (registros) que finalmente se van a insertar en el almacén de datos, se construyan a partir de otros registros de distintas fuentes. El proceso de integración sigue una serie de reglas que se diseñan para garantizar que el dato que se va a cargar en el almacén es correcto.

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

Existen dos formas básicas de desarrollar el proceso de carga:

- Acumulación simple: La acumulación simple es la más sencilla y común, y consiste en realizar un resumen de todas las transacciones comprendidas en el período de tiempo seleccionado y transportar el resultado como una única transacción hacia el almacén de datos, almacenando un valor calculado que consistirá típicamente en un sumatorio o un promedio de la magnitud considerada.
- Rolling: El proceso de Rolling por su parte, se aplica en los casos en que se opta por mantener varios niveles de granularidad. Para ello se almacena información resumida a distintos niveles, correspondientes a distintas agrupaciones de la unidad de tiempo o diferentes niveles jerárquicos en alguna o varias de las dimensiones de la magnitud almacenada (por ejemplo, totales diarios, totales semanales, totales mensuales, etc.).

La fase de carga interactúa directamente con la base de datos de destino. Al realizar esta operación se aplicarán todas las restricciones y triggers (disparadores) que se hayan definido en ésta (por ejemplo, valores únicos, integridad referencial, campos

obligatorios, rangos de valores). Estas restricciones y triggers (si están bien definidos) contribuyen a que se garantice la calidad de los datos en el proceso ETL, y deben ser tenidos en cuenta.

Almacén de datos: Repositorio de datos que contiene la información que ha sido extraída de los diferentes sistemas operacionales. Este repositorio de datos puede ser consultado directamente por los analistas de la organización.

Explotación: Consiste en la consulta y análisis de los datos en el AD. Desde el punto de vista del usuario, el único proceso visible es el de la explotación del AD, aunque la calidad del AD y su éxito radican en los dos procesos anteriores, que durante el desarrollo del AD, consumen la mayor parte de los recursos.

Presentación: El componente más externo en un sistema de almacén de datos es el componente de presentación, éste elabora una presentación amigable de los datos para los usuarios, ocultando la complejidad del esquema del almacén de datos.

1.2.2 Explotación de un almacén de datos: herramientas OLAP

La explotación de los AD está basada en la utilización de las herramientas OLAP (*Procesamiento Analítico en línea*), estas constituyen una tecnología de software específica para el análisis de datos en un sistema de AD. Aunque las herramientas clásicas de consulta y explotación en bases de datos (generadores de informes) podrían ser utilizadas para este fin, las características de uso y explotación específicas de los sistemas de ADs, han favorecido el desarrollo de herramientas orientadas específicamente al análisis de datos.

El objetivo de las herramientas OLAP es proveer un análisis multidimensional de la información. Para lograr esta meta, dichas herramientas emplean unmodelo multidimensional para el almacenamiento y para la presentación de los datos. Los datos objeto de análisis son almacenados en cubos que son definidos en un espacio multidimensional que facilita el análisis desde diferentes puntos de vista, independientemente del servidor que soporte el AD.

La idea básica de la perspectiva multidimensional, consiste en presentar al usuario

los datos sobre la actividad objeto de análisis en relación con los parámetros o dimensiones que caracterizan dicha actividad, en un espacio multidimensional. Cada eje del espacio representa una dimensión de la actividad y los puntos del espacio (celdas) los datos sobre la actividad (medidas) para cada combinación de valores.

En esta representación multidimensional de los datos, las dimensiones juegan un papel muy importante ya que representan los puntos de vista del análisis. En la presentación multidimensional, las dimensiones son representadas al nivel de detalle (gránulo) al que se desea registrar la actividad en el AD. Para enriquecer las posibilidades de análisis, las dimensiones se completan con un conjunto de atributos descriptivos. Estos atributos van a realizar distintas funciones durante el análisis de los datos: permitir establecer condiciones para filtrar los datos, definir criterios de agregación para obtener datos resumidos, etc.

Cuando una dimensión se completa con atributos descriptivos, es usual que estos atributos no sean independientes entre sí, entre ellos suelen aparecer dependencias funcionales que definen jerarquías dentro de la dimensión, estas jerarquías van a desempeñar un papel importante durante el análisis. Estas jerarquías definidas van a permitir, además cambiar dinámicamente el nivel de agregación al que se observan los datos de una consulta.

Para ello las herramientas de OLAP, introducen un tipo nuevo de operadores, los operadores de DRILL (disgregación) y de ROLL (agregación), que permiten cambiar el nivel de agregación de los datos de una consulta, “navegando” a través de las jerarquías.

ROLL-UP: El operador roll-up, permite reducir el nivel de detalle al que se consultan los datos, realizando agregaciones a través de las jerarquías de las dimensiones. Por ejemplo: resumiendo los datos semanales en trimestrales o en anuales.

DRILL-DOWN: Esta operación, es la inversa de la operación roll-up, es decir permite aumentar el detalle al que se consultan los datos, al ir a un nivel más bajo dentro de la jerarquía. Por ejemplo: disgregando las ventas nacionales en ventas por regiones y después éstas en ventas por subregiones.

Otros operadores son SLICE y DICE que permiten reducir el conjunto de datos consultados, por medio de operaciones de proyección y selección de los datos basándose en los atributos de las dimensiones.

SLICE: Consiste en eliminar una dimensión de la consulta activa restringiendo los valores de dicha dimensión a un valor o a un rango de valores. Por ejemplo: si de la dimensión tiempo tomamos únicamente el mes de ¨Marzo¨.

DICE: Consiste en focalizar la consulta en un subcubo del cubo de datos, restringiendo los valores en varias de las dimensiones.

Las operación **FILTRAR** consiste en realizar una selección sobre los datos de un cubo utilizando alguna constante mientras que la operación **PIVOT** sirve para visualizar desde distintos ángulos el cubo, permite rotar los ejes del cubo para examinarlo desde ese punto de vista.

Resumiendo, una herramienta de OLAP, es una herramienta para el análisis de datos en un sistema de AD, que tiene las siguientes características:

- Presentación multidimensional de los datos objeto de análisis. En un esquema multidimensional, se presenta la actividad objeto de análisis (**hechos**) descrita por un conjunto de indicadores (**medidas**) y las **dimensiones** que caracterizan la actividad al nivel de detalle (**gránulo**) al que se almacena, estando estas dimensiones descritas por un conjunto de atributos **(atributos de dimensión**).
- Simetría en el conjunto de dimensiones: una dimensión no es más importante que otra.
- Posibilidad de definir jerarquías dentro de las dimensiones para aplicar operadores de DRILL y ROLL.
- Posibilidad de aplicar operaciones de selección (filtros) en los datos.
- Facilidades para el análisis de datos: funciones estadísticas, visualización gráfica, etc.

Debido a que las herramientas OLAP siguen una perspectiva multidimensional, las metodologías de modelado que se han desarrollado para el diseño de ADs han adoptado también esta perspectiva, y por ello se habla de "modelado multidimensional" y de "modelo multidimensional de datos".

1.3 Diseño de un Almacén de Datos

La tecnología de Almacenes de Datos debido a su orientación analítica, impone un procesamiento y pensamiento distinto, la cual se sustenta por un modelamiento de Bases de Datos propio, conocido como Modelamiento Multidimensional, el cual busca ofrecer al usuario su visión respecto de la operación del negocio.
El Modelamiento Dimensional es una técnica para modelar bases de datos simples y entendibles al usuario final. La idea fundamental es que el usuario visualice fácilmente la relación que existe entre las distintas componentes del modelo.
Para construir un AD se debe primero tener claro que existe una diferencia entre la estructura de la información y la semántica de la información, y que esta última es mucho más difícil de abarcar y que también es precisamente con ella con la que se trabaja en la construcción de un AD.
Aquí se encuentra la principal diferencia entre los sistemas operacionales y el AD: Cada uno de ellos es sostenido por un modelo de datos diferente. Los sistemas operacionales se sustentan en el Modelo Entidad Relación (MER) y DDW trabaja con el Modelo Multidimensional.

1.3.1 Modelos Multidimencionales

Un modelo de datos es una representación de los datos y sus relaciones con otros datos que se utiliza para conocer cómo se organizará los datos en bases de datos u otro medio de almacenaje y administración de datos.
Podemos definir de forma general a un modelo de datos como:

"*Una serie de conceptos que puede utilizarse para describir un conjunto de datos y operaciones para manipular los mismos*" (Batini, Ceri, &Navathe, 1992).

Un modelado de datos multidimensional o cubo es una colección de medidas las cuales dependen de un conjunto de dimensiones, es una representación de los datos

que permite organizarlos en la forma de hechos y dimensiones. Los hechos representan una actividad objeto de análisis que contienen medidas, es decir, la información a nivel transacción que vamos a analizar, por ejemplo: compra, ventas, préstamos, etc. Las dimensiones contienen información descriptiva de esas transacciones que se representan por un conjunto de atributos, por ejemplo: fecha, cliente, producto, etc.

Para entender la definición presentada así como el modelo multidimensional se deben comprender tres conceptos: cubo, medida, dimensión.

CUBO

Un modelo de datos multidimensional soporta el manejo de una vasta cantidad de datos empresariales y temporales. De esta forma surge la instancia del modelo multidimensional, también conocido como cubo o hipercubo.

Para clasificarlo un poco se puede imaginar un cubo con tres dimensiones: producto, tiempo, región; donde cada dimensión tiene diferentes niveles o hechos, para finalmente intersectar estos valores y obtener una medida. La medida es el índice de un producto como puede ser el huevo en el mes de mayo y en la zona centro del país, ver figura5.

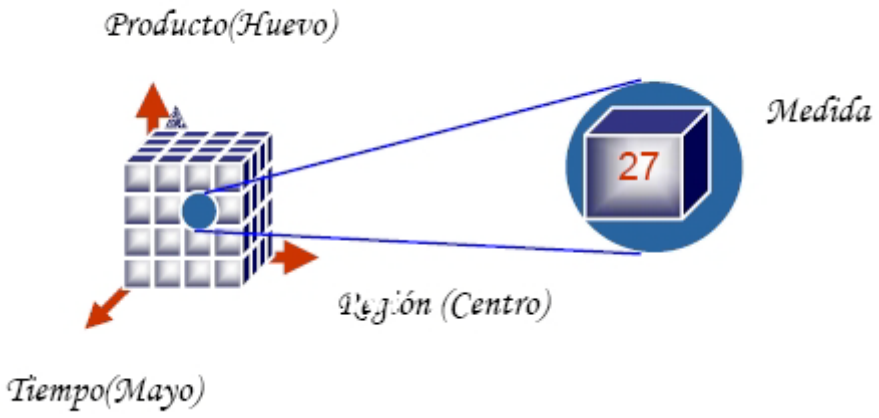

Figura 5. Esquema del cubo.

En base a esta estructura el esquema del cubo, se enriquece permitiendo el manejo de restricciones. Estas restricciones pueden ser ampliamente clasificadas como restricciones de llaves, restricciones de integridad referencial, restricciones no nulas. Las restricciones pueden ser clasificadas en dos categorías:

- Restricciones intra-cubo: definen restricciones dentro del cubo, explotando la relación que existe dentro de los distintos atributos del cubo.
- Restricciones inter-cubo: define restricciones entre dos o más cubos.

Dimensión

Las dimensiones son un concepto esencial de bases de datos multidimensionales. Las dimensiones son usadas para seleccionar y agregar datos a un cierto nivel de detalle. Cada dimensión está descrita por un conjunto de atributos organizados en jerarquías de niveles, atendiendo a las dependencias funcionales entre estos atributos.

Una jerarquía implica una organización de niveles dentro de una dimensión, con cada nivel representando el total agregado de los datos del nivel inferior. Las jerarquías definen cómo los datos son sumarizados desde los niveles más bajos hacia los más altos. Una dimensión típica soporta una o más jerarquías naturales. Una jerarquía puede pero no exige contener todos los valores existentes en la dimensión.

Viendo los datos dentro de un cubo se tiene la ventaja de que se puede manejar cualquier número de dimensiones.

Medida

La medida o hecho es un dato numérico que representa una actividad específica de un negocio, mientras que una dimensión representa una perspectiva de los datos. Cada dimensión está descrita por un conjunto de atributos (datos agregados). A su vez se pueden intersectar estas dimensiones para obtener un valor, llamado medida.

Los hechos representan el patrón de interés o el evento dentro de una empresa que necesita ser analizado. Los hechos son implícitamente definidos por la combinación de valores de las dimensiones. Un AD comúnmente maneja tres tipos de hechos:

- Eventos: Con la granularidad más fina, típicamente modela eventos del mundo real.
- Fotos fijas: Modelan entidades en un punto dado en el tiempo.
- Fotos fijas acumulativas: Modelan actividades en un punto dado en el tiempo.

Las medidas son resultados cuantificables, o indicadores clave de desempeño usados para determinar el éxito de una operación de negocios. Orientan las respuestas a preguntas relacionadas con cuestiones numéricas como la cantidad valor o costo. Un

cubo puede contener una o varias medidas, dependiendo del diseño y los requerimientos. Existen dos tipos de medidas:

- Medida regular: toma su dato directamente de una fuente disponible. Es un compendio de información que ya se tiene, tal como el número de unidades vendidas, ingresos, gastos, etc.
- Medida calculada: obtiene como resultado un nuevo dato numérico para medidas que no están en una fuente directa disponible. Es derivada de otras medidas, por ejemplo: ganancias (ingresos- costos), margen de ganancias (ingreso- costos/ingresos) , tiempo promedio de espera(fecha de entrega- fecha de la orden), etc.

1.3.1.1Caracteristicas de los Modelos Multidimensionales.

La estructura básica de un AD para el modelo multidimensional está definida por dos elementos esénciales las tablas y los esquemas.

- Tabla en AD: como cualquier BD relacional, un AD se compone de tablas. Tabla de hechos y dimensiones.

- Esquema AD. La colección de tablas en el AD se conoce como esquema. El esquema multidimensional es un esquema relacional compuesto de una tabla de hecho y **n** tablas de dimensiones. Los esquemas caen dentro de tres categorías básicas: Esquema de Estrella, Esquema de Copo de Nieve y Esquema Constelación.

Tabla Fact o de Hechos.

Es la tabla central en un esquema dimensional. Es en ella donde se almacenan las mediciones numéricas del negocio. Estas medidas se hacen sobre el grano, o unidad básica de la tabla.

El grano o la granularidad de la tabla queda determinada por el nivel de detalle que se almacenará en la tabla. El grano revierte las unidades atómicas en el esquema dimensional.

Cada medida es tomada de la intersección de las dimensiones que la definen. Idealmente está compuesta por valores numéricos, continuamente evaluados y

aditivos. La razón de estas características es que así se facilita que los miles de registros que involucran una consulta sean comprimidos en unas pocas líneas en un set de respuesta.

La clave de la tabla fact recibe el nombre de clave compuesta o concatenada debido a que se forma de la composición (o concatenación) de las llaves primarias de las tablas dimensionales a las que está unida.

Así entonces, se distinguen dos tipos de columnas en una tabla fact: columnas fact y columnas key. Donde la columna fact es la que almacena alguna medida de negocio y una columna key forma parte de la clave compuesta de la tabla.

Tabla Lock_up o de Dimensiones.

Estas tablas son las que se conectan a la tabla de hecho, son las que alimentan a la tabla de hecho. Una tabla de dimensiones almacena un conjunto de valores que están relacionados a una dimensión particular. Las tablas de dimensiones no contienen hechos, en su lugar los valores en las tablas de dimensiones son los elementos que determinan la estructura de las dimensiones. Así entonces, en ellas existe el detalle de los valores de la dimensión respectiva.

Una tabla de dimensiones está compuesta de una llave primaria que identifica unívocamente una fila en la tabla junto con un conjunto de atributos, y dependiendo del diseño del modelo multidimensional puede existir una llave foránea que determina su relación con otra tabla de dimensiones.

Para decidir si un campo de datos es un atributo o un hecho se analiza la variación de la medida a través del tiempo. Si varía continuamente implicaría tomarlo como un hecho, caso contrario será un atributo.

Los atributos dimensionales son un rol determinante en el diseño del AD. Ellos son la fuente de todas las necesidades que debieran cubrirse. Esto significa que la base de datos será tan buena como lo sean los atributos dimensionales, mientras más descriptivos, manejables y de buena calidad, mejor será el diseño del AD.

Modelo o Esquema de Estrella.

El esquema de estrella (starschema) es una técnica de modelado de datos usada para hacer corresponder un modelo multidimensional a una base de datos relacional, debe su nombre debido a que gráficamente parece una estrella. El esquema de estrella

tiene cuatro componentes: hechos, dimensiones, atributos y jerarquías de atributos. Los hechos y dimensiones son representados por tablas físicas en el almacén de datos, la tabla de hechos está relacionada a cada dimensión en una relación uno a muchos. Las tablas de hechos y dimensiones están relacionadas por llaves foráneas y están sujetas a las restricciones de llaves foráneas y primarias.

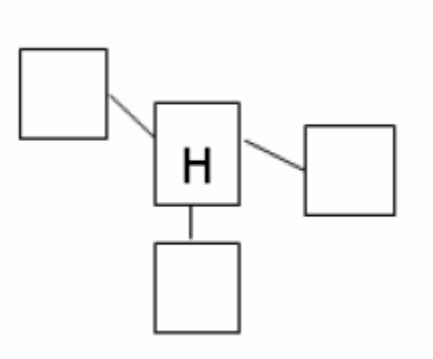

Figura 6: Esquema de estrella.

Modelo de Copo de Nieve

El esquema de copo de nieve (snowflakeschema) es una variación del modelo de estrella, lo que se hace es que en cada dimensión se almacenan jerarquías de atributos o bien simplemente se separan atributos en otra entidad por razones de desempeño y mejor utilización del espacio. En el esquema copo de nieve las tablas de dimensiones son normalizadas para simplificar las operaciones de selección de datos, con lo que logra presentar la información sin redundancia, evitando así las anomalías.

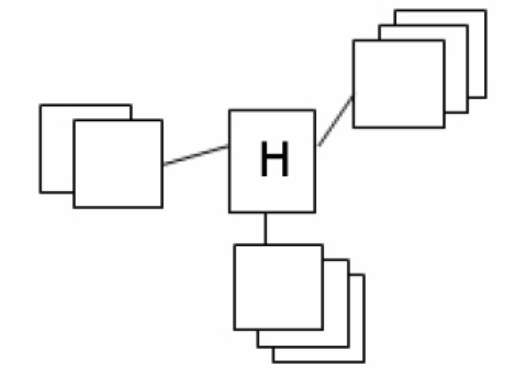

Figura 7: Esquema copo de nieve.

Modelo de Constelación de hechos.

Para cada esquema estrella o esquema del copo de nieve en un almacén de datos es posible construir un esquema de constelación de hechos.
Este esquema es más complejo que las otras arquitecturas debido a que contiene múltiples tablas de hechos. Con esta solución las tablas de dimensiones pueden estar compartidas por más de una tabla de hecho.
El esquema de constelación de hechos tiene mucha flexibilidad y este es su grande virtud. Sin embargo, el problema es que cuando el número de las tablas vinculadas aumenta, la arquitectura puede llegar a ser muy compleja y difícil para mantener.
En un esquema de constelación de hechos las distintas tablas de los hechos están asignadas a las dimensiones relevantes para cada uno de los hechos. Esto puede ser útil cuando los hechos están asignados a un nivel de una dimensión y los otros hechos a otro nivel de detalle de una dimensión.
La utilidad principal de este modelo es que al tener dimensiones que puede ser compartidas por diferentes cubos se tendrá un mejor uso de espacio de almacenamiento evitando la redundancia.

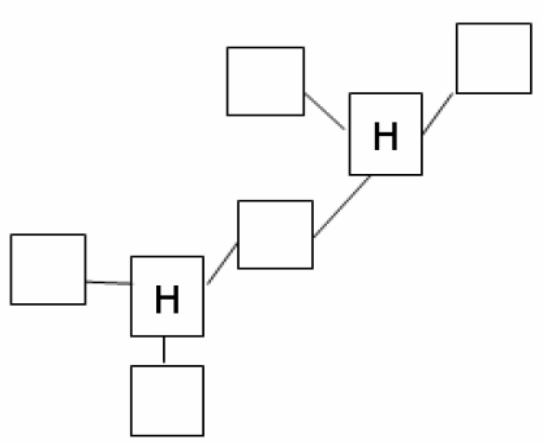

Figura 8: Esquema constelación de hechos.

1.3.2 Niveles para la modelación de los datos en los AD.

La modelación de los datos en el diseño de un AD al igual que el proceso de diseño de una Bases de datos transita a través de una serie de pasos en los cuales se va avanzando de un nivel de abstracción menor a otro más profundo, mediante la elaboración de una sucesión de modelos

Como en los sistemas de bases de datos tradicionales, la modelación de los datos del DW puede dividirse en tres niveles secuenciales: diseño conceptual, diseño lógico y diseño físico (Batini, Ceri, & Navathe, 1992). En la Figura 9 se muestran las etapas con sus respectivas entradas y salidas de información.

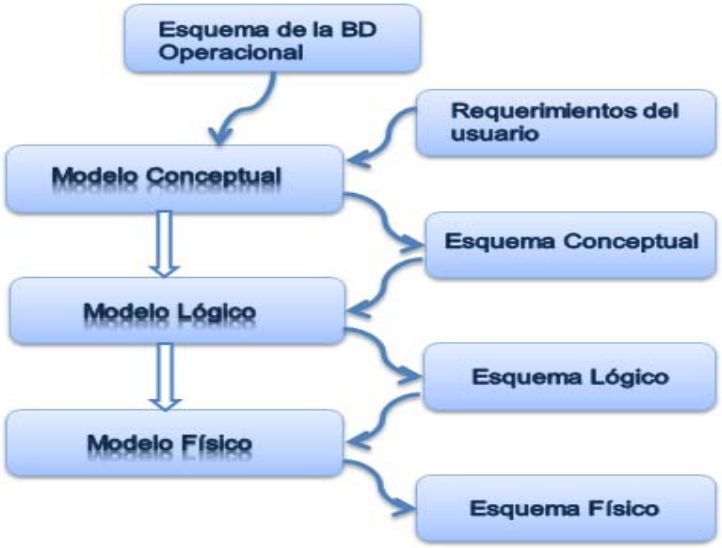

Figura 9: Etapas del DAD.

Los modelos conceptuales, están más próximos al usuario y son independientes de la tecnología que se vaya a utilizar, los modelos lógicos dependen del tipo de gestor de bases de datos, y los modelos físicos dependen de los sistemas comerciales particulares.

En un contexto de bases de datos, en el modelado conceptual el dominio se representa por medio de las primitivas de algún lenguaje de modelado conceptual: diagramas Entidad Relación (modelo ER), diagramas de clases (UML), etc.; una vez que se tiene el esquema conceptual éste se traduce a un modelo lógico implementado en algún Sistema Gestor de Base de Datos (SGBD) comercial (modelo relacional, modelo red o modelo jerárquico), lo que se conoce como modelado lógico y por último el esquema lógico se implementa en un SGBD comercial (modelado físico).

1.3.2.1. Modelos conceptuales.

El modelo conceptual es un modelo intermedio entre la realidad informativa y el modelo lógico, el modelo conceptual se define fuera del SGBD que se emplee. El

proceso de modelación conceptual también es denominado modelación semántica, ya que con este modelo se pretende reflejar, en mayor medida, la semántica o significado de los datos y sus interrelaciones.

Es ampliamente aceptado que en el diseño de un AD a nivel conceptual se sigua un modelo multidimensional (Kenan, 1996), (Carpani, 2000), este modelado multidimensional se basa en la dualidad hecho - dimensión, un hecho representa una actividad objeto de análisis, actividad que está caracterizada por un conjunto de dimensiones. En un esquema multidimensional se representa un hecho y las dimensiones que lo caracterizan. Esta representación es normalmente en forma de estrella: el hecho se representa en la parte central y las dimensiones en las puntas de la estrella.

El modelo conceptual de un AD se define desde el mayor nivel de abstracción y contiene únicamente los objetos y relaciones más importantes. (Luján Mora, 2005). A este nivel resurgen los modelos relacionales multidimensionales ((Kenan, 1996), (Carpani, 2000)), que representan la información como matrices multidimensionales o cuadros de múltiples entradas denominados cubos. A los ejes de la matriz se los llama dimensiones y representan los criterios de análisis, y a los datos almacenados en la matriz se los llama medidas y representan los indicadores o valores a analizar.

A nivel conceptual un componente adicional a tener en cuenta son las bases fuentes. Un AD no es una base de datos para construir desde cero, sino que debe construirse con información extraída de un cierto conjunto de bases fuente.

Consideramos que el diseño conceptual es la base necesaria para la construcción de un sistema de información bien documentado y que responda a los requisitos del usuario. Generalmente en las empresas los sistemas de información se encuentran implementados en sistemas de bases de datos operacionales que en el 90% utilizan el modelo Entidad / Relación (E/R) por lo que se han desarrollado innumerables esfuerzos para poder obtener el modelo conceptual del AD a partir de los esquemas E/R de los sistemas operacionales.

1.3.2.2 Modelo lógico.

Una de las tareas más importantes en la construcción de un DW es la construcción de su esquema lógico. El esquema lógico es una especificación más detallada que el esquema conceptual, donde se incorporan nociones de almacenamiento, performance y estructuración de los datos.

A nivel lógico surgen implementaciones de los cubos tanto para bases de datos relacionales como multidimensionales. El nivel lógico es un modelado dependiente de la tecnología utilizada, ROLAP (relacional) o MOLAP (multidimensional) y más recientemente HOLAP (un hibrido).

Durante el diseño lógico deben considerarse las bases de datos operacionales y cómo se corresponden con el esquema conceptual. Por lo tanto es esencial poder relacionar los elementos del esquema conceptual con las tablas y atributos de las bases fuentes. También a nivel lógico a nuestra consideración también debe de modelarse el proceso de limpieza, transformación y cargas.

1.3.2.3. Modelo Físico.

A nivel físico se visualiza un modelado dependiente del gestor comercial que soporta la implementación. Durante la etapa de diseño físico se incorporan elementos específicos de almacenamiento y performance, como son la elección de índices, almacenamiento especializado, parámetros de sistemas, etc.

La estructura física puede ser representada con diferentes configuraciones:

- Arquitectura centralizada. Todo el Almacén de datos se encuentra en un único servidor.

- Arquitectura distribuida. Los datos del Almacén se reparten entre varios servidores. Asignando cada servidor a uno o varios temas lógicos.

- Arquitectura distribuida por niveles. Refleja la estructura lógica del Almacén, asignando los servidores en función del nivel de agregación de los datos que contienen. Un servidor está dedicado para los datos de detalle, otro para los

resumidos y otro para los muy resumidos. Cuando los datos muy resumidos se duplican en varios servidores para agilizar el acceso se habla de Supermercados de Datos (*Data Marts*).

1.4 Sistemas ROLAP y sistemas MOLAP.

Así como para las metodologías de diseño de almacenes de datos se ha consensuado y adoptado un modelo multidimensional de datos, respecto a la implementación (esquema físico) del esquema conceptual (multidimensional), se han seguido dos aproximaciones: los sistemas ROLAP y los sistemas MOLAP.

- Sistemas ROLAP (Relational On-Line AnalyticalProcessing): se trata de usar gestores de bases de datos relacionales, con ciertas extensiones y facilidades nuevas. En estos sistemas, las herramientas OLAP son las encargadas de transformar el esquema relacional del almacén en un esquema multidimensional para el usuario, se ilustra en la Figura 10.
- Sistemas MOLAP (Multidimensional On Line AnalyticalProcessing): se trata de construir gestores de propósito específico, basados en el uso de estructuras de almacenamiento de tipo matrices multidimensionales (Figura 10), que favorezcan el tipo de análisis que se hace en estos sistemas.

La principal ventaja de los sistemas MOLAP es que mantienen una correspondencia directa entre los datos almacenados y la vista que de ellos tiene el usuario (multidimensional). Se ha demostrado que las matrices multidimensionales son estructuras de datos muy adecuadas para el análisis de datos.

La ventaja principal de los sistemas ROLAP reside en la posibilidad de utilizar una tecnología ampliamente extendida y utilizada para la gestión de datos, los sistemas relacionales (Becker, 2003).

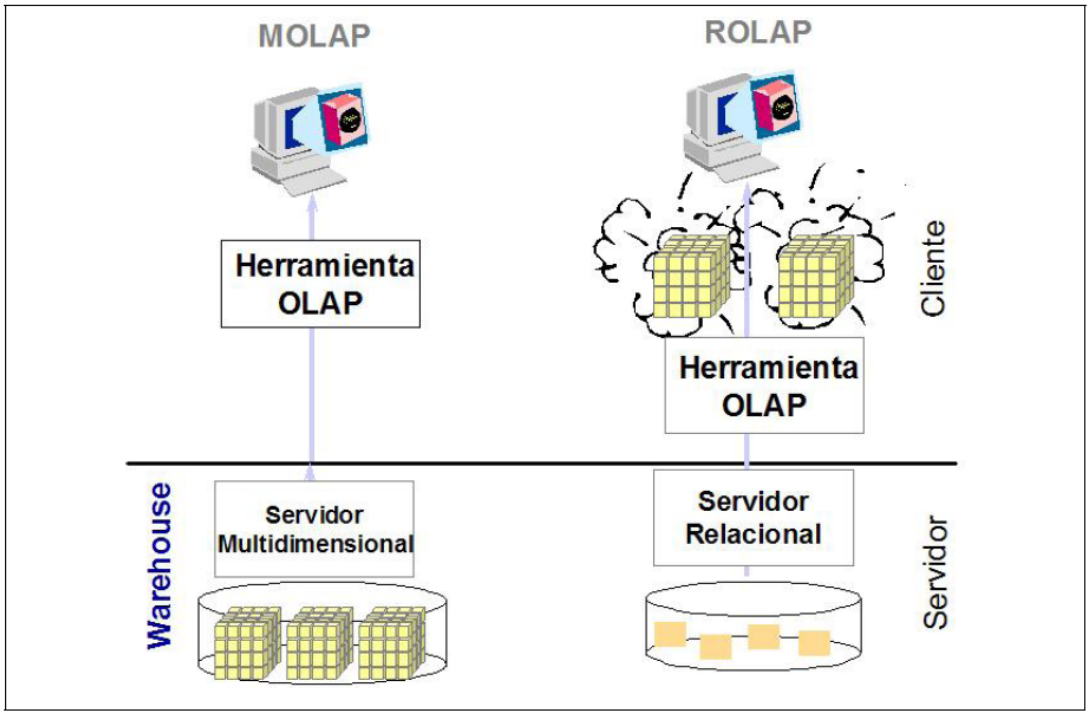

Figura 10: Sistemas MOLAP y Sistema ROLAP.

1.5 Mantenimiento.

La posibilidad de tener “datos frescos”, es importante para las aplicaciones de los negocios. Existen dos formas de refrescar datos: la primera es llevando los datos al AD segundos después de que las fuentes fueron actualizadas, un ejemplo claro de esto son las transacciones de un banco. La segunda es acumulando y almacenando los datos ya integrados y transformados, en un sitio intermedio para finalmente de forma periódica pasar dicha información ya al AD.

Refrescar un AD consiste en propagar las actualizaciones de las fuentes. Hablando del refrescado, hay dos cuestiones que debemos considerar: ¿Cuándo refrescar? y ¿Cómo refrescar?

Respondiendo al cómo refrescar, se puede realizar un refrescado incremental o bien un recalculado de los datos; y respondiendo al cuándo, se puede hacer a solicitud explicita del administrador, o periódicamente con un tiempo determinado. Uno de los

métodos más empleados es el refrescado periódico, pero esto depende mucho de las cualidades de los datos que maneje el AD.

Así pues, el refrescado de un AD es considerado como un problema crítico y difícil debido a tres principales razones. Primero, el volumen de datos almacenados en el AD es muy grande y crece cada vez más. La segunda razón, es ya que el refrescado debe ser accesible a los diferentes cambios de desempeño o ejecución del AD. Y por último, el refrescado envuelve transacciones que acceden múltiples datos, lo que implica cálculos complejos que producen un alto nivel de agregación.

El componente de extracción debe ser capaz de recuperar y guardar todos los cambios que ocurren en las fuentes. Un monitor es una parte del componente de extracción que es responsable de la actualización de los datos, es decir, el refrescado. El refrescar implica dos problemas. El primero es que la integración debe ser incremental, y segundo es el de reducir la cantidad de información que debe ser incorporada en el AD.

1.6 Conclusiones Parciales.

En este capítulo abordamos sobre los principales conceptos y términos que son necesarios tener en cuenta a la hora de diseñar un almacén de datos. Losalmacenes de datos son una buena técnica para el problema al que nos enfrentamos ya que estos son una colección de datos donde se encuentra integrada la información de estas, proporcionando una herramienta para que puedan hacer un mejor uso de la información y para el soporte al proceso de toma de decisiones. Una vez mostrada la técnica DW en el siguiente capítulo se muestra el diseño del modelo dimensional del almacén propuesto para la actividad de postgrado.

CAPITULO 2. Análisis y Diseño del Almacén de Datos.

Introducción

En este capítulo realizaremos el análisis del sistema, para esto se utilizó el Lenguaje Unificado de Modelado (UnifiedModeling Language, UML) ya que es un lenguaje gráfico estándar para escribir planos de software. UML prescribe un conjunto de notaciones y diagramas estándar para modelar sistemas orientados a objetos, y describe la semántica esencial de lo que estos diagramas y símbolos significan. También abordaremos sobre el diseño Lógico del Almacén de Datos, para esto se utilizó la herramienta Embarcadero ER/Studio ya que ofrece a los administradores y desarrolladores de bases de datos la posibilidad de modelado de datos de forma visual, permitiendo el diseño y mantenimiento de bases de datos transaccionales, de soporte a la toma de decisiones y para Web. ER/Studio también soporta diseño multinivel y ofrece la capacidad de controlar, documentar y desplegar rápidamente cambios en el diseño en las principales plataformas.

2.1 Conocimiento de los requerimientos

Los requerimientos son una descripción de las necesidades o deseo de un producto. La meta primaria de la fase de requerimientos es identificar y documentar lo que en realidad se necesita, en una forma que claramente se lo comunique al cliente y a los miembros del equipo de desarrollo. (Larman, ed. P. Hall. 1999.)

2.2 Descripción de los requerimientos

Se necesita crear un almacén de datos que partiendo de la información histórica asociada al Sistema de Gestión de Postgrado, permita el almacenamiento de la misma en formato dimensional y a partir de allí posibilite elaborar consultas y generar informes ad hoc que brinden tendencias del aprovechamiento académico de graduados universitarios, instituciones de la producción y los servicios, etc., y posibilite valorar la eficiencia y productividad asociada a las actividades de postgrado.

2.3 Diagrama de Casos de Uso

En UML, un Diagrama de casos de uso es una especie de diagrama de comportamiento que muestra la relación entre los actores y los casos de uso del sistema. Un diagrama de casos de uso consta de los siguientes elementos: Actor, Casos de Uso y Relaciones.

Actor

Un Actor es un rol que un usuario juega con respecto al sistema. Es importante destacar el uso de la palabra rol, pues con esto se especifica que un Actor no necesariamente representa a una persona en particular, sino más bien la labor que realiza frente al sistema.

Casos de Uso

Un caso de uso es una secuencia de interacciones que se desarrollarán entre un sistema y sus actores en respuesta a un evento que inicia un actor principal sobre el propio sistema. Es una operación/tarea específica que se realiza tras una orden de algún agente externo, sea desde una petición de un actor o bien desde la invocación desde otro caso de uso.

Relaciones

Una relación es una conexión entre los elementos del modelo: Actor y Casos de Uso.

2.3.1 Casos de Uso del Almacén de Datos.

Los casos de usos asociados al almacén de datos se basa en poder elaborar consultas ad hoc o sea sobre la marcha sobre aspectos del postgrado a diferentes niveles de abstracción, que a su vez incluye generar informes. También existe el caso de uso extraer, transformar y limpiar datos pero este es solamente para el actor encargado de mantener el almacén.

2.3.2 Actores del Almacén de Datos.

Los actores del almacén de datos son:

- Actor Facultad:
  - -Vicedecano de PG
  - -Decano

- -Director CE/I

➢ Actor VRector:

- -Rector
- -Asesor de PG
- -ViceRector de PG
- -Director de PG

➢ Administrador

2.3.3 Diagrama de Casos de Uso del Almacén de Datos

A continuación se muestra los casos de uso principales cuyo desarrollo es necesario para la implementación satisfactoria de los requerimientos del almacén de datos para las actividades de postgrado.

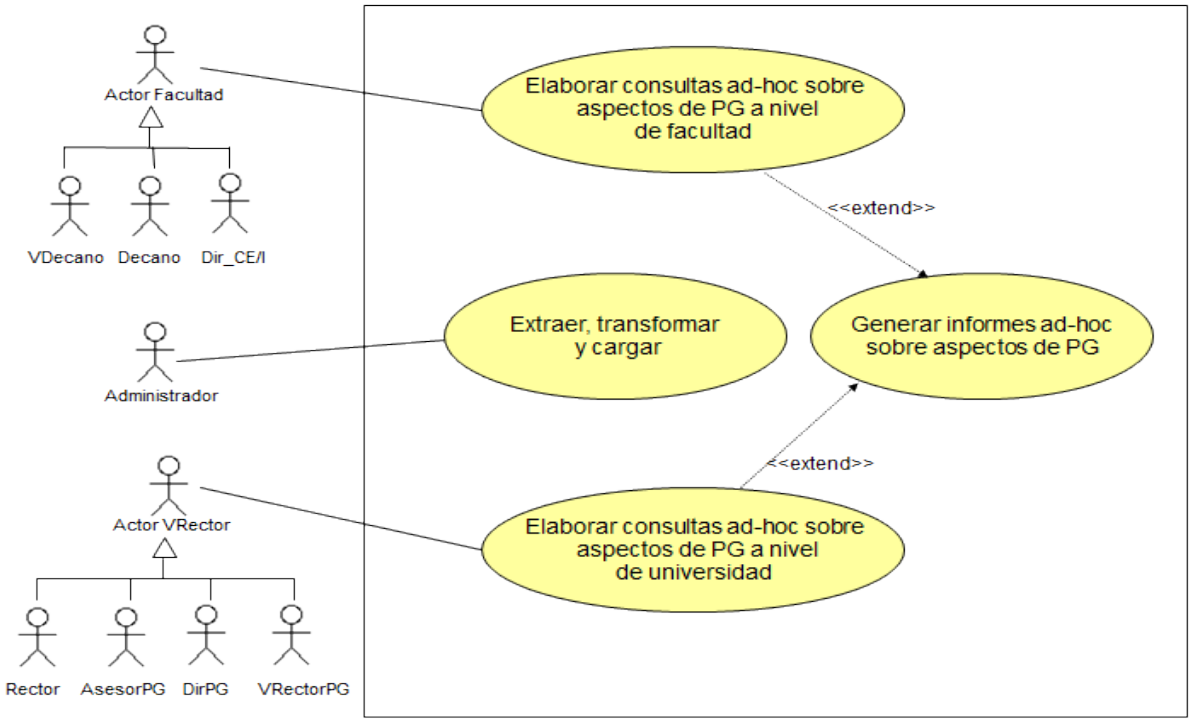

Figura: 11 Diagrama de Casos de Usos.

2.4 Diagrama de Actividades

En UML un diagrama de actividades muestra la serie de actividades que deben ser realizadas en un caso de uso, así como las distintas rutas que pueden irse desencadenando en el caso de uso. Los diagramas de actividades muestran el flujo de trabajo desde el punto de inicio hasta el punto final detallando muchas de las rutas de decisiones que existen en el progreso de eventos contenidos en la actividad. Estos también pueden usarse para detallar situaciones donde el proceso paralelo puede ocurrir en la ejecución de algunas actividades.

Un diagrama de actividad es utilizado en conjunción de un diagrama de caso de uso para auxiliar a los miembros del equipo de desarrollo a entender como es utilizado el sistema y cómo reacciona en determinados eventos.

Los elementos fundamentales que constituyen un Diagrama de Actividades son:

Inicio

El inicio de un diagrama de actividad es representado por un círculo de color negro sólido.

Actividad

Una actividad representa la acción que será realizada por el sistema.

Transición

Una transición ocurre cuando se lleva acabo el cambio de una actividad a otra.

Ramificación (Branch)

Una ramificación ocurre cuando existe la posibilidad que ocurra más de una transición (resultado) al terminar determinada actividad.

Bifurcación (Fork)

Un fork representa una necesidad de ramificar una transición en más de una posibilidad. Aunque similar a una ramificación (Branch) la diferencia radica en que un fork representa más de una ramificación obligada, esto es, la actividad debe proceder por ambos o más caminos, mientras que una ramificación (Branch) representa una transición u otra para la actividad (como una condicional).

Unión (Join)

Un join ocurre al fusionar dos o más transiciones provenientes de un fork, y es empleado para dichas transiciones en una sola, tal y como ocurría antes de un fork .

Fin

El fin de un diagrama de actividad es representado por un círculo, con otro círculo concéntrico de color negro sólido.

2.4.1 Diagramas de Actividades del Almacén de Datos

Al iniciar el sistema cualquiera de los actores anteriormente señalados deben identificarse (con nombre y contraseña). Esto lo verifica el sistema y a partir de una comprobación válida, debe presentarse la interfaz de usuario o ventana principal asociada a cada uno. Una vez que el sistema verifique al usuario este puede comenzar a elaborar consultas sobre los aspectos de postgrado que necesite analizar para así poder tomar una mejor decisión, cuando se termina de elaborar la consulta esta se ejecuta mostrando una tabla con los resultados de la misma pudiendo así generar un informe con ellos si se desea. Cuando termina se puede realizar otra consulta para hacer otro análisis desde otro punto de vista, o dimensiones sino se sale del sistema. El siguiente diagrama de actividad figura 12, muestra la actividad de Elaborar Consulta.

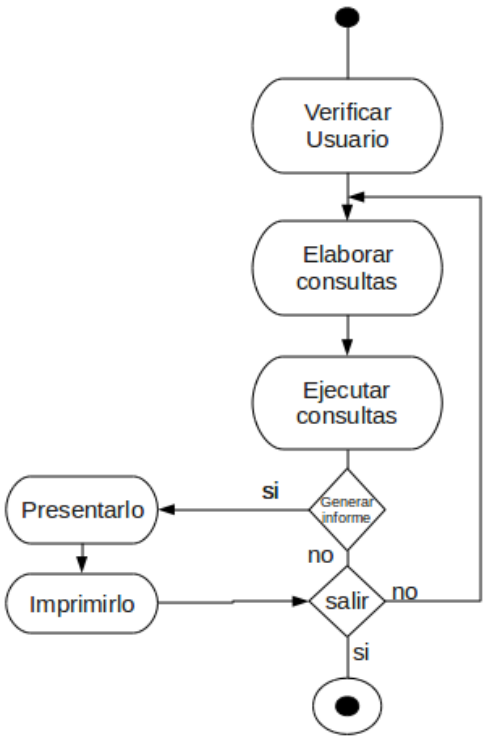

Figura: 12 Actividad de Elaborar Consulta.

El administrador del sistema es el encargado de realizar el mantenimiento del almacén de datos para las actividades de postgrado; para esto se determinó una regla: el proceso de ETL se realizara una vez al año después de que se cierra el informe de postgrado del área y se tomaran solo los postgrados cerrados. En la figura 13 se muestra el diagrama de actividad referente a este cao de uso.

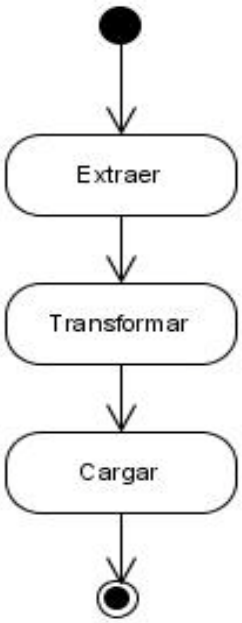

Figura: 13 Actividad de ETL.

2.5 Diagramas de Componentes.

Un diagrama de componentes representa cómo un sistema de software es dividido en componentes y muestra las dependencias lógicas entre estos componentes. Los diagramas de componentes describen los elementos físicos del sistema (incluyen archivos, cabeceras, bibliotecas compartidas, módulos, ejecutables, o paquetes.) y sus relaciones muestran las opciones de realización incluyendo código fuente, binario y ejecutable. Son utilizados para modelar la vista estática y dinámica de un sistema.

2.5.1 Diagrama de Componentes del Almacén de Datos

El almacén de datos está compuesto por diferentes componentes como son: BD SPG, Almacén, Kettle, SGBD Postgres y la aplicación Web, ver figura 14. El componente Kettle es el encargado de realizar el proceso de las ETL, este toma los datos del

componente BD SPG que no es más que la base dato operacional del Sistema de Postgrado y luego de realizar las transformaciones necesarias los cargas en el componente Almacén. Se tomó como Sistema de Gestor de Base Datos el Postgres por ser software Libre que es uno de los requerimientos del sistema. La aplicación Web es la interfaz gráfica que les permitirá a los usuarios realizar las consultas ad-hoc, pero eso se deja para trabajos futuros.

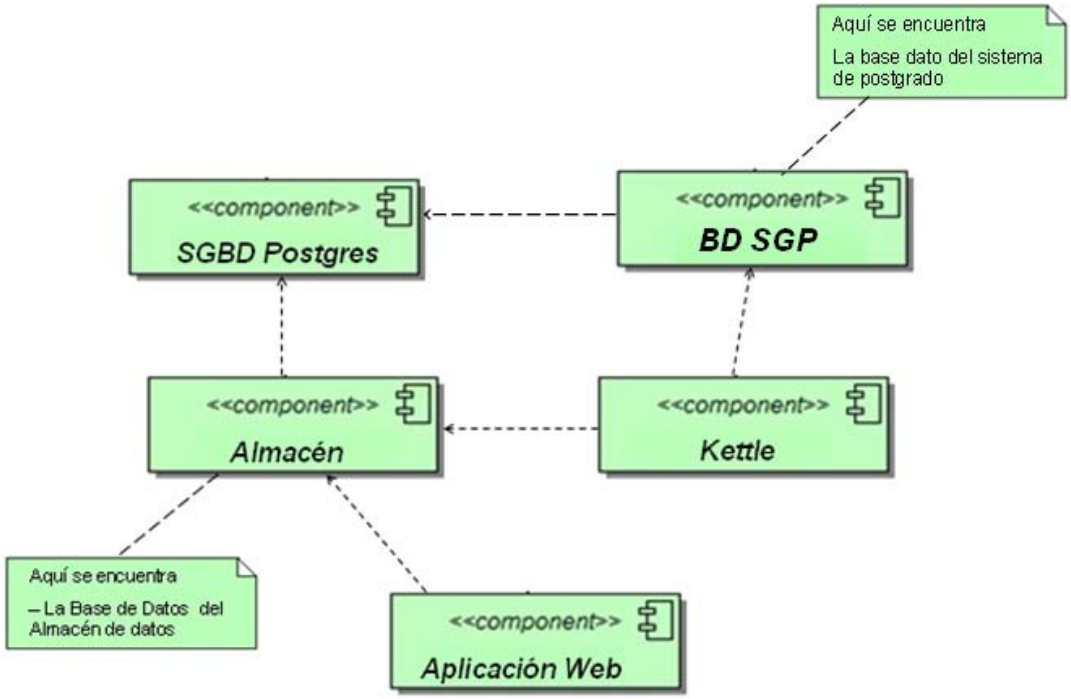

Figura 15: Diagrama de Componentes.

2.6 Diseño del Almacén de Datos.

El modelo multidimensional **MD** es el fundamento de los almacenes de datos; en él los datos se estructuran mediante hechos y dimensiones. Los hechos representan el objeto de análisis en el proceso de toma de decisión y contienen, generalmente, medidas que representan los elementos específicos de análisis. Las dimensiones permiten explorar las medidas desde diferentes perspectivas de análisis.

En nuestro almacén el objeto de análisis es la participación tanto del estudiante como la del profesor en el postgrado, por lo que tendríamos dos tablas de hechos. Para el diseño del almacén se decidió utilizar el esquema Constelación de hechos figura 17,

ya que está compuesto por dos estrellas en las cuales las tablas de hechos estudiantes y profesores comparten varias tablas de dimensiones como son: la dimensión tiempo, la dimensión tipo_pg, la dimensión postgrado, la dimensión área y la dimensión zona.

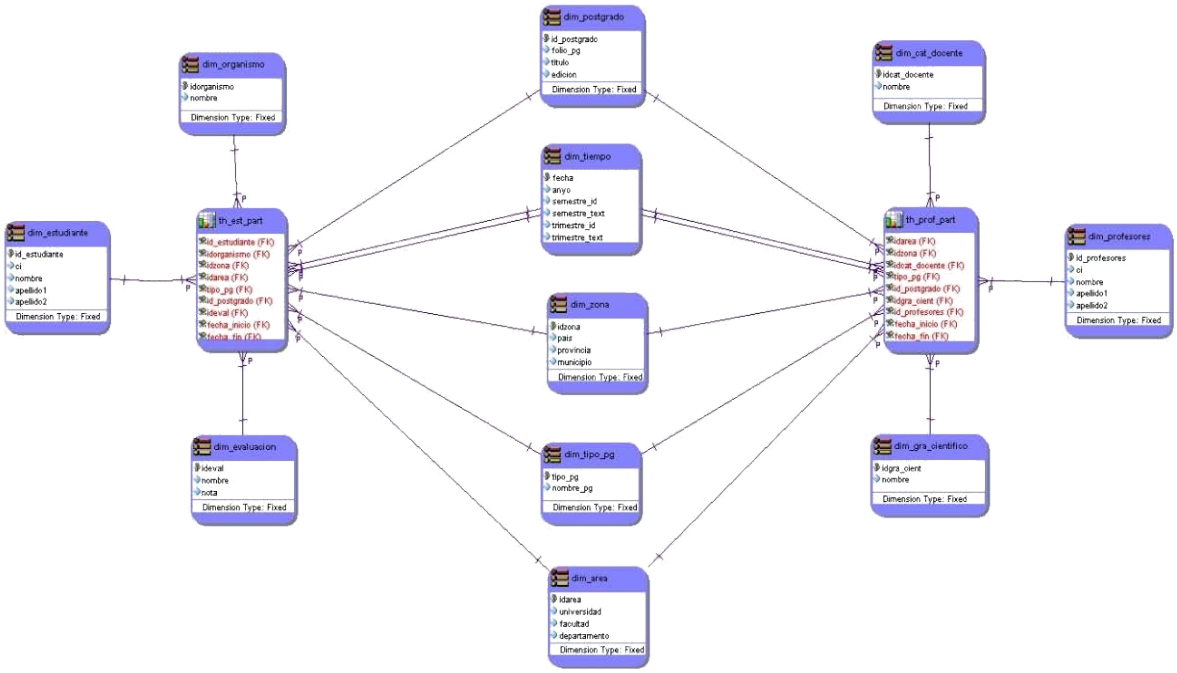

Figura 17: Constelación de Hechos.

A continuación veremos la estrella asociada al objeto de análisis estudiantes figura 18.

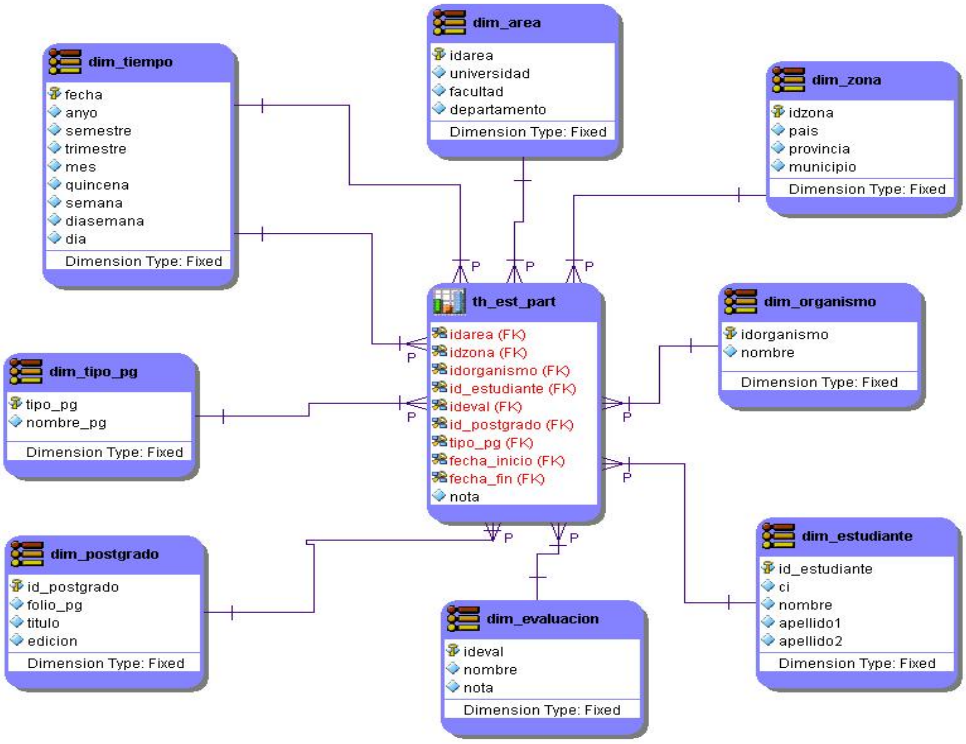

Figura 18: Estrella Estudiantes.

En la figura 18 se observa que la tabla de hechos "th_est_part" forma un cubo de ocho dimensiones: dim_tiempo, dim_area, dim_zona, dim_orgnismo, dim_estudiante, dim_evaluación, dim_postgrado y dim_tipo_pg. Además se puede apreciar que la medida de la tabla de hehos es la nota, ya que un estudiante haya terminado el curso por el cual estaba pasando lo da la nota de este.

En la figura 19 se puede apreciar la estrella asociada a los profesores.

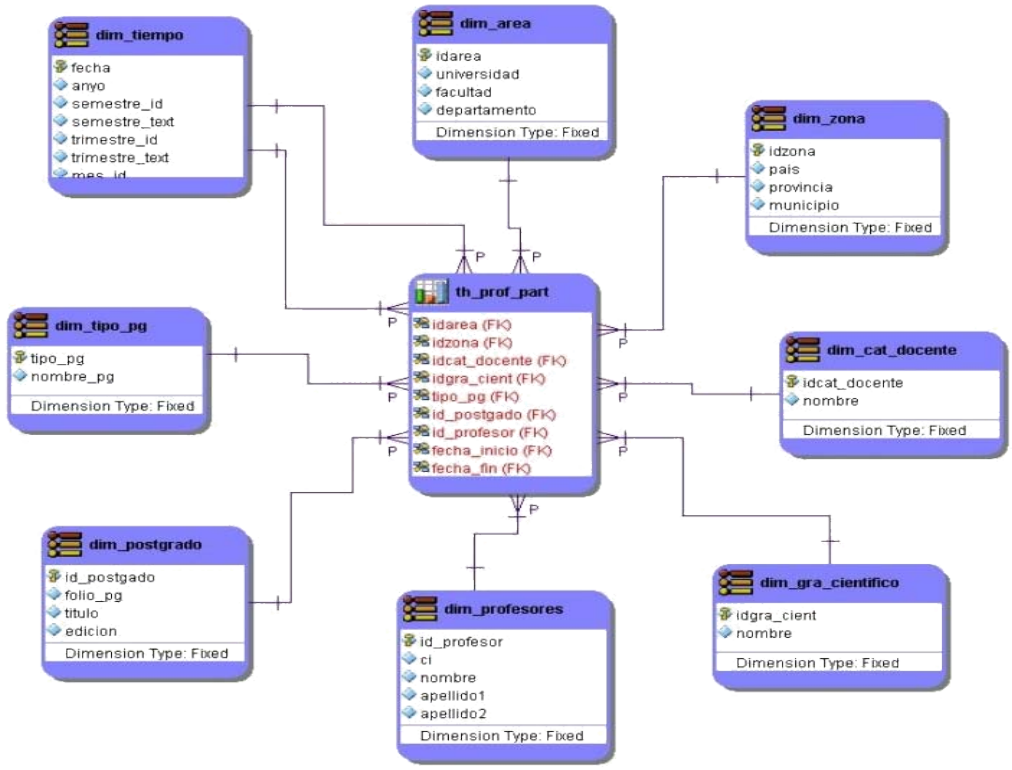

Figura 19: Estrella Estudiantes.

En la figura 19 se observa que la tabla de hechos "th_est_part" forma un cubo de ocho dimensiones: dim_tiempo, dim_area, dim_zona, dim_cat_docente, dim_gra_cientifico, dim_profesores, dim_postgrado y dim_tipo_pg. Además se puede apreciar que la medida de la tabla de hechos es la fecha en la que el profesor participó en el postgrado.

2.7 Conclusiones Parciales

Este capítulo hizo un resumen del análisis de requisitos y el diseño propuesto para la modelación del problema y dar solución a las preguntas de investigación. Para esto se hizo un diseño en UML utilizando varios tipos de diagramas. Partiendo de los casos de usos, se describen con diagramas de actividad los principales escenarios. Se emplearon los diagramas de componentes y de despliegue para documentar las partes del sistema y como se interrelacionan. El diseño del almacén se hizo mediante un modelado multidimensional para eso nos ayudamos de la herramienta Embarcadero ErEstudio.

CAPITULO 3. Modelo Físico del Almacén de Datos y Modulo ETL.

Introducción

En este capítulo veremos el modelo físico y la implementación del almacén de datos en el SGBD PostgreSQL, así como el módulo ETL correspondiente que permita transformar y almacenar la información histórica del SGP.

3.1 Modelo Físico del Almacén de Datos

Conforme a las exigencias de la Dirección de Postgrado de la UCLV con relación a la información requerida para la toma de decisiones, se ha implementado el almacén de datos de la siguiente manera:

La dimensión área es útil para visualizar el análisis en función de la universidad, de la facultad así como en función del departamento. Esta dimensión es común para las tablas de hechos estudiantes y profesores.

	Column	Datatype
1	idarea	integer
2	universidad	varchar(255)
3	facultad	varchar(50)
4	departamento	varchar(50)

Figura 20: Dimensión Área.

En la dimensión cat_docente se almacenan las categorías docentes por las cuales se puede analizar el profesor. Esta dimensión es valida solamente para el profesor.

	Column	Datatype
1	idcat_docente	int4
2	nombre	varchar(20)

Figura 21: Dimensión Categoría Docente.

En la dimensión estudiantes se almacenan toda la información que es de importancia para los directivos referente al estudiante. Esta dimensión es valida solamente para el estudiante.

	Column	Datatype
1	id_estudiante	integer
2	ci	varchar(15)
3	nombre	varchar(20)
4	apellido1	varchar(20)
5	apellido2	varchar(20)

Figura 22: Dimensión Estudiantes.

La dimensión evaluación contiene la descripción de la nota obtenida por el estudiante luego de haber pasado el postgrado. Esta dimensión es valida solamente para el estudiante

	Column	Datatype
1	ideval	int4
2	nombre	varchar(15)
3	nota	int4

Figura 23: Dimensión Evaluación.

En la dimensión gra_cientifico se almacenan los grados científicos por las cuales se puede analizar el profesor. Esta dimensión es valida solamente para el profesor.

	Column	Datatype
1	idgra_cient	int4
2	nombre	varchar(20)

Figura 24: Dimensión Evaluación.

La dimensión organismoes útil para visualizar el análisis en función del organismo al que pertenece el estudiante. Esta dimensión es valida solamente para el estudiante.

	Column	Datatype
1	idorganismo	int4
2	nombre	varchar(50)

Figura 25: Dimensión Organismo.

En la dimensión postgrado se almacena la información relevante para los directivos asociada con los diferentes tipos de postgrado, es útil para visualizar el análisis del postgrado en función del folio, del titulo y de la edición. Esta dimensión es común para las tablas de hechos estudiantes y profesores.

	Column	Datatype
1	id_postgrado	integer
2	folio_pg	varchar(10)
3	titulo	varchar(256)
4	edicion	varchar(50)

Figura 26: Dimensión Postgrado.

En la dimensión profesores se almacenan toda la información que es de importancia para los directivos referente al profesor. Esta dimensión es valida solamente para el profesor.

	Column	Datatype
1	id_profesores	integer
2	ci	varchar(15)
3	nombre	varchar(20)
4	apellido1	varchar(20)
5	apellido2	varchar(20)

Figura 27: Dimensión Postgrado.

La dimensión tipo_pgcontiene la descripción de todos los tipos de postgrado: curso, maestría, entrenamiento, etc. Esta dimensión es común para las tablas de hechos estudiantes y profesores.

	Column	Datatype
1	tipo_pg	int4
2	nombre_pg	varchar(20)

Figura 28: Dimensión Tipo Postgrado.

La dimensión zona es útil para visualizar el análisis en función del país, de la provincia así como en función del municipio. Esta dimensión es común para las tablas de hechos estudiantes y profesores.

	Column	Datatype
1	idzona	integer
2	pais	varchar(255)
3	provincia	varchar(50)
4	municipio	varchar(50)

Figura 29: Dimensión Tipo Postgrado.

La dimensión tiempoes básica para cualquier modelo, pues el tiempo siempre es una de las perspectivas por las que queremos analizar la información. Los datos que forman esta dimensión los generaremos para un periodo de tiempo determinado: del

1 de enero de 1994 al 31 de diciembre del 2020. Esta dimensión es común para las tablas de hechos estudiantes y profesores.

	Column	Datatype
1	fecha	date
2	anyo	int4
3	semestre_id	int4
4	semestre_text	char(15)
5	trimestre_id	int4
6	trimestre_text	char(15)
7	mes_id	int4
8	mes_text	char(15)
9	quincena_id	int4
10	quincena_text	char(15)
11	semana_id	int4
12	semana_text	char(15)
13	diasemana_id	int4
14	diasemana_text	char(15)
15	dia	int4

Figura 30: Dimensión Tiempo.

A continuación se puede a preciar las tablas de hechos estudiante y profesores con sus medidas y llaves foráneas respectivamente.

	Column	Datatype
1	id_estudiante	int4
2	idorganismo	int4
3	idzona	int4
4	idarea	int4
5	tipo_pg	int4
6	id_postgrado	int4
7	ideval	int4
8	fecha_inicio	date
9	fecha_fin	date
10	nota	int4

Figura 31: Tabla de Hecho Estudiantes.

	Column	Datatype
1	idarea	int4
2	idzona	int4
3	idcat_docente	int4
4	tipo_pg	int4
5	id_postgrado	int4
6	idgra_cient	int4
7	id_profesores	int4
8	fecha_inicio	date
9	fecha_fin	date

Figura 32: Tabla de Hechos Profesores.

3.2 Módulo ETL.

Para el desarrollo del módulo de ETL se utilizó la herramienta Kettle por ser Software Libre que es uno de los requerimientos del sistema. Ademáses muy intuitiva, y con unos conceptos básicos se pueden realizar las transformaciones. Conceptualmente es muy sencilla y potente.Kettle no es solo una herramienta de ETL sino que está integrada a toda una suite de inteligencia empresarial denominada Pentaho.

A continuación veremos las transformaciones realizadas en nuestro trabajo para llevar a cabo la carga en el Almacén de Datos.

Transformación de Postgrado

En la transformación asociada a la dimensión de postgrado es necesario realizar una consulta en la base de datos operativa para obtener los atributos de los diferentes tipos de postgrados que son de interés a la Dirección de Postgrado de la UCLV, luego se genera la llave primaria la cual es una llave subrogada, después se realiza una correspondencia entre los atributos seleccionados con los de la dimensión de postgrado y luego se almacenan.

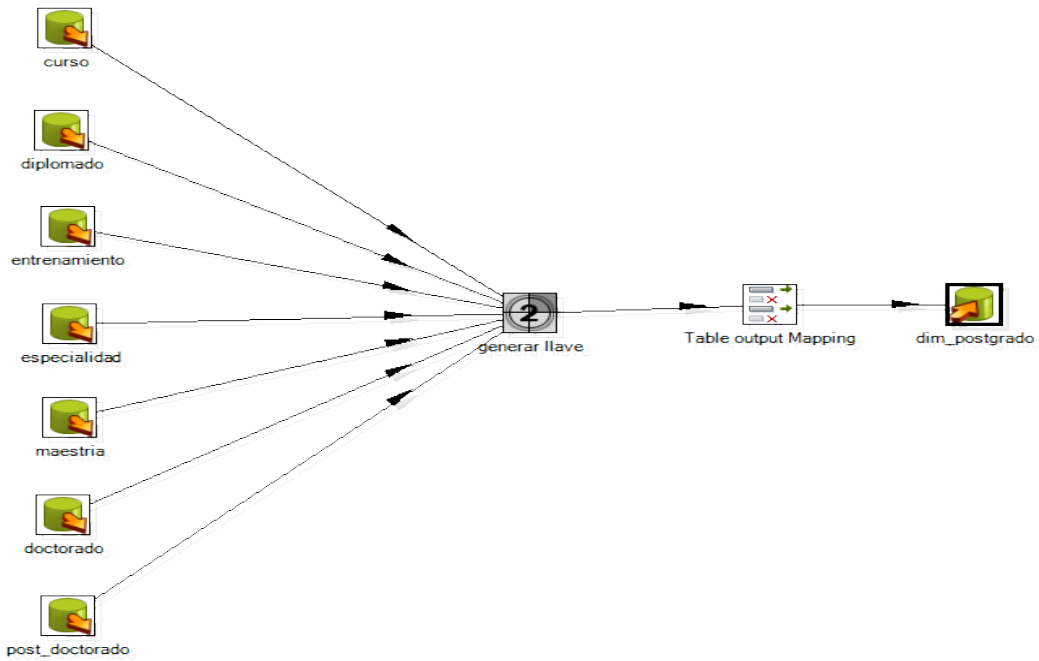

Figura 33: Transformación Postgrado.

En la figura 34 se puede observar la consulta realizada a los diferentes tipos de postgrado.

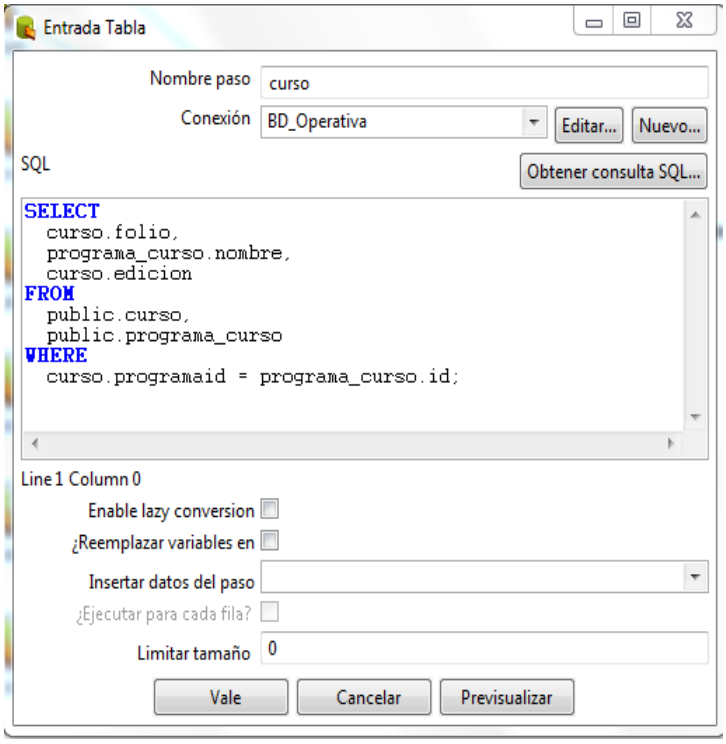

Figura 34: Transformación Postgrado.

Transformación de Área.

En la transformación de la dimensión área primeramente se hace la relación de las universidades con las facultades y los departamentos, luego se genera su llave primaria y por último se hace la correspondencia entre los atributos.

Figura 35: Transformación Área.

Esta transformación es igual a la transformación de la dimensión zona exceptuando los atributos particulares de esta.

Las siguientes transformaciones (grado científico, organismo, evaluación y categoría docente) son muy sencillas ya que se corresponden con las tablas clasificadoras de la base de datos operativa por lo cual lo único que se hizo en estos caso fue lograr la correspondencia entre los atributo

Figura 36: Transformación Grado Científico.

En el caso de las transformaciones estudiantes y profesores se selecciona de la base dato operativa los atributos que son de interés para la dirección, se genera la llave primaria y luego se hace la correspondencia de los atributos.

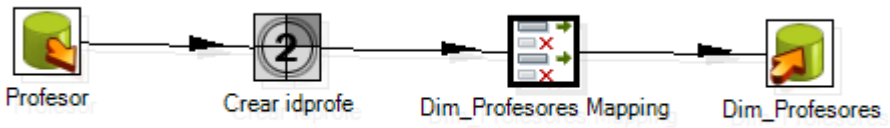

Figura 37: Transformación Profesores.

La dimensión tiempo es una dimensión estática, pues se crea una vez, y no se vuelve a tocar, a no ser que queramos añadir algún atributo adicional.La transformación va a generar todos los datos vistos para cada fecha, desde el 01 de Enero de 1994 (para los datos históricos anteriores que también cargaremos en nuestro AD) hasta el 31 de Diciembre de 2020. El diseño de nuestra transformación será el de la figura 37:

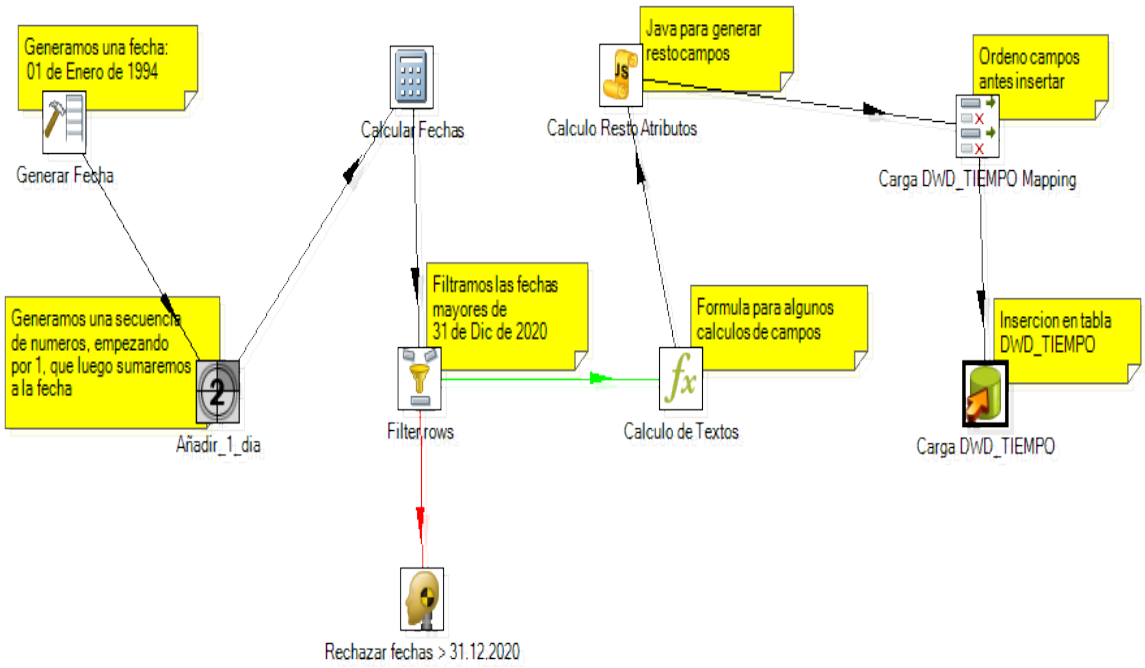

Figura 38: Transformación Tiempo.

Para la dimensión tipo_pg se decidió crear un fichero texto que contenga los nombres de los diferentes tipos de postgrados ya que en la base dato operativa no existe ninguna entidad que brinde esta información, luego se genera la llave primaria y se cargan en el almacén.

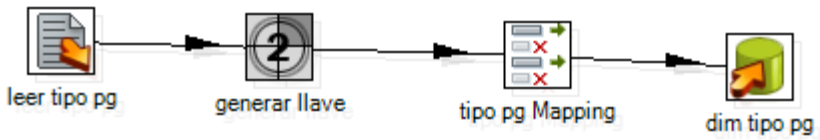

Figura 39: Transformación Tipo Postgrado.

Para la transformación de la tabla de hechos estudiantes y la transformación de la tabla de hechos profesores es necesario primeramente obtener la información asociada a cada uno de ellos por el transcurso de los diferentes tipos de postgrados. Esta consulta se realiza por separado para cada postgrado ya que en la base de dato

operativa ya que la información de los estudiantes y profesores en relación con el postgrado esta por separado por cada uno de ellos, luego se realiza una búsqueda de la informacion obtenida con la consulta por las diferentes dimensiones obteniendo asi las llaves primarias de cada dimensión para de esta forma almacenarla en la tabla de hechos.

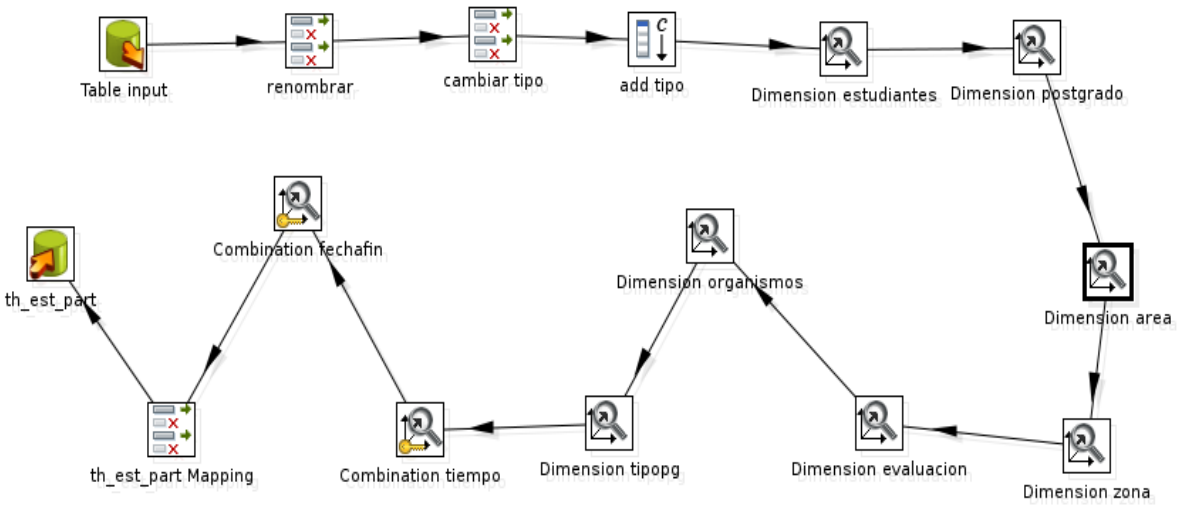

Figura 40: Transformación Tipo Postgrado.

En la figura 41 se puede observar la consulta realizada para obtener los datos de los estudiantes similares para los profesores.

Figura 41: Transformación Tipo Postgrado.

Conclusiones parciales

En este capítulose concluyó la exposición del modelo físico del almacén de datos así como el módulo ETL que llevan a cabo las transformaciones necesarias para almacenar la información histórica del SGP.

Conclusiones

- Se creó mediante el modelo dimensional una estructura para el almacén de datos asociado a las actividades de postgrado.
- Se implementó el mismo sobre el SGBD PostgreSql para compatibilidad con el proyecto SIGENU.
- Se desarrolló las transformaciones necesarias para llevar la carga del almacén dejándolo listo para el análisis.

Recomendaciones

Se presentan las siguientes recomendaciones:

1. Implementar una interfaz Web que permita recuperar datos desde el almacén, permitiendo consultas ad-hoc.
2. Extender la implantación del almacén a todos los CES del país.
3. Continuar el perfeccionamiento del almacén teniendo en cuenta los cambios y las necesidades de la dirección de Postgrado.

Referencias Bibliográficas

BATINI, C., CERI, S., & NAVATHE, Conceptual Database Design. An Entity-Relationship Approach Benjamin/Cummings Publishing, 1992.

CARPANI, F., CMDM: A conceptual multidimensional model for Data Warehouse. *Master Thesis* . Universidad de la República, Uruguay, 2000.

FINLAY, P. N.,*Introducing decision support systems.* Oxford, UK Cambridge: Blackwell Publishers, 1994.

GLORIA PONJUAN, M. M. Sistemas de Informacion: Principios y Aplicaciones.

GUPTA, H. HARINARAYA, V. RAJARAMAN, A. & ULLMAN,J. Index Selection for OLAP.Proceeding ICDE '97. 1997.

HARINARAYA, V., RAJARAMAN, A. & ULLMAN, J.D. Implementing data cubes efficiently. ACM SIGMOD Record, 25(2): 205--216. 1996.

INMON, W.H, *Building the Data Warehouse* (Fourth Edicción ed.). Indianapolis, Indiana: Wiley, 2005.

JHON D. PORTER and JHON J. ROME, The Data Warehouse: 2 year late, Lesson Learned, Arizona State University, 1994.

KENAN, T., An Introduction to Multidimensional Databases. *White Paper, Kenan Technologies,* 1996 .

KIMBALL, R.,*The Data Warehouse Toolkit.*(J. Wiley, & Son, Edits.), 1996.

LAUDON, JANE y KENNETH, Sistemas de información gerencial- Administración de la empresa digital.Pearson Educación- Prentice Hall, 2006.

LUJAN MORA, S., Diseño de Almacenes de Datos con UML. *Resumen para la admisión de Tesis Doctorales en lenguas no formales,* 2005.

PATRIC ZIEGLER and KLAUS R. PITTRICH, Three Decades of Data Integration – All problem solved?, WCC2004, 3-12, 2004.

SHIRLEY BECKER, Libro “Data Warehousing and Web Engineering”, página 20, idea Group publishing, ISBN 1931777020, 2003.

TURBAN, E., *Decision support and expert systems.* (m. s. systems, Ed.) Englewood Cliffs, N.J.: Prentice Hall, 1995.

Bibliografía

1. Arnold, K., J. Gosling, and D. Holmes, *Java (TM) Programming Language, The*. 2005: Addison-Wesley Professional.
2. Date, C., *Introduccion a los Sistemas de Bases de Datos*. 2001: Pearson Publications Company.
3. Douglas, K., *PostgreSQL*. 2005: Sams Indianapolis, IN, USA.
4. Kimball, R., M. Ross, and R. Merz, *The data warehouse toolkit: the complete guide to dimensional modeling*. 2002: Wiley.
5. Momjian, B., *PostgreSQL: introduction and concepts*. Vol. 192. 2001: Addison-Wesley.
6. Senn, J.A., E.G.U. Medal, and O.A.P. Velasco, *Analisis y diseno de sistemas de informacion*. 1992: McGraw-Hill.
7. Serrano, M., et al., *Metricas de calidad para almacenes de datos.*
8. Simitsis, A., P. Vassiliadis, and T. Sellis, *Optimizing ETL processes in data warehouses.*2005.
9. Trujilla, M.P.J.C., *Diseno de Almacenes de Datos*. 2002.
10. Villarroel, R., et al., *Un profile de UML para disenar almacenes de datos seguros.*
11. Wolff, C.G., *La Tecnología Datawarehousing.* Ingeniería informática, 1999(3): p. 4.