Universidad Autónoma de Querétaro

Facultad de Informática

"IMPLEMENTACIÓN DE UN DATA MART PARA EL CONTROL DE ASPIRANTES, MATRICULADOS Y EGRESADOS DE LA UNIVERSIDAD AUTÓNOMA DE QUERÉTARO".

TRABAJO DE INVESTIGACIÓN

Que para obtener el TÍTULO de:
INGENIERO EN COMPUTACIÓN

Presenta:
JOSÉ JOAQUÍN AGUILAR GUERRERO

Asesor:
M.S.I. Gerardo Rodríguez Rojano

Santiago de Querétaro, Qro., Septiembre 2009

Resumen

El presente trabajo de investigación describe el análisis, diseño e implementación de un conjunto de Data Marts que servirán de apoyo para el proceso de toma de decisiones de la Universidad Autónoma de Querétaro.

La implementación de los Data Marts permitirá conocer de forma oportuna, centralizada y de fácil acceso, información referente a los aspirantes y alumnado de la Universidad, por ejemplo, el total de aspirantes a cualquier facultad, el número de aspirantes aceptados por periodo, total de aspirantes por sexo y grupo de edad, etc. Además, ayudarán a la creación de datos históricos (si no existen), los cuales se utilizaran para realizar comparaciones con periodos pasados, y de esta manera obtener la tendencia de los diferentes indicadores.

Por lo general, los registros almacenados en las Bases de Datos no tienen utilidad alguna si no son transformados en información que sea la base para tomar decisiones. Por tal motivo es necesario que todos los datos históricos sean sometidos a un proceso de limpieza para poder garantizar su integridad y confiabilidad.

Para abordar lo anterior, en primer lugar, se realiza una revisión teórica de los diversos modelos conceptuales, la definición del proceso de extracción-transformación-carga de datos (ETL) y de las herramientas utilizas para la realización de este tipo de sistemas. En cuanto a las herramientas, la investigación se centra en software Open Source, aún así, se mencionan algunas características de las herramientas con licencia de paga; al terminar de revisar la teoría se explica el procedimiento para el desarrollo de los Data Marts.

En la implementación del presente trabajo de investigación se realizaron casi todos los pasos de un proyecto de Inteligencia de Negocios (Business Intelligence): Análisis de las fuentes de datos, creación de los procesos ETL, diseño y construcción de los Data Marts. Los Data Marts creados son la materia prima para elaborar una gran cantidad de informes que pueden ser presentados a las personas encargadas de la toma de decisiones de la Institución.

Dedicatorias y agradecimientos

Este trabajo de investigación está dedicado de manera muy especial a mis padres y hermanos, que en todo momento me brindaron su apoyo y confianza.

Al M.S.I. Gerardo Rodríguez Rojano, por asesorarme a lo largo del presente trabajo de investigación.

A la Dirección de Innovación y Tecnologías de Información de la Universidad Autónoma de Querétaro, por brindarme la oportunidad de llevar a la práctica los conocimientos adquiridos a lo largo de mi formación personal y académica.

A mis profesores y compañeros, por compartir sus conocimientos y experiencias que ayudaron a mi formación profesional.

Índice

Índice de figuras

Índice de tablas

CAPÍTULO 1. INTRODUCCIÓN

1.1 Título del trabajo de investigación

Implementación de un *Data Mart* para el control de aspirantes, matriculados y egresados de la Universidad Autónoma de Querétaro.

1.2 Definición del problema

La sociedad e industria queretana esta en un cambio constante, por lo que sus necesidades académicas han evolucionado conforme al paso de los años. Estos cambios han sido graduales e imperceptibles que solo se pueden apreciar con la comparación entre diversos intervalos de tiempo. En consecuencia las obligaciones de la Universidad hacia la sociedad han cambiado, por tanto es relevante conocer estos cambios y tendencias para seguir cumpliendo con los compromisos como institución educativa.

El estudio e implementación de un Data Mart para la Universidad Autónoma de Querétaro, brindará información acerca de las necesidades que tiene la propia institución, así como la sociedad e industria queretana, pues con los datos contenidos en el Data Mart se podrá obtener información de aspirantes, carreras, alumnos y egresados de la Universidad por periodo escolar. Asimismo, al contar con los datos históricos, se puede hacer una comparación con periodos pasados, obteniendo la tendencia de estos indicadores, dichas tendencias son la base para la toma de decisiones, además permitirán a los administrativos tomar acciones mejor fundamentadas para satisfacer las necesidades de la sociedad.

Este proyecto, con la documentación que genere, será la base fundamental para la futura creación de un Data Warehouse (DW) para la Universidad. Este DW no solo brindara el estudio de mercado de los aspirantes y alumnos, sino también se podrá tener estadística de costos, rendimiento académico de los alumnos, por mencionar algunos.

Además al analizar los registros de la Base de Datos permitirá conocer cuáles son sus inconsistencias, por que se generan, como se pueden evitar y en un determinado caso corregirlos.

1.3 Objetivos

1.3.1 Objetivo general

Desarrollar un prototipo de un *Data Mart* para la automatización de indicadores sobre el control de aspirantes, matriculados y egresados de la Universidad Autónoma de Querétaro.

1.3.2 Objetivos específicos

- Revisar los principales conceptos sobre *Business Intelligence* y *Data Warehouse*.
- Explorar las técnicas de diseño y construcción de un *Data Mart*.
- Seleccionar, a partir de los indicadores propuestos por la Dirección de Innovación y Tecnologías de la Información, los datos que formarán parte del *Data Mart*.
- Implementar un Data Mart con herramientas *Business Intelligence Open Source*.
- Elaborar un informe de los resultados obtenidos por el sistema y del uso de la herramienta BI.

1.4 Hipótesis

La implantación de un Data Mart permite concentrar y organizar los datos de diferentes fuentes para tomar mejores decisiones y no por meras especulaciones, ahorrando tiempo de procesamiento y no afectando a los sistemas transaccionales de la institución.

1.5 Organización del documento

La elaboración del presente trabajo de investigación se divide en los siguientes capítulos:

Capítulo 2:

Objetivo: Describir los aspectos teóricos de Business Intelligence.

Descripción: Se realiza un estudio sobre el contexto y los aspectos teóricos de Business Intelligence. Se presentan las etapas que se deben cumplir para que el sistema sea considerado como una solución de Business Intelligence. Estos conceptos son la base para comprender básico de este trabajo de investigación.

Capítulo 3:

Objetivo: Definir de manera clara y precisa los conceptos y características de un Data Mart.

Descripción: En esta sección se define el concepto de Data Warehouse y/o Data Mart. La investigación realizada en este apartado describe detalladamente cada uno de los elementos que conforman al Data Mart. Se dan a conocer los diferentes modelos para su implementación, las características y sus fases de desarrollo.

Capítulo 4:

Objetivo: Describir el proceso de extracción, transformación y carga de datos.

Descripción: Los elementos de Business Intelligence de mayor importancia son la creación del Data Warehouse y/o Data Mart (Capítulo 3) y las metodologías de extracción, transformación y carga de datos. Por tanto, el Capítulo 4 se refiere a la descripción de las herramientas y técnicas ETL. También se mencionan algunos ejemplos del software que se utiliza para realizar este tipo de tareas.

Capítulo 5:

Objetivo: Mencionar el apoyo brindado por parte de la institución y describir las características del software y hardware utilizado.

Descripción: Se menciona el apoyo que se tuvo por parte de la Universidad Autónoma de Querétaro, además de señalar la misión y visión de dicha Institución. También se describen las características físicas (hardware) y lógicas (software) utilizadas a lo largo del proyecto.

Capítulo 6:

Objetivo: Describir la metodología para la creación de un Data Mart para la Universidad Autónoma de Querétaro.

Descripción: Es el pilar de la investigación, en este capítulo se desarrolla la metodología y los resultados obtenidos en el proyecto. Describe los indicadores utilizados, el software a utilizar, etc. Además se realiza un análisis de los datos de entrada, dicho análisis incluye una serie de filtros que permiten conocer aquellos que no cumplen con un criterio establecido, y por tanto, son datos con algún tipo de inconsistencia y deben ser tratados para poder ser aceptados en las dimensiones o tablas de hecho de los Data Marts.

También explica la creación de los diferentes Data Marts, sus características principales y el contenido de cada uno. Por último, se plantean una serie de imágenes que describen el proceso de migración de datos hacia los Data Mart.

Capítulo 7:

Objetivo: Presentar las conclusiones logradas.

Descripción: Se presentan las conclusiones y recomendaciones a las que se llegaron después de la elaboración del proyecto.

CAPÍTULO 2. BUSINESS INTELLIGENCE

2.1 Introducción

El brindar información que sea útil en el análisis y fundamento para la toma de decisiones no es tarea sencilla, mucho menos cuando se tienen gran cantidad de datos de forma dispersa, para lograrlo es necesario utilizar herramientas y metodologías que ayuden a acceder a cada una de las fuentes de datos (sin importar los diferentes formatos y fuentes heterogéneas de las que provienen), importarlos e integrarlos en un repositorio diseñado especialmente para su posterior estudio [20].

Es importante resaltar que dichas herramientas, aparte de mostrar los datos, permitan hacer análisis sobre ellos y, por tanto, proceder a tomar decisiones [1], siendo este el principal objetivo del proceso de almacenamiento tan meticuloso de los datos [7]. Los datos deben ser mostrados de forma interactiva y sencilla a las personas encargadas de la toma de decisiones dentro de la organización.

A lo anterior se le denomina *Business Intelligence* (BI), y desde un punto de vista más pragmático, y asociándolo directamente con las tecnologías de la información, podemos definir BI como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la compañía) en información estructurada, para su explotación directa (*reporting*, análisis OLTP u *OnLine Transaction Processing* / OLAP u *On-Line Analytical Processing*, alertas, etc.) o para su análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio [9][6].

Las aplicaciones BI han evolucionado vertiginosamente y en muchas direcciones, tanto en la necesidad de acceso a los distintos datos que existen en las compañías, como en el crecimiento exponencial que éstos han tenido. Desde los informes operacionales a los modelos estadísticos para campañas publicitarias, los ambientes multidimensionales de OLAP para analistas, los tableros de control para ejecutivos; las compañías comenzaban a demandar más opciones de reporte y análisis de datos. La expansión de almacenamiento de datos, combinada con la adopción extendida de aplicaciones empresariales, tales como los *Enterprise resource planning* (ERP) y *Customer relationship management* (CRM), así como el aumento en la cantidad de usuarios capaces de utilizar una comparadora, abasteció la demanda exponencial para reportes BI y aplicaciones analíticas [21].

Hasta el momento, la mayoría de las empresas líderes en su ramo han comprado diversas herramientas BI a distintos proveedores; cada herramienta enfocada a una nueva aplicación de BI [21]. Esto no significa que empresas medianas y pequeñas no puedan aprovechar las ventajas de BI, ya que existe software con licencia *Open Source* y no genera ningún costo extra.

2.2 Elementos de Business Intelligence

En general un modelo integral de una solución de BI se describe en cuatro fases:

2.2.1 Fuentes de datos (Obtención de datos):

Las fuentes de datos utilizadas en el desarrollo de aplicaciones BI se dividen en internos (Datos Operacionales) y externos [16]. Los internos son todos aquellos datos que la organización tiene almacenados en sus sistemas DB2, ORACLE, INFORMIX, SQL Server, ERP, EDI, CRM, OLTP, Hojas de cálculo, etc.; las fuentes de datos externos son textos, documentos *HyperText Markup Language* (HTML), etc., no pertenecientes directamente a la organización.

Dichos datos se encuentran normalmente de forma heterogénea y en diferentes fuentes y formatos, en modelos estructurados o simples archivos, para utilizar estos datos en un modelo de BI es necesario darles un formato homogéneo, cuidando su integridad.

2.2.2 Herramientas de ETL:

El proceso *Extract*, *Transform* y *Load* (ETL) es el responsable de identificar, extraer y dar formato a los datos de mayor relevancia de los sistemas transaccionales, adaptando y sincronizando los datos de distintas fuentes y plataformas tecnológicas [23][27]; además se depuran y preparan (homogeneización de los datos), respetando los esquemas de seguridad e integridad de la fuente, para cargarlos en un Data Warehouse (DW) o *Data Mart* según el diseño preestablecido [3][17]; dichas herramientas son las responsables de que el proceso de constitución del DW se realice con éxito.

- *Extract*: Extraer, es la obtención de los datos de las distintas fuentes tanto internas como externas. Ésta etapa convierte los datos a un formato preparado para iniciar el proceso de transformación. Además verifica si los datos extraídos cumplen con la estructura preestablecida, de no ser así, los datos son rechazados [27].

- *Transform*: Transformar, es el filtrado, limpieza, depuración, homogeneización y agrupación de los datos. Algunas de las transformaciones más utilizadas son:
  - Seleccionar solo ciertas columnas,
  - Obtener nuevos valores calculados,
  - Unir datos de múltiples fuentes,
  - Dividir una columna en varias.

- *Load*: Carga, es la organización y actualización de los datos y metadatos[1] en la base de datos destino [28].

Se estima que el proceso ETL consume un 70% a 80% de tiempo y esfuerzo en la construcción del DW o *Data Mart* [12] [11].

Hoy en día existen un sinfín de productos ETL en el mercado, pero cuando se desea seleccionar alguno se deben considerar los siguientes puntos [16]:

- Base de Datos soportadas
- Capacidad de procesamiento
- Adquisición de datos en tiempo real
- Metadatos
- Ambiente de desarrollo
- Plataformas soportadas
- Precio

2.2.3 Data Warehouse (DW)

Un DW o bodega de datos es una colección de datos, orientados a hechos relevantes del negocio, integrados, que incluyen el tiempo como característica importante de referencia y no volátiles para el proceso de toma de decisiones [14]. Este subtema se abordara a fondo en el *Capítulo 3*.

2.2.4 Herramientas para la toma de decisiones[2]

- Cubos OLAP: Este término fue presentado en 1993, publicado por Codd y asociados y apoyados por Arbor Software Corporation, compañía que creó ESSBASE una de las primeras herramientas OLAP que aparecen en el mercado, adquirida luego por Hyperion Software [5]. Procesamiento Analítico en Línea (*On-Line Analytical Processing*) realiza análisis de datos empresariales y proporciona la capacidad para cálculos complejos, análisis de tendencias, etc., su uso facilita respuestas rápidas a consultas analíticas complejas [24].

[1] Los Metadatos se comportan como un índice del contenido dentro del DW. Normalmente los Metadatos almacenan la definición de las tablas, columnas, vistas y cualquier otro objeto dentro del DW o Data Mart [22].

[2] El estudio de las herramientas para la toma de decisiones, si bien se basan en los datos almacenados en el Data Mart, queda fuera del alcance de este trabajo de investigación.

Para poder navegar entre los datos, OLAP ofrece un conjunto de operadores que facilitan la concepción de consultas, entre ellos están *Slice & Dice, Swap, Drill Down, Drill Up, Roll Up, Drill Across, Drill-Through* [4]. Existen tres tipos de cubos OLAP, Relational OLAP, Multidimensional OLAP y Hybrid OLAP, su estudio queda fuera del alcance de este proyecto.

- Data Mining: Es un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos [8], y desde el punto de vista empresarial, se define como la integración de un conjunto de áreas, que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos, que aporten bases para la toma de decisiones.

Con la ayuda de Data Mining es posible identificar tendencias y comportamientos, extraer información e identificar patrones de comportamiento [26].

CAPÍTULO 3. DATA WAREHOUSE Y/O DATA MART

3.1 Data Warehouse

Un Data Warehouse, Data Mart o Almacén de Datos es una colección de datos históricos específicos de un sistema de bases de datos transaccional y operacional, ofrece una vista del estatus de la empresa; se trata de una tecnología que permite desarrollar un análisis multidimensional de los datos, y que además, favorece la organización de la información tomando en cuenta ciertos parámetros establecidos, sobre los que se lleva a cabo un análisis de información [15].

El DW se crea a partir de la de extracción, transformación y carga de datos (Sistemas ETL), dicho proceso no afecta a las bases de datos operacionales. Los datos contenidos en el almacén se quedan a disposición del usuario final para su consulta, favoreciendo el análisis y la divulgación eficiente de los mismos. El DW se apoya en herramientas OLAP, Data Mining y herramientas administrativas, las cuales tienen la capacidad de crear información con el objetivo de mejorar la toma de decisiones en la organización; un DW debe entregar la información correcta a la gente indicada en el momento óptimo y en el formato adecuado [13][29].

Un DW no se puede comprar, se tiene que construir [18]. La construcción e implementación de un DW es un proceso evolutivo e iterativo. Este proceso surge de la definición del objetivo del DW, los requerimientos que se planean satisfacer, el diseño y modelado del DW, la implementación y revisión del mismo [25]. Además, con ayuda de los DW los sistemas operacionales se pueden deslindar de la responsabilidad de representar el pasado de la organización.

3.2 Data Mart

El Data Mart al igual que el DW es un sistema orientado a la consulta, donde se producen procesos de carga de datos (altas), muy rara vez se eliminan registros. Es consultado mediante herramientas OLAP que ofrecen una visión multidimensional de la información. Sobre estas bases de datos se pueden construir Sistemas de Información para Directivos (EIS) y Sistemas de Ayuda a la toma de Decisiones (DSS).

Cuenta con las mismas características del DW, algunos autores determinan la creación de un Data Mart como algo consecuente del DW con la finalidad de ofrecer una mejor presentación de los datos. Hay otros autores que parten de la creación de Data Marts para formar el DW porque el conjunto de los Data Marts formara al DW. Esto varía de acuerdo a las necesidades de la organización debido a que los resultados del DW se desarrollaran a largo plazo y los del Data Mart, a consecuencia de ser de un área específica del negocio, son a corto plazo. [13][14][27].

La diferencia entre el DW y Data Mart es el menor coste para la gestión y explotación eficiente de los datos y una oportunidad para empezar a familiarizarse con esta nueva tecnología.

***Nota**: En adelante se tomara el termino Data Mart como un sinónimo del Data Warehouse, porque, como ya se menciono, poseen las mismas características pero con diferente alcance.*

3.3 Características de Data Mart

Orientado al tema: Un Data Mart está diseñado para ayudar al análisis de información de la empresa, el tema de cada organización influye en el diseño del Data Mart. Por ejemplo: venta, inventario, selección de aspirantes, etc...

Integrado: Se refiere a la integración de los datos de las diferentes fuentes. Los datos tienen que ser consistentes, de no ser así, pasan por un proceso de transformación, lo que garantiza su integridad dentro del Data Mart.

De tiempo variante o temporal: Los datos almacenados están referidos a un período de tiempo específico, por ejemplo, el día, la semana, el mes, semestre, etc.

No Volátil: Una vez almacenados los datos en el Data Mart no deberían ser modificados [14][27].

3.4 Metadatos

Los Metadatos permiten al usuario tener una mayor visión de la información. El usuario puede explorar el Data Mart y averiguar qué datos están allí y qué otros no [13][30]. Los Metadatos se comportan como un índice del contenido dentro del Data Mart, en general en los Metadatos se almacenan:

- La estructura de datos como los ve el programador.
- La estructura de datos como los ve el analista DSS.
- Datos de origen que alimentan al Data Mart.
- La transformación y carga de datos cuando pasan al Data Mart.
- Las vistas definidas, índices, etc.
- La relación entre las dimensiones y tablas de facto del Data Mart.
- Las políticas de seguridad para acceder o ingresar datos.

3.5 Arquitectura del Data Mart

La *arquitectura* de Data Mart se divide en cuatro fases (Figura 3.1):

3.5.1 Sistemas Operacionales:

Es la materia prima del Data Mart, en estos sistemas se encuentra los registros de cada operación transaccional de la organización, se debe considerar que se encuentra fuera del Data Mart, en esta parte se tiene un control limitado del formato de los datos. Si estos sistemas cuentan con una integridad en sus datos, la construcción del Data Mart será mucho más sencilla, de lo contrario la elaboración del Data Mart será más lenta.

3.5.2 Área de depuración de los datos:

Se hace la integración, unificación y limpieza de los datos que vienen de los diferentes sistemas operacionales.

- El primer paso es la *extracción* en esta fase se pretende comprender los datos de origen.
- El segundo paso es la *transformación* de los datos, donde se da un formato preestablecido a los datos.
- Por último, la *carga* de datos en estructuras normalizadas, aquí se depositan los datos transformados para la disposición del usuario. Ver capítulo 4.

3.5.3 Presentación de los datos:

Es el área donde se almacenaron los datos ya organizados. Se encuentran en esquemas dimensiónales (cubos) normalmente llamados Data Marts, Éste modelo dimensional, es una estructura simple para presentar los datos en *dimensiones* y *tablas de hecho*, permitiendo un mejor rendimiento en las consultas realizadas por el usuario.

Estos datos, al estar interrelacionados mediante dimensiones, permiten al usuario tener un mayor nivel de detalle en las consultas que realice, favoreciendo una mejor interpretación de la información obtenida.

3.5.4 Herramientas de acceso a los datos:

Existe un sin número de herramientas que pueden aprovechar el área de presentación de los datos, éstas pueden variar en cuanto a su complejidad, puede ser una herramienta de consulta o de extracción. Aproximadamente del 80 a 90 por ciento de los usuarios

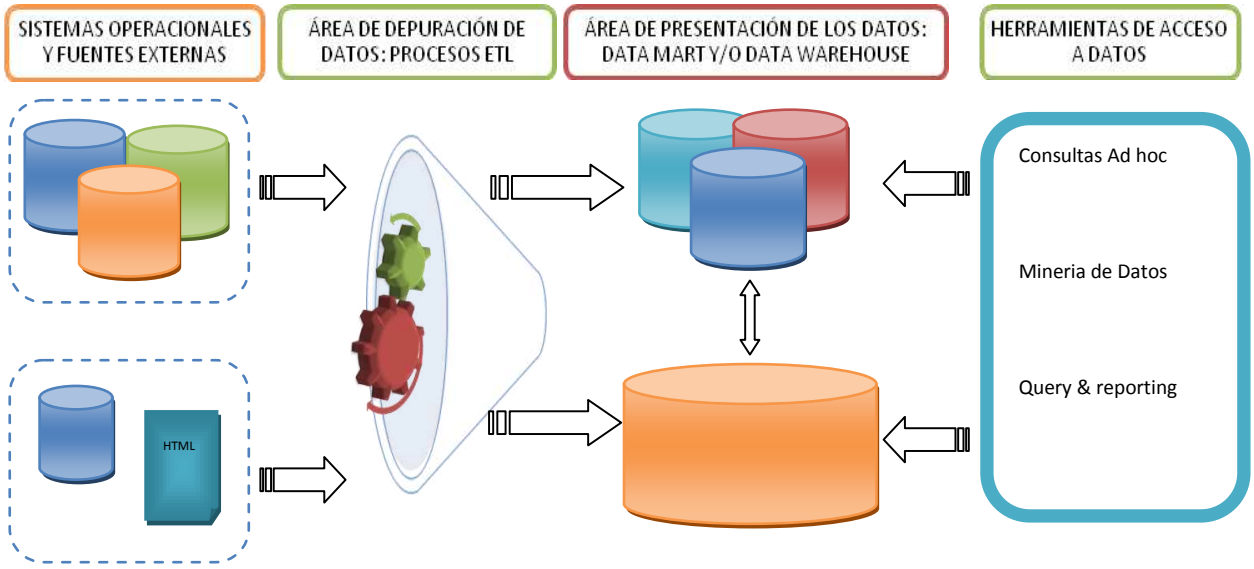


Figura 3.1 Arquitectura de un DW y/o Data Mart.

potenciales usaran herramientas de consultas que no son más que plantillas y por tantp, no exigen a los usuarios construir las consultas relacionales directamente. Algunas de las herramientas de acceso a la información más sofisticadas realizan actividades como planear o prever; las herramientas variarán según las necesidades y gustos del usuario [14][19][13].

3.6 Diseño lógico de un Data Mart

Existen dos principales estrategias para la metodología del diseño de un Data Mart, las cuales surgen de acuerdo a las necesidades de la organización:

Metodología descendente (top-down), considera la creación del Data Warehouse como base del proyecto y después la creación de los Data Marts [15] a partir del DW, el tiempo de desarrollo del proyecto es muy amplio para tener resultados tangibles, algunas organizaciones consideran esta estrategia como una inversión a largo plazo (Figura 3.2).

Metodología ascendente (bottom-up), considera primeramente la creación de Data Marts, y con la unión de todos los Data Marts se forma el DW [13]. Esta metodología tiene una gran ventaja con respecto al tiempo ya que se pueden tener resultados por áreas del negocio, las cuales son representados por cada uno de los Data Marts (Figura 3.3).

La decisión sobre cual metodología utilizar depende del tipo de organización (institución educativa, bancarias, gubernamentales, hospitales, etc.), de los recursos a invertir en el desarrollo del Data Mart y/o DW y en qué tiempo se desean los resultados.

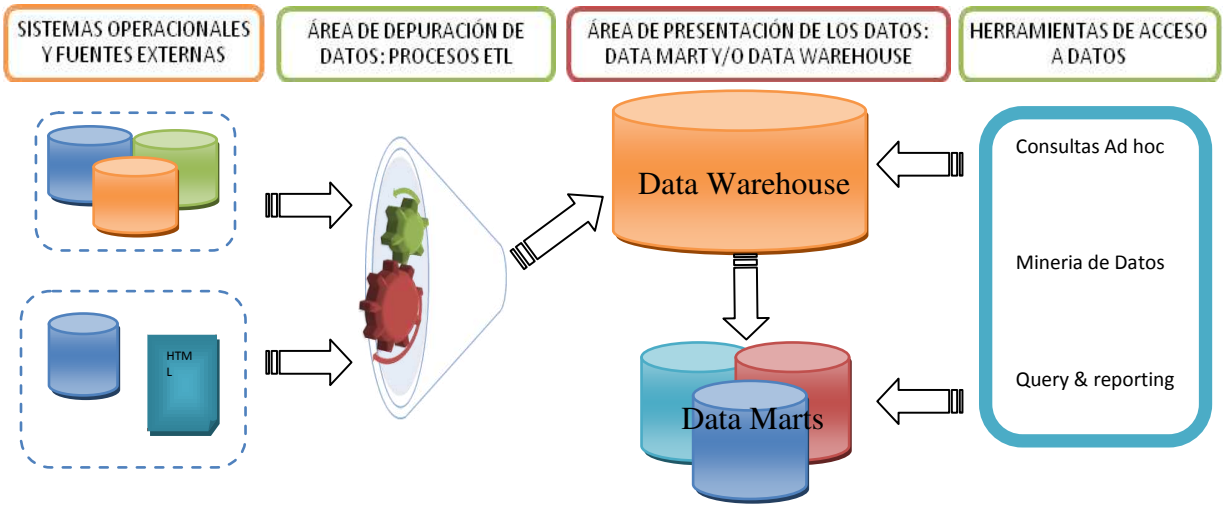


Figura 3.2 La arquitectura top-down describe la creación de un Data Warehouse ya partir de éste crear un conjunto de Data Marts.

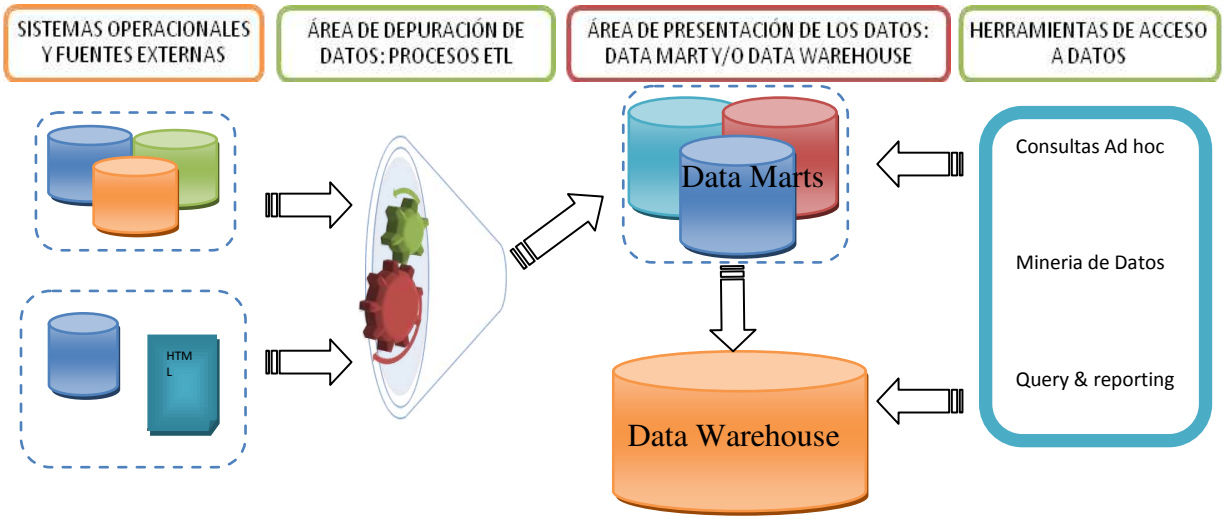


Figura 3.3 La arquitectura bottom-up describe la creación de un conjunto de Data Marts, formando así el Data Warehouse.

Independiente de la metodología a utilizar, el primer paso para la creación de un Data Mart es definir su uso y las necesidades que cubrirá (análisis de requerimientos); una vez establecido esto se analiza la información con la que se va a trabajar y además se propone la estructura de la base de datos a implementar.

Dicha estructura se desarrolla por medio de un Análisis Multidimensional o hipercubos; esto es, en lugar de buscar unidades atómicas de información como entidades, atributos y todas las relaciones entre ellos, los datos se organizan alrededor de los temas de la organización (ventas, control de inventarios, control de aspirantes, etc.), Figura 3.4. Un hipercubo consiste en un conjunto de celdas, cada una se identifica por la combinación de los miembros de las diferentes dimensiones y contiene el valor de la medida analizada para dicha combinación de dimensiones. Se caracteriza por la presencia de una tabla central normalizada, llamada tabla de facto o hecho (fact table) y un conjunto de tablas pequeñas, generalmente desnormalizadas y llamadas dimensiones, las cuales contienen las descripciones de las características de los datos [13][24][31].

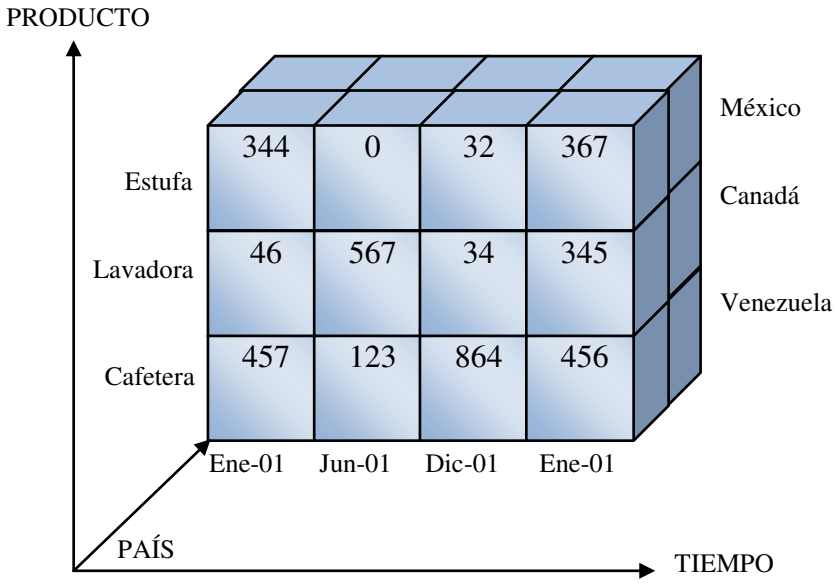


Figura 3.4 Hipercubo sobre las ventas de productos en diversos periodos y por diferentes países.

El Análisis Multidimensional ofrece dos cualidades sobre la información, la primera consiste en profundizar sobre el tema, es decir poder ir de lo general a lo particular, y la segunda es la posibilidad de ver la información desde varios ángulos.

El Análisis Multidimensional combina los datos provenientes de distintas áreas de la organización en múltiples dimensiones, y así ubicar cierto tipo de información importante que revele el comportamiento de la institución o negocio. Por ejemplo para observar las cifras de un negocio; se analiza la cifra global de ingresos, pero se necesita bajar hasta el nivel de detalle, es decir logrando ver los ingresos de cada uno de los vendedores o por cada uno de los productos, o por cada una de las semanas transcurridas, de esta manera se visualiza con mayor detalle el comportamiento de una determinada variable [1][13][31].

Adicionalmente, al analizar las cifras desde diferentes perspectivas, se puede obtener la información de ventas por producto, por geografía, por tiempo, por vendedor, entre otros aspectos, Figura 3.5.

Al identificar las necesidades de información de la organización, se logran resolver las preguntas de negocios estratégicas, tácticas u operativas. Aunado a esto, se logra conocer el comportamiento de la organización, para poder anticiparse y tomar decisiones adecuadas y oportunas [1].

De la definición anterior surgen los siguientes conceptos:

3.6.1 Fact Table

Tablas de facto: Contiene dos tipos de atributos, uno con las medidas o valores numéricos y el segundo se refiere a las llaves foráneas de las tablas de dimensión. Sus medidas se obtienen generalmente por la aplicación de una función estadística que resume un conjunto de valores en un único valor. Por ejemplo: ventas, inventarios, cantidad de productos vendidos, horas trabajadas, etcétera [1][2][13][31].

3.6.2 Dimensional Table

Tablas de dimensión: Representan cada uno de los ejes en un especio multidimensional. Son elementos que contienen atributos (o campos) que se utilizan para restringir y agrupar los datos almacenados en una tabla de facto, es decir, suministran el contexto en el que se obtienen las medidas de un hecho. Las dimensiones se utilizan para seleccionar y agrupar los datos en un nivel de detalle deseado [1][2][31].

3.6.3 Jerarquías

Las jerarquías son estructuras lógicas que utilizan niveles como un medio de organizar los datos. Una jerarquía puede usarse para definir la agregación de los datos. Por ejemplo, en una dimensión temporal, una jerarquía podría agregar los datos del día, mes o año.

Dentro de una jerarquía, cada nivel se conecta lógicamente a los niveles sobre y debajo de él. Los valores de datos de los niveles bajos se agregan en los valores de datos a los niveles más altos. Una dimensión puede componerse de más de una jerarquía. Las herramientas de consulta dan uso de las jerarquías para obtener búsquedas con diferente granularidad, Figura 3.5.

Las jerarquías imponen una estructura familiar en valores de la dimensión. Para un valor nivelado particular, un valor al próximo nivel más alto es su padre y los valores al próximo más bajo nivel son sus hijos. Estas relaciones familiares les permiten a analistas tener acceso a los datos rápidamente [1][13].

- N*ivel:* Representa una posición en una jerarquía. Por ejemplo, una dimensión temporal podría tener una jerarquía que representa los datos del día, mes, y año. Los niveles van del general al específico, con el nivel de la raíz como el más alto.

- *Relaciones de niveles*: Detallan la raíz para adquirir más información específica. Definen la relación del padre-hijo entre los niveles de una jerarquía. Las jerarquías también son los componentes esenciales más complejos. Por ejemplo, la Base de Datos puede agregar una renta de ventas existente, calculando los valores trimestralmente, bajo una agregación anual, cuando las dependencias dimensiónales como el año y los valores trimestrales del mismo son conocidos, podemos obtener las rentas anuales, Figura 3.5.

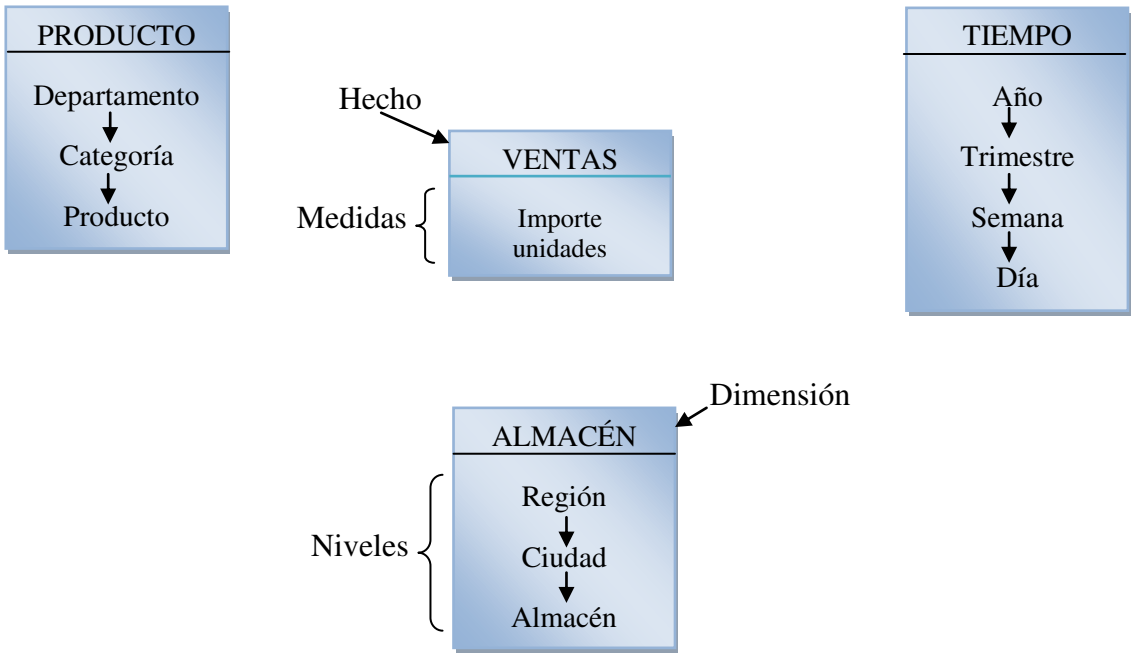


Figura 3.5 Modelo multidimensional con diferentes jerarquías y dimensiones.

3.7 Esquemas multidimensionales

Los esquemas de la colección de objetos de la Base de Datos, incluso de las tablas, vistas, índices y sinónimos, se pueden organizar de una gran variedad de maneras dentro del modelo general del Data Mart. La mayoría de los almacenes de datos usan un modelo dimensional [24][14].

Los modelos multidimensionales consisten en un conjunto de celdas, cada una se identifica por la combinación de los miembros de las diferentes dimensiones y contiene el valor de la medida analizada para dicha combinación de dimensiones [24]. Las formas de representarlo se describe a continuación:

3.7.1 Esquema en estrella

Es un modelo de datos que tiene una tabla de hechos (o tabla fact) que contiene los datos para el análisis, rodeada de las tablas de dimensiones.

Un esquema en estrella perfecciona el rendimiento guardando las consultas simples y proporcionando el tiempo de la contestación rápido. Toda la información sobre cada nivel se almacena en una fila (figura 3.6).

Este esquema es ideal por su simplicidad y velocidad para ser usado en análisis multidimensionales (OLAP, Data Marts, EIS). Permite acceder tanto a datos agregados como de detalle. El diseño de esquemas en estrella permite implementar la funcionalidad de una base de datos multidimensional utilizando una clásica base de datos relacional (más extendidas que las multidimensionales) [27].

Otra razón para utilizar los esquemas en estrella es su simplicidad desde el punto de vista del usuario final. Las consultas no son complicadas, ya que las condiciones y las uniones (*JOIN*) necesarias sólo involucran a la tabla de hechos y a las de dimensiones, no haciendo falta que se encadenen uniones y condiciones a dos o más niveles como ocurriría en un esquema en copo de nieve [24][27].

3.7.2 Modelo de copo de nieve

Surge cuando alguna de las dimensiones tiene más de una tabla de datos, es decir, esta normalizada. Al normalizar las tablas se elimina la redundancia de datos; pero tiene la contrapartida de generar peores rendimientos al tener que crear más tablas de dimensiones y más relaciones entre las tablas (*JOINS*) lo que tiene un impacto directo sobre el rendimiento. En éste tipo de esquemas se tiene una tabla central de hechos en la que se guardan las medidas del negocio que se desean analizar y en las tablas adyacentes se tendrán las dimensiones (parámetros) [27][13].

El único argumento a favor de los esquemas en copo de nieve es que al estar normalizadas las tablas de dimensiones, se evita la redundancia de datos y con ello se ahorra espacio. Pero, si tonemos en cuenta que, hoy en día, el espacio en disco no suele ser un problema y si el rendimiento se presenta con una mala opción en Data Mart, ya que el hecho de disponer de más de una tabla por cada dimensión, implica tener que realizar código más complejo para realizar una consulta que a su vez se ejecutará en un tiempo mayor, debido en parte al mayor número de uniones (*JOINS*) que habrá que realizar. Se puede usar un esquema de copo de nieve en un Data Mart, aunque éstos sean realmente grandes y complejos, pero nunca en sistemas donde el tiempo de respuesta sea un factor crítico para los usuarios [27]. Ver figura 3.7

3.7.3 Modelo de constelación

Es una variación del esquema de estrella tradicional, en este modelo algunos atributos de las dimensiones se separan formando una nueva entidad que puede ser compartida con otros cubos. La utilidad principal de este modelo es que al tener dimensiones que pueden ser compartidas por diferentes cubos se tendrá un mejor uso del espacio de almacenamientos evitando la redundancia [20]. Figura 3.8.

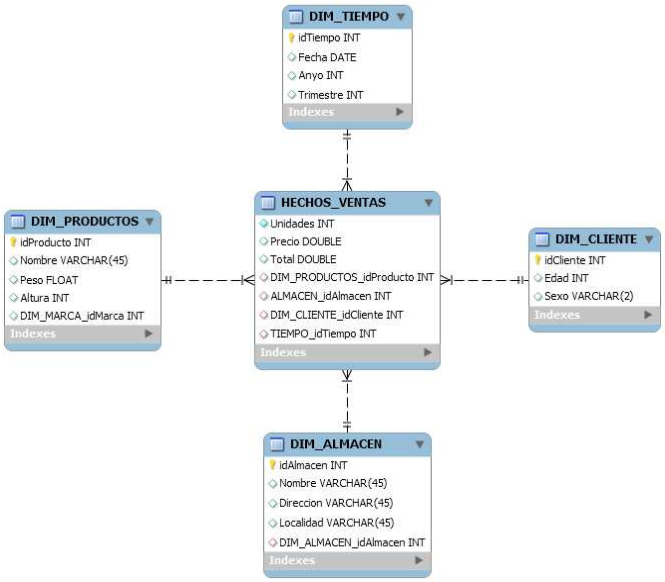


Figura 3.6 Modelo en Estrella con cuatro dimensiones y una tabla de hechos o facto.

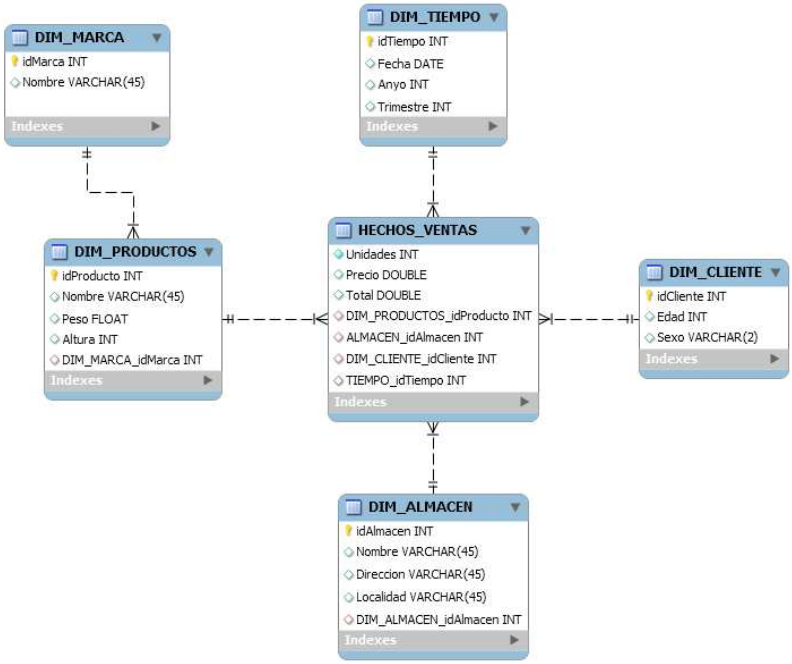


Figura 3.7 Modelo de copo de Nieve

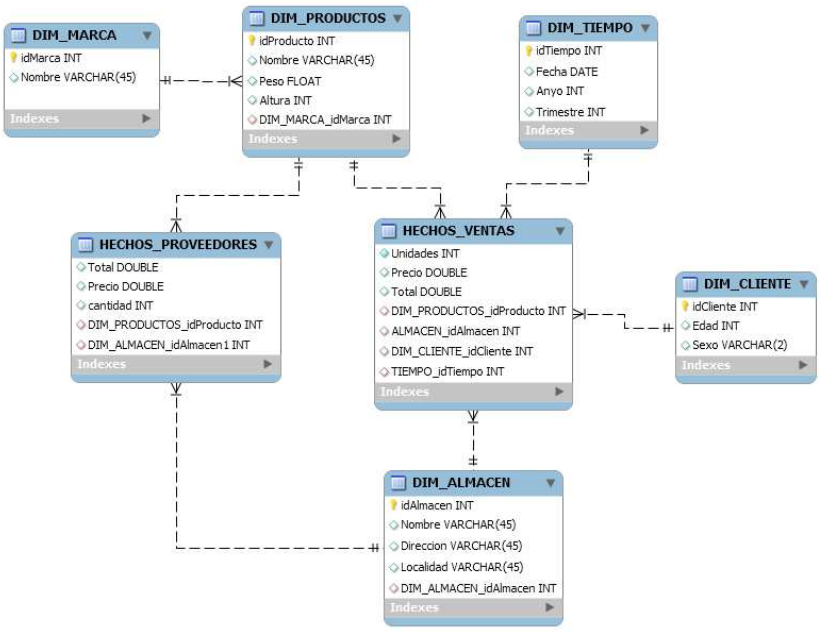


Figura 3.8 Modelo de Constelación

CAPÍTULO 4. HERRAMIENTAS Y METODOLOGÍAS ETL

4.1 Herramientas y Metodologías

Son los elementos más importantes y de valor añadido de la arquitectura Business Intelligence, ya que absorben más del 70% del desarrollo del proyecto, teniendo como principal objetivo transferir datos desde diferentes sistemas transaccionales a los sistemas de Business Intelligence [27][30].

Las principales características de este tipo de herramientas son:

- *Accesibilidad a la información*: Los datos son la fuente principal de este concepto. Lo primero que deben garantizar éste tipo de herramientas y técnicas será el acceso de los usuarios a los datos con independencia de la procedencia de estos.

- *Apoyo en la toma de decisiones*: Se busca ir más allá en la presentación de la información, de manera que los usuarios tengan acceso a herramientas de análisis que les permitan seleccionar y manipular sólo aquellos datos que les interesen.

- *Orientación al usuario final*: Se busca independencia entre los conocimientos técnicos de los usuarios y su capacidad para utilizar estas herramientas.

4.2 Herramientas ETL

El proceso *Extract*, *Transform* y *Load* (ETL) es el responsable de identificar, extraer y dar formato a los datos de mayor relevancia de los sistemas transaccionales, adaptando y sincronizando los datos de distintas fuentes y plataformas tecnológicas [23][27]; además se depuran y preparan (homogeneización de los datos), respetando los esquemas de seguridad e integridad de la fuente, para cargarlos en un DW o *Data Mart* según el diseño preestablecido [3][17]; dichas herramientas son las responsables de que el proceso de desarrollo del Data Mart se realice con éxito.

Se puede decir que, el análisis, diseño y la implementación de estas herramientas es:

- *Crítica*: Porque el resto de las fases del proyecto se alimentan de ella y además los siguientes procesos no se pueden llevar a cavo hasta concluir satisfactoriamente este proceso.

- *Difícil pero no imposible*: Porque conlleva el extraer e integrar datos de diversas fuentes y plataformas muchas veces heterogéneas y acceder a

información contenida en sistemas que no están concebidos ni diseñados para las exigencias de un proceso masivo de análisis de datos.

- *Ágil*: Porque es una fase que permite adquirir una visión profunda y extendida de la institución.

- *Especifica*: Porque los datos a extraer serán utilizados para mejorar la toma de decisiones y se tienen que ajustar a criterios de contenido, calidad y formato.

4.2.1 Extract (Extracción)

Es la obtención de los datos de las distintas fuentes, tanto internas como externas. Esta etapa verifica que datos extraídos cumplan con la estructura preestablecida, de no ser así, los datos son rechazados [27][30]. Es paso incluye:

- Leer los datos de origen,
- Acceso y conectar a los datos,
- Capturar los cambios en los datos y
- Extraer los datos.

4.2.2 Transform (Transformación)

Filtrado, limpieza, depuración, homogeneización y agrupación de los datos. Algunas de las transformaciones más usuales son:

- Seleccionar solo ciertas columnas,
- Obtener nuevos valores calculados,
- Unir datos de múltiples fuentes,
- Dividir una columna en varias.
- La limpieza de datos, consiste en procesar los datos eliminando los atributos que sean erróneos o redundantes,
- Datos incorrectos o inconsistentes, ocurre cuando las base de datos no están bien modelados y permiten el ingreso de cualquier valor,
- Conversión de un tipo de dato a otro. Por ejemplo se podría cambiar el formato EBCDIC a ASCII; un dato de tipo decimal a float; y otras como cambiar de “M” a “Masculino” [27][30].

4.2.3 Load (Carga)

Organización y actualización de los datos y metadatos en la base de datos destino. Existen dos formas de desarrollar el proceso de carga:

- *Acumulación simple*: Es la más usual y consiste en realizar un resumen de todas las transacciones ocurridas durante un lapso de tiempo y transportar solo una transacción, es decir, una sumatorio y/o promedio.

- *Rolling*: Se aplica solo en casos donde se desea tener un grado de granularidad alto, para ello se almacenan los datos de forma resumida en diferentes niveles, correspondientes a distintas jerarquías (Ejemplo: totales diarios, totales semanales, etc.).

Esta fase interactúa directamente con la base de datos destino, la cual, al recibir los datos, aplicará todas las restricciones y triggers que tenga definidos. Los triggers y las restricciones contribuyen, en gran medida, a que se garantice la calidad de los datos en el proceso ETL [27][30].

En general, la idea es que una aplicación ETL lea los datos primarios de unas bases de datos de sistemas principales, realice transformación, validación, el proceso cualitativo, filtración y al final escriba datos en el almacén y en este momento los datos son disponibles para analizar por los usuarios.

Las herramientas con mayor aceptación en el mercado son:

- IBM Websphere DataStage (anteriormente Ascential DataStage y Ardent DataStage)
- Pentaho Data Integration (Kettle ETL) - Una herramienta Open Source Business Intelligence
- SAS ETL Studio
- Oracle Warehouse Builder
- Informatica PowerCenter
- Cognos Decisionstream
- Ab Initio
- BusinessObjects Data Integrator (BODI)
- Microsoft SQL Server Integration Services (SSIS)

Para mayor información es estas herramientas ver Anexo A.

CAPÍTULO 5. CASO DE ESTUDIO

5.1 Caracterización del área donde se participo

Institución:

Universidad Autónoma de Querétaro.

Dirección:

Centro Universitario, Cerro de las Campanas S/N, Santiago se Querétaro, Qro., México. C.P. 76010.

Misión de la Institución:

Impartir educación universitaria de calidad, en sus distintas modalidades en los niveles medio superior y superior; formar profesionales competitivos al servicio de la sociedad; llevar a cabo investigación humanística, científica y tecnológica, generadora de bienestar y progreso en su ámbito de influencia; difundir y extender los avances del humanismo, la ciencia, la tecnología y el arte, contribuir en un ambiente de participación responsable, apertura, libertad, respeto y crítica propositiva al desarrollo al logro de nuevas y mejores formas de vida y convivencia humana.

Visión de la Institución:

La UAQ es una institución de educación superior con pertinencia social, financieramente viable, que centra la atención en la formación de sus estudiantes para asegurar su permanencia y su desarrollo integral, con programas educativos reconocidos por su buena calidad. Genera y aplica el conocimiento, forma recursos humanos en investigación, con cuerpos académicos consolidados, integrados en redes de colaboración a nivel nacional e internacional; con procesos de gestión, eficaces y eficientes, contribuyendo a la preservación y difusión de la cultura, estrechamente vinculada con los diferentes sectores de la sociedad, promoviendo la pluralidad y libertad de pensamiento.

Apoyo institucional:

Dirección de Innovación y Tecnologías de Información:

- Director: M.C. Luis Fernando Saavedra.
- Administrador de la Base de Datos: L.I. Jesús Martín Jaramillo Morales.

Departamento de Finanzas:

- L.I. Maribel Noguez Mondragón.

Departamento de Servicios Escolares:

- L.I. Janet Carolina Ordoñez Loyola.

Desarrolladores:

Aguilar Guerrero José Joaquín

Moctezuma Martínez Homero Emmanuel (colaborador)

5.2 Recursos Físicos y lógicos:

- Computadora Personal
- Servidor
- Conexión a Internet

Hardware y software utilizado:

Especificaciones Técnicas y software de la Computadora

Marca: COMPAQ

Monitor 17''

Memoria RAM: 512 MB

Disco Duro: 80 GB

Procesador: Pentium III

Velocidad del Procesador: 1.6 GHz.

Unidad Lectora de DVD: 1

SOFTWARE:

Sistema Operativo: Debian 4 Etch

Base de Datos: MySQL Server 5.0, MySQL Client 5.0, MySQL Administrator

Pentaho: KETTLE->Spoon

Especificaciones Técnicas y software del Servidor

Marca: SUN MICROSYSTEM

Modelo: E250

Monitor 15''

Memoria RAM: 1024 MB

Disco Duro: 2 con 18 GB c/u.

Procesador: Ultra SPARC II

Velocidad del procesador: 400 MHz.

Unidad Lectora de CD: 1

SOFTWARE

Sistema Operativo: Ubuntu SPARC 7.10

Base de Datos: MySQL Server 5.0, MySQL Client 5.0

CAPÍTULO 6. METODOLOGÍA Y RESULTADOS

Los pasos a seguir para la elaboración del Data Mart se presentan a continuación:

- ❖ Análisis de Requerimientos
- ❖ Selección e instalación de Software
- ❖ Análisis de la base de datos origen
- ❖ Diseño de las tablas de hechos y dimensiones
- ❖ Migración de los datos a utilizar (ETL)

6.1 Análisis de Requerimientos

Para describir los requerimientos que la institución demanda primero se presenta una breve explicación sobre el manejo de indicadores en los Data Mart.

6.1.1 Indicadores[3]

- ❖ **¿Qué es un indicador?**

No existe una definición oficial por parte de algún organismo nacional o internacional, sólo algunas referencias que los describen como: "Herramientas para clarificar y definir, de forma más precisa, objetivos e impactos (...) son medidas verificables de cambio o resultado (...) diseñadas para contar con un estándar contra el cual evaluar, estimar o demostrar el progreso (...) con respecto a metas establecidas, facilitan el reparto de insumos, produciendo (...) productos y alcanzando objetivos" [34].

Una de las definiciones más utilizadas por diferentes organismos y autores es la que Bauer dio en 1966: "Los indicadores son estadísticas, serie estadística o cualquier forma de indicación que nos facilita estudiar dónde estamos y hacia dónde nos dirigimos con respecto a determinados objetivos y metas, así como evaluar programas específicos y determinar su impacto" [27][9].

A continuación se presenta una descripción de los indicadores utilizados en el proyecto.

[3] Estos indicadores se diseñaron para uso específico del Departamento de Innovación y Tecnología de Información, quedando prohibido el uso en cualquier otro lugar y/o proyecto.

Aspirantes:

Los aspirantes, son personas que pretenden ingresar a la Universidad y que para lograrlo, han realizado el proceso necesario para tal fin, por ejemplo: inscribirse al curso propedéutico (donde así se requiera) y entregar la documentación que acredite los requisitos solicitados.

Frecuencia: Cada ciclo escolar (periodo)

Distribución UAQ: Periodo, Campus-Plantel, Nivel, Esc-Fac., Plan de Estudios

Dist. por Variables: Sexo, Edad, Entidad de Nacimiento, Entidad de Estudios del Nivel Anterior

Conteo: Por Periodo.

Aceptados (Nuevo Ingreso):

Los aspirantes que han acreditado el proceso de nuevo ingreso, incluido el examen de admisión, obteniendo en él el puntaje mínimo para ser aceptado; es considerado como aspirante aceptado.

Frecuencia: Cada ciclo escolar (periodo)

Distribución UAQ: Periodo, Campus-Plantel, Nivel, Esc-Fac., Plan de Estudios

Dist. por Variables: Sexo, Edad, Entidad de Nacimiento, Entidad de Estudios del Nivel Anterior

Conteo: Primera Inscripción (vector), Por Periodo.

Matricula:

Se considera matrícula, a los alumnos inscritos en la Universidad, para el periodo en que se realice la consulta.

Frecuencia: Cada ciclo escolar (periodo)

Distribución UAQ: Periodo, Campus-Plantel, Nivel, Esc-Fac., Plan de Estudios, Áreas o Líneas Terminales, Semestre, Año.

Dist. por Variables: Sexo, Edad

Conteo: Inscritos por Nivel.

Egreso:

Los Alumnos que han concluido con su plan de estudios, sin contrariar ninguna de las restricciones establecidas para su plan de estudios, son considerados Egresados.

Frecuencia: Cada ciclo escolar (periodo)

Distribución UAQ: Periodo, Campus-Plantel, Nivel, Esc-Fac Plan de Estudios.

Dist. por Variables: Sexo, Edad

Conteo: Egresados por Periodo

Titulados:

Los Alumnos que concluido su plan de estudios, acreditan los requisitos para obtener el Título ofrecido por su plan de estudios.

Frecuencia: Cada ciclo escolar (periodo)

Distribución UAQ: Periodo, Campus-Plantel, Nivel, Esc-Fac., Plan de Estudios.

Dist. por Variables: Sexo, Edad

Conteo: Egresados por Periodo

Los indicadores propuestos se analizaron con el administrador de la base de datos de la institución con el objetivo de conocer qué tipo de estadística se necesita obtener. Con esto se crearon las tablas de hechos y las dimensiones de los Data Mart.

En las tablas de hecho se almacenara lo referente al *conteo* descrito en cada indicador, siendo esto lo que la institución requiere conocer acerca de sus datos.

En resumen, con los datos contenidos en las tablas de hecho se podrá obtener información de aspirantes, carreras, alumnos y egresados de la Universidad por periodo escolar, por ejemplo:

- Total de aspirantes en la Universidad por sexo, promedio de edad,
- Total de aspirantes aceptados y no aceptados en la universidad,
- Total de aspirantes aceptados y no aceptados por facultad,
- Total de aspirantes por Facultad y carrera en un periodo determinado,
- Total de aspirantes por sexo y edad promedio,
- Total de aspirantes por institución de origen, etc.

Asimismo, las tablas de hecho permitirán la creación de datos históricos, los cuales ayudarán a realizar comparaciones con periodos pasados, y de esta manera obtener la tendencia de los diferentes indicadores.

Nota: Para fines del presente trabajo de investigación y por cuestiones de seguridad solo se presentará información referente a los dos primeros indicadores, además la información sobre las tablas de la base de datos origen son ficticios y solo se presentan aquellas que son necesarios para la elaboración de este documento.

6.2 Selección del Software

Al elegir el software se tomaron en cuenta los siguientes criterios:

- *Herramientas Open Source*: Se decidió utilizar software Open Source para no generar gastos extras.

- *Documentación*: Es de vital importancia, ya que sin ella no se podría conocer a fondo las propiedades de las herramientas. La documentación de las herramientas debe ser clara y especificar los pros y contras de dicho software.

- *Especificaciones técnicas*: Cantidad de datos soportados, seguridad, tiempo-rendimiento, etc.

- Facilidad de manejo

Ventajas que ofrece el software elegido:

- Reducción de costos.

- Flexibilidad en la adaptación de los productos.

- Independencia de proveedores.

- Empezar una solución con bajos costos y con gran escalabilidad.

- Multiplataforma.

Descripción general del software utilizado:

6.2.1 Sistema Operativo:

GNU/Linux es el término empleado para referirse al sistema operativo similar a Unix que utiliza como base las herramientas de sistema de GNU y el núcleo Linux. Su desarrollo es uno de los ejemplos más prominentes de software libre; todo el código fuente puede ser utilizado, modificado y redistribuido libremente por cualquiera bajo los términos de la GPL de GNU (**L**icencia **P**ública **G**eneral de GNU) y otras licencias libres [34].

Desde sus comienzos, Linux se diseñó para que fuera un sistema multi-tarea y multi-usuario. Estos hechos son suficientes para diferenciar a Linux de otros sistemas operativos más conocidos. Sin embargo, Linux es más diferente de lo que pueda imaginar. Nadie es dueño de Linux, a diferencia de otros sistemas operativos. Gran parte de su desarrollo lo realizan voluntarios de forma altruista.

Es menos probable que un sistema Linux se colapse, además tiene mejor capacidad para ejecutar múltiples programas al mismo tiempo y es más seguro que muchos otros sistemas operativos. Debido a estas ventajas, Linux es el sistema operativo que ha experimentado

mayor crecimiento en el mercado de los servidores. Últimamente, Linux está empezando a ser popular entre los usuarios domésticos y en empresas.

Las variantes de este sistema se denominan distribuciones GNU/Linux (o *distribuciones Linux*) y su objetivo es ofrecer una edición que cumpla con las necesidades de determinado grupo de usuarios.

Algunas distribuciones GNU/Linux son especialmente conocidas por su uso en servidores y supercomputadoras. No obstante, es posible instalar Linux en una amplia variedad de hardware como computadoras de escritorio y portátiles. [34]

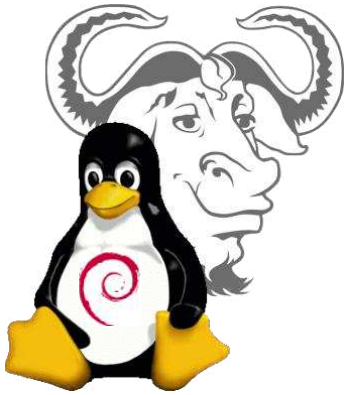


Figura 6. 1 Símbolos utilizados por GNU/Linux & Debian

Se decidió utilizar la distribución Debian 4 Etch por las siguientes ventajas:

- Estabilidad,
- Integración de paquetes,
- Documentación y soporte,
- Actualización sencillas,
- Sistema de seguimiento de errores,
- Rápido y ligero en memoria y
- Seguridad.

6.2.2 Gestor de Base de Datos:

MySQL es un sistema de gestión de base de datos relacional, multihilo y multiusuario con más de seis millones de instalaciones. MySQL AB —desde enero de 2008 una subsidiaria de Sun Microsystems y ésta a su vez de Oracle Corporation desde abril de 2009— desarrolla MySQL como software libre en un esquema de licenciamiento dual.

Por un lado se ofrece bajo la GNU GPL para cualquier uso compatible con esta licencia, pero para aquellas empresas que quieran incorporarlo en productos privativos deben comprar a la empresa una licencia específica que les permita este uso. Está desarrollado en su mayor parte en ANSI C.

MySQL Server 5.0. Es una de las utilizadas, por ello, posee gran soporte tanto extra oficial como oficial proveyendo gran cantidad de herramientas graficas para mantenimiento y administración, ejemplo de ellas es MySQL Administrator, la cual permite administrar de manera visual y remotamente todo lo que se encuentra en la Base de Datos MySQL Server. Por ejemplo, la creación de usuarios, tablas, hacer backup, etc.

6.2.3 Plataforma BI

Suite PENTAHO BI: Se define a sí mismo como una plataforma de BI orientada a la solución y centrada en procesos. Se ha concebido desde un principio para incluir los principales componentes para implementar soluciones basadas en procesos.

La plataforma Open Source Pentaho Business Intelligence cubre muy amplias necesidades de Análisis de los Datos y de Informes empresariales. Las soluciones de Pentaho están escritas en Java y tienen un ambiente de implementación también basado en Java. Eso hace que Pentaho sea una solución muy flexible para cubrir una amplia gama de necesidades empresariales – tanto las típicas como las sofisticadas y especificas al negocio.

Kettle: Es una solución que ofrece Pentaho Business Intelligence (Pentaho Data Integrator), su acrónimo es “Kettle E.T.T.L. Environment” esto significa que realiza funciones de extracción de la fuente de datos (transaccionales o externas), transformación (limpieza, consolidación) y la carga al Data Warehouse o Data Mart.

Kettle ofrece una herramienta que permite diseñar transformaciones y trabajos de manera grafica denominada Spoon.

Spoon es un componente de la Suite Pentaho BI que permite desarrollar procesos ETL de forma simple y rápida. Provee de una interfaz gráfica intuitiva. El principal objetivo de esta herramienta es ayudar a migrar los datos de Oracle a MySQL, para hacerlo cuenta con varios objetos (filtros, actualizaciones, alertas, errores, datos de entrada-salida), que se

pueden entrelazar para obtener solo los datos que realmente se necesiten con la mayor integridad y consistencia posible.

Figura 6. 2 Logotipo de Pentaho B.I.

6.2.4 TORA

Es un conjunto de herramientas multiplataforma de software libre creado para ayudar a los administradores y desarrolladores de aplicaciones de bases de datos Oracle. También suministra soporte para MySQL y PostgreSQL.

Es una herramienta muy valorada por los mismos ya que la herramienta que proporciona Oracle, el Enterprise Manager, no es demasiado intuitiva. Los desarrolladores para Oracle de Red Hat Linux afirman hacer un uso intensivo de la misma.

Independientemente del sistema operativo sobre el que se instale, necesita como paso previo que esté instalado el cliente de Oracle correspondiente a la misma o superior versión de base de datos sobre la que queremos actuar. Dicho cliente se puede instalar en su forma mínima, de modo *Runtime*.

Figura 6. 3 Logotipo del software

6.3 Análisis de la Base de Datos origen

En esta etapa se presentan las tablas y atributos utilizados para la implementación del Data Mart. Se describen las inconsistencias que se pueden encontrar en el momento de revisar el estado de los datos en la base de datos origen.

Por cuestiones de seguridad el nombre de la base de datos, de las tablas y atributos mencionados en este documento son ficticios.

6.3.1 Descripción de las Tablas utilizadas

- DM_EASPIRANTEH

Almacena los datos generales del aspirante, tal como nombre, apellidos, fecha de nacimiento, estado de nacimiento, periodo actual, etc. Solo se consideraron aquellos atributos relevantes para la estadística.

- DM_EASPIRANTED

A partir de DM_EASPIRANTEH se genera un detalle donde se almacena la clave de la carrera a la cual el aspirante pretende ingresar. Se genera cuando el aspirante realiza el pago correspondiente.

- DM_FACULTAD

Es un catalogo con todas las facultades de la Universidad. Para términos prácticos solo se eligieron dos atributos, clave de facultad y su nombre.

- DM_CARRERA

Son las carreras ofertadas en un periodo determinado. Describe el nombre, clave, periodo y nivel.

- DM_EESTADOS

La vista de DM_ESTADO contiene los estados de la república mexicana, tiene 2 atributos y una llave foránea (FK). Su identificador único (PK) es numérico y el nombre del estado es una cadena de caracteres.

- DM_EINSTITUCION

Catalogo con las instituciones de origen de los aspirantes. Los atributos relevantes de esta tabla son, nombre, nivel, clave y estado de origen.

Las tablas que a continuación se presentan muestran los atributos que conforman a cada una de las tabas mencionadas en párrafos anteriores, también describen el tipo de dato que manejan dichos atributos y para qué son utilizados en la base de datos.

VISTA DM_EASPIRANTEH	ATRIBUTO	TIPO DE DATO	DESCRIPCIÓN
 DM_EASPIRANTEH	AH_ASPIRANTE	INTEGER	PK, llave primaria de la tabla.
	AH_SEXO	VARCHAR	Permite valores F para femenino y M para masculino.
	AH_PERIODO	VARCHAR	Periodo actual.
	AH_FECNACIMIE	DATE	Fecha de nacimiento del aspirante.
	AH_FEC	DATE	Fecha de registro del aspirante.
	AH_EDONACIMIE	INTEGER	FK, Estado de nacimiento.
	AH_CLAVE_INST	VARCHAR	FK, Institución de origen.
	…		

Tabla 6. 1 Vista de la tabla DM_EASPIRANTEH de la base de datos origen.

VISTA DM_EASPIRANTED	ATRIBUTO	TIPO DE DATO	DESCRIPCIÓN
 DM_EASPIRANTED	AD_ASPIRANTE	INTEGER	FK, Relación con DM_EASPIRANTEH
	AD_FOLIO	VARCHAR	PK. Llave primaria para el detalle de aspirante.
	AD_PERIODO	VARCHAR	Periodo actual
	AD_ACEPTADO	VARCHAR	Permite dos valores S para aceptado y N para no aceptado.
	AD_OPCION	VARCHAR	Permite dos valores 1 para la primera opción y 2 para segunda.
	AD_CARR_PERI	VARCHAR	FK, Relación con DM_CARRERA.
	…		

Tabla 6. 2 Vista de la tabla DM_EASPIRANTED de la base de datos origen.

VISTA DM_FACULTAD	ATRIBUTO	TIPO DE DATO	DESCRIPCIÓN
 DM_FACULTAD	FC_CLAVE	VARCHAR	PK
	FC_NOMBRE	VARCHAR	Nombre de la Facultad.
	…		
	…		

Tabla 6. 3 Vista de la tabla DM_FACULTAD de la base de datos origen.

VISTA DM_CARRERA	ATRIBUTO	TIPO DE DATO	DESCRIPCIÓN
 DM_CARRERA	CR_CLAVE	VARCHAR	PK
	CR_CAMPUS	VARCHAR	Muestra los diferentes campus de la UAQ, por ejemplo, Amealco, Querétaro, Cadereyta, etc.
	CR_NIVEL	VARCHAR	Permite tres valores: Licenciatura, Bachillerato, Técnico
	CR_FAC	VARCHAR	FK, llave foránea a la facultad de la carrera.
	CR_EAREAS	VARCHAR	…….
	CR_NOMBRE	VARCHAR	Nombre de la carrera
	CR_PERIODO	VARCHAR	Periodo en que se oferta la carrera.
	…		
	…		

Tabla 6. 4 Vista de la tabla DM_CARRERA de la base de datos origen.

VISTA DM_EESTADOS	ATRIBUTO	TIPO DE DATO	DESCRIPCIÓN
 DM_EESTADOS	ET_CLAVE	INTEGER	PK, llave primaria del estado.
	ET_NOMBRE	VARCHAR	Nombre del estado.
	…		
	…		

Tabla 6. 5 Vista de la tabla DM_EESTADOS de la base de datos origen.

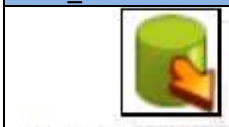
VISTA DM_EINSTITUCION	ATRIBUTO	TIPO DE DATO	DESCRIPCIÓN
 DM_EINSTITUCION	IT_CLAVE	VARCHAR	PK, llave primaria
	IT_ESTADO	INTEGER	FK, Relación DM_EESTADOS.
	IT_NOMBRE	VARCHAR	Nombre de la institución de origen
	IT_NIVEL	VARCHAR	Nivel de la institución: Preparatorio o Secundaria.
	…		
	…		

Tabla 6. 6 Vista de la tabla DM_EINSTITUCION de la base de datos origen.

El modelo Entidad-Relación de las tablas descritas se presenta en la figura 6.4.

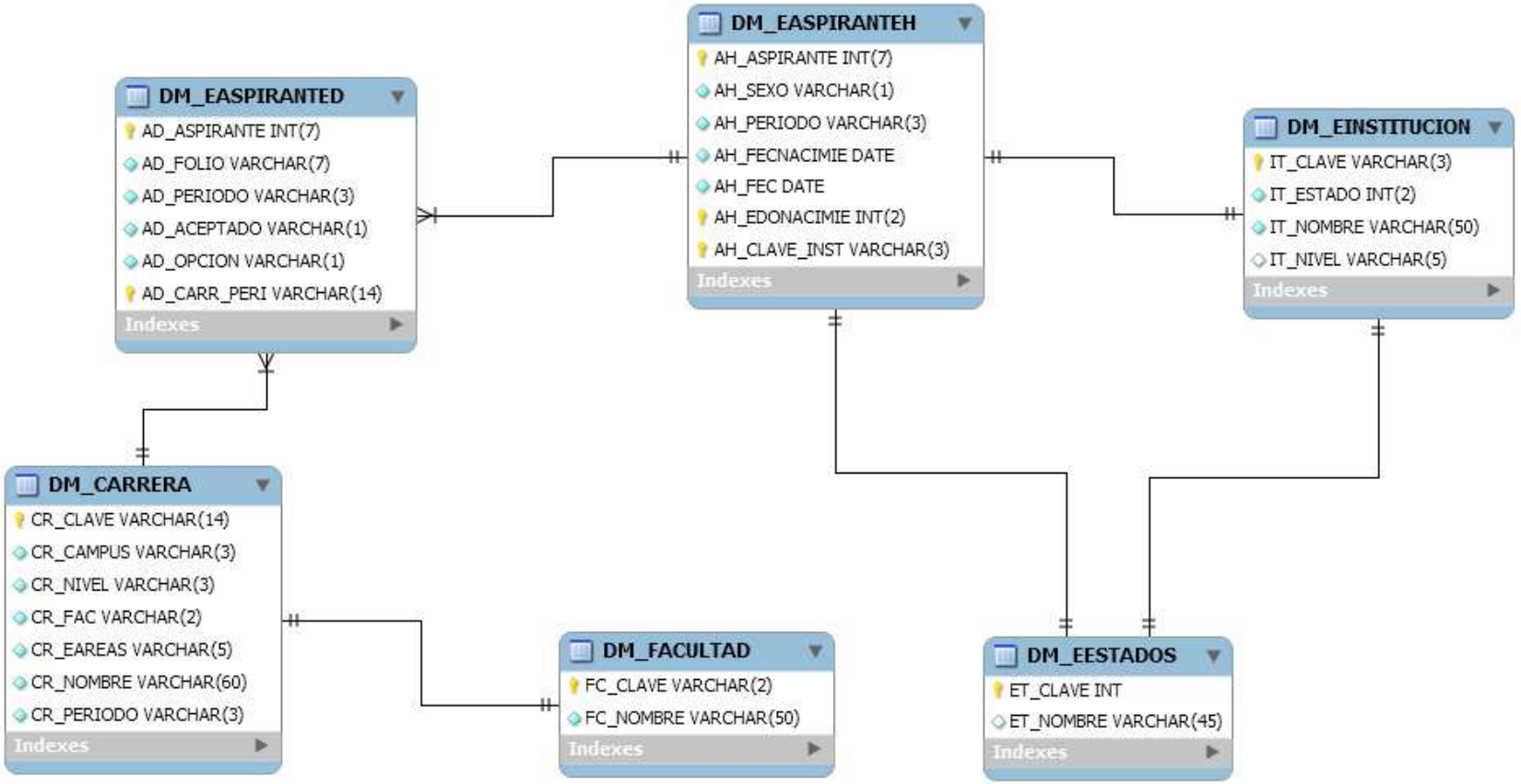


Figura 6. 4 Modelo Entidad-Relación de la Base de Datos Origen.

6.3.2 Inconsistencias y tratamiento

La búsqueda de errores o posibles inconsistencias se realizo con cada uno de los atributos de las tablas, esta parte se realiza con el objetivo de que los datos almacenados en el Data Mart sean lo más íntegros posibles.

- DM_EASPIRANTEH

Inconsistencia

Registros nulos en cada atributo: Se buscaron registros nulos que pudieran afectar a la estadística, por ejemplo, en el caso de estado de nacimiento, sexo, fecha de nacimiento.

La siguiente consulta permite obtener los estados de nacimiento con valor NULL, además se verifica que los estados estén dentro del rango de 1 a 34, ya que estos son los registrados en la tabla DM_EESTADOS.

```
SELECT
  AH_ASPIRANTE, AH_EDONACIMIE, AH_PERIODO
FROM BASE_DE_DATOS.DM_EASPIRANTEH
WHERE AH_EDONACIMIE IS NULL
     OR AH_EDONACIMIE <=0
     OR AH_EDONACIMIE >34
```

Solución

Se buscaron datos en otras tablas que podrían contenerlos, con esto se dio solución a algunos casos; los restantes se optó por asignarles un valor por default:

- En el caso de los estados de nacimiento se asigno de acuerdo a la institución de procedencia.

VISTA				
	DM_EASPIRANTEH		DM_EINSTITUCION	
#	AH_ASPIRANTE	AH_EDONACIMIE	IT_ESTADO	IT_CLAVE
1	******	22	22	?
2	******	22	22	?
3	******	22	22	?
4	******	22	22	?
5	******	22	22	?

Figura 6. 5 Muestra la clave de estado por institución.

Inconsistencia

Fechas de nacimiento erróneas: No existe un atributo en la base de datos que permita conocer la edad de los aspirantes, por ello es necesario calcularla a partir de fechas de nacimiento y la fecha de registro como aspirante. A continuación se presenta la consulta que permitió conocer lo descrito en este párrafo.

```
SELECT AH_ASPIRANTE, AH_FECNACIMIE, AH_FEC
,TRUNC(MONTHS_BETWEEN(AH_FEC,AH_FECNACIMIE)/12) EDAD
FROM BASE_DE_DATOS.DM_EASPIRANTEH
WHERE (AH_FECNACIMIE >= AH_FEC
     OR TRUNC(MONTHS_BETWEEN(AH_FEC,AH_FECNACIMIE)/12)
     < 12
     OR TRUNC(MONTHS_BETWEEN(AH_FEC,AH_FECNACIMIE)/12)
     > 60)
```

Solución

Las edades resultantes deben ser lógicas y de acuerdo al nivel académico que desean cursar, en el caso de PREPARATORIA no deben ser menores a 12 años ni mayores a 19; para el nivel UNIVERSITARIO no deben ser menores a 18 años ni mayores a 60 años.

Inconsistencia

Registros validos: Se verifico que los datos almacenados en los atributos correspondan solo a lo que deben guardas, por ejemplo, el campo sexo solo permite valores F para femenino y M para masculino.

```
SELECT AH_ASPIRANTE, AH_SEXO
FROM BASE_DE_DATOS.DM_EASPIRANTEH
WHERE AH_SEXO NOT IN ('M' ,'F')
```

Solución

No se encontró ningún caso.

Inconsistencia

Registros duplicados: En esta tabla se encontraron pocas filas con la misma información.

```
SELECT AH_ASPIRANTE, AH_?, AH_PERIODO, AH_FEC,
AH_EDONACIMIE
FROM BASE_DE_DATOS.DM_EASPIRANTEH
WHERE AH_?||AH_PERIODO IN (SELECT AH_?||AH_PERIODO
                FROM BASE_DE_DATOS.DM_EASPIRANTEH
                GROUP BY AH_?||AH_PERIODO
                HAVING Count(*) > 1)
ORDER BY AH_?||AH_PERIODO
```

Solución

Eliminar aquellos que no contarán con la asignación de alguna carrera.

Inconsistencia

Registros de prueba: No se pudo realizar una búsqueda a fondo de este tipo de registros almacenados en la base de datos debido a que no se tiene un formato establecido para la inserción de pruebas; aún así se encontraron algunos de estos registros.

```
SELECT AH_ASPIRANTE
FROM BASE_DE_DATOS.DM_EASPIRANTEH
WHERE AH_?||AH_PERIODO IN
(SELECT AH_?||AH_PERIODO
                FROM BASE_DE_DATOS.DM_EASPIRANTEH
                GROUP BY AH_?||AH_PERIODO
                HAVING COUNT(*) > 1)
ORDER BY AH_?||AH_PERIODO
```

Solución

Eliminarlos de la base de datos.

- DM_EASPIRANTED

Al igual que la tabla de encabezado (DM_EASPIRANTEH), en la tabla de DM_EASPIRANTED se analizaron los registros para encontrar valores nulos, registros no validos y registros duplicados; los datos inconsistentes fueron tratados de manera similar a la tabla de encabezado.

- DM_FACULTAD

Se buscaron las mismas inconsistencias que en las tablas anteriores pero no se encontró inconsistencia alguna.

- DM_CARRERA

Se buscaron las mismas inconsistencias que en las tablas anteriores pero no se encontró inconsistencia alguna.

- DM_EESTADOS

Se buscaron las mismas inconsistencias que en las tablas anteriores pero no se encontró inconsistencia alguna.

- DM_EINSTITUCION

Se buscaron las mismas inconsistencias que en las tablas anteriores pero no se encontró inconsistencia alguna.

La búsqueda de inconsistencias y su solución se realizo minuciosamente debido a que esta parte permite definir los filtros que serán aplicados a los datos antes de migrar al Data Mart.

El análisis presentado en este apartado es un resumen del trabajo realizado por él colaborador del proyecto. Para mayor detalle sobre las inconsistencias y tratamiento de datos revisar el documento titulado: "Detección y depuración de inconsistencias en Base de Datos para el uso de indicadores de un Data Mart" realizado por Homero Emmanuel Moctezuma Martínez.

6.4 Diseño de las tablas de hechos y dimensiones

Después del análisis del estado de los datos, se procede a crear las tablas de hecho y sus dimensiones, cuidando que se acoplen en un cien por ciento a los objetivos de cada Data Mart.

Se opto por utilizar el modelo en constelación porque permite la reutilización de tablas de dimensiones en los diferentes Data Mart, es decir, permite utilizar los registros de las diferentes dimensiones en cualquier Data Mart, evitando de esta manera la redundancia en los datos.

6.4.1 Dimensiones

- DM_CARRERA

Como su nombre lo indica, en esta dimensión de almacenan los datos de las diferentes carreras por periodo. Se utiliza para conocer las carreras a las que se aspira en un periodo determinado.

DIMENSIÓN DIM_CARRERA	ATRIBUTO	TIPO DE DATO	DESCRIPCIÓN
	CR_CLAVE	VARCHAR (14)	PK, llave primeria de carrera.
	CR_CAMPUS	VARCHAR (3)	Campus al que pertenece la carrera, por ejemplo: Jalpan, Querétaro, etc.
	CR_NIVEL	VARCHAR (3)	Nivel de la carrera, permite tres opciones: Licenciatura, Bachillerato, Técnico.
	CR_FAC	VARCHAR (2)	FK, llave foránea de la facultad.
	CR_EAREAS	VARCHAR (5)	...
	CR_NOMBRE	VARCHAR (60)	Nombre de la carrera.
	CR_PERIODO	VARCHAR (3)	Periodo en que se abrió la carrera.

Tabla 6. 7 DIM_CARRERA, dimensión que almacena las carreras de todos los períodos y facultades.

❖ DM_FACULTAD

Almacena los datos de las diferentes facultades de la Universidad.

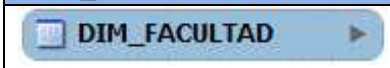
DIMENSIÓN DIM_FACULTAD	ATRIBUTO	TIPO DE DATO	DESCRIPCIÓN
	FC_CLAVE	VARCHAR (2)	PK, llave primaria de la facultad.
	FC_NOMBRE	VARCHAR (50)	Nombre de la facultad.

Tabla 6. 8 DIM_FACULTAD, almacena las facultades existentes en la Universidad.

❖ DM_EINSTITUCION

Almacena las instituciones de origen de los aspirantes.

DIMENSIÓN DIM_EINSTITUCION	**ATRIBUTO**	**TIPO DE DATO**	**DESCRIPCIÓN**
	IT_CLAVE	VARCHAR (3)	PK, llave primeria de institución.
	IT_ESTADO	INT (2)	FK, llave foránea del estado de origen de la institución.
	IT_NOMBRE	VARCHAR (50)	Nombre de la institución.
	IT_NIVEL	VARCHAR (5)	Nivel de la institución: Secundaria o Preparatoria.

Tabla 6. 9 DIM_EINSTITUCION, almacena los registros de las instituciones de origen por estado.

- DM_EESTADOS

Almacena los estados de la república mexicana.

DIMENSIÓN DIM_EESTADOS	ATRIBUTO	TIPO DE DATO	DESCRIPCIÓN
	ET_CLAVE	INT (2)	PK, llave primaria de Estado.
	ET_NOMBRE	VARCHAR (30)	Nombre del Estado.

Tabla 6. 10 DIM_EESTADOS, almacena los estados de la República Mexicana.

- DM_TIEMPO

Contiene los registros de los semestres y años en los que se registran los aspirantes.

DIMENSIÓN DIM_TIEMPO	ATRIBUTO	TIPO DE DATO	DESCRIPCIÓN
	TIM_CLAVE	INT (5)	PK, llave primaria del tiempo.
	TIM_YEAR	INT (4)	Se guardan los años que hasta el momento se tienen registrados.
	TIM_PERIODO	VARCHAR (3)	Almacena los periodos que hasta el momento se tienen registrados.

Tabla 6. 11 DIM_TIEMPO, almacena los años y periodos que hasta el momento se tienen registrados.

- DM_EASPIRANTEH

Son los datos generales del aspirante. Con ellos se podrán obtener consultas por sexo, edad, periodo, etc.

DIMENSIÓN DIM_EASPIRANTEH	ATRIBUTO	TIPO DE DATO	DESCRIPCIÓN
	AH_ASPIRANTE	INT (7)	PK, llave primaria del aspirante
	AH_SEXO	VARCHAR (1)	Sexo del aspirante, puede se F para femenino o M para masculino.
	AH_PERIODO	VARCHAR (3)	Periodo en que se dio de alta.
	AH_FECNACIMIE	DATE	Fecha de nacimiento del aspirante.
	AH_FEC	DATE	Fecha de registro del aspirante.
	AH_EDAD	INT (2)	Edad del aspirante.
	AH_EDONACIMIE	INT (2)	FK, llave foránea del estado de nacimiento del aspirante.
	AH_CLAVE_INST	VARCHAR (3)	FK, llave foránea de la institución de origen del aspirante.

Tabla 6. 12 DIM_EASPIRANTEH, almacena la información general del aspirante.

- DM_EASPIRANTED

Almacena la clave de la carrera a la que aspiran.

DIMENSIÓN DIM_EASPIRANTED	ATRIBUTO	TIPO DE DATO	DESCRIPCIÓN
	AD_ASPIRANTE	INT (7)	FK, llave foránea del aspirante con sus datos generales.
	AD_FOLIO	VARCHAR (7)	PK, llave primaria del detalle del aspirante.
	AD_PERIODO	VARCHAR (3)	Periodo donde se creó el detalle.
	AD_ACEPTADO	VARCHAR (1)	Guarda S si el aspirante fue aceptado o N si no fue aceptado.
	AD_CARR_PERI	VARCHAR (14)	FK, llave foránea para la carrera que se desea.
	AD_OPCION	VARCHAR (1)	Guarda 1 si es primera opción o 2 se es segunda.

Tabla 6. 13 DIM_EASPIRANTED, almacena la carrera elegida por el aspirante.

6.4.2 Hechos

- FACT_GENERAL_BAC

Se almacenan los datos generales referentes al nivel PREPARATORIA, por ejemplo, total de aspirantes a los diferentes planteles, total de aspirantes por sexo, total de aspirantes por grupo de edad, etc. En la Tabla 6.14 se describen los atributos que la conforman.

- FACT_GENERAL_LIC

Se almacenan los datos generales referentes al nivel LICENCIATURA, se realiza un análisis similar al anterior. Ver Tabla 6.15.

Las siguientes tablas de Hecho se realizaron exclusivamente para el nivel LICENCIATURA, debido a que es en este rubro se concentra la mayor cantidad de aspirantes y, por tanto, se desea tener un control más eficaz.

- FACT_FF_FACULTAD

Se almacenan las estadísticas de los aspirantes a un nivel de detalle mayor a FACT_GENERAL_LIC, ya que se realiza por facultad. Se toman las mismas métricas que en FACT_GENERAL_LIC, pero a nivel facultad.

- FACT_FC_CARRERA

Se almacenan los datos a nivel Facultad-Carrera. Con esto se podrá conocer la demanda que existe por carrera.

- FACT_FCI_INSTITUCION

Se almacenan los datos a nivel Facultad-Carrera-Institución de origen del aspirante.

- FACT_FCIE_ESTADO

Se almacenan los datos a nivel Facultad-Carrera-Institución-Estado de procedencia del aspirante.

La figura 6.6 representa los niveles de detalle descritos en las tablas de hecho.

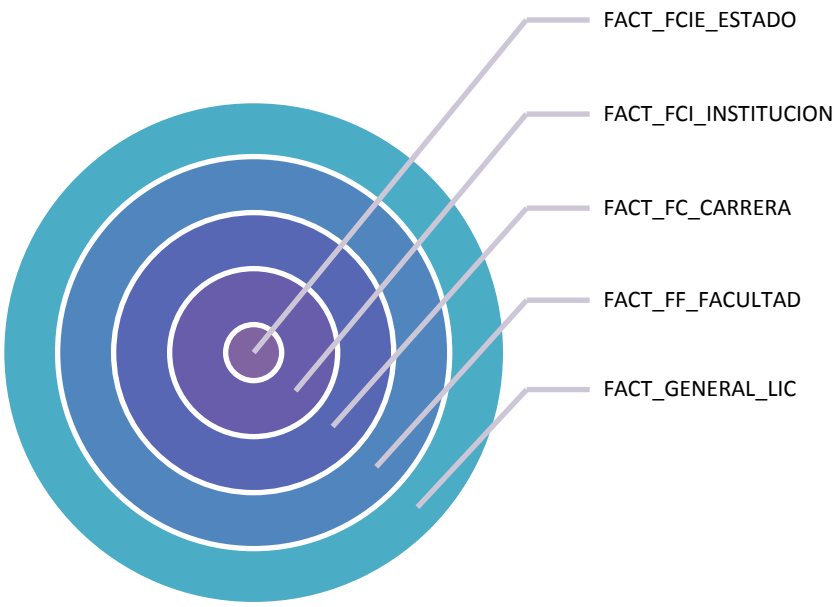


Figura 6. 6 Los indicadores se muestran en cinco niveles de detalle con el objetivo de obtener la mayor información posible de los datos almacenados en las tablas de la base de datos origen.

Tabla 6. 14 Descripción de los atributos que conforman las estadísticas generales de los aspirantes a PREPARATORIA.

HECHO	**INFORMACIÓN DE LOS ASPIRANTES PARA BACHILLERATO.**		
FACT_GENERAL_BAC	**ATRIBUTO**	**TIPO DE DATO**	**DESCRIPCIÓN**
	FGB_TIM	INT(5)	FK, llave foránea para conocer el periodo o año de los registros.
	FGB_TOTAL_ASPIRANTES	INT(5)	Almacena el total de aspirantes para bachillerato en un determinado tiempo.
	FGB_FEMENINO	INT(4)	Total de aspirantes del sexo Femenino.
	FGB_FPROM_EDAD	INT(4)	Promedio de edad de los aspirantes del sexo Femenino.
	FGB_MASCULINO	INT(4)	Total de aspirantes del sexo Masculino.
	FGB_MPROM_EDAD	INT(4)	Promedio de edad de los aspirantes del sexo Masculino.
	FGB_ACEPTADO	INT(4)	Número de aspirantes aceptados.
	FGB_AC_FEMENINO	INT(4)	Total de aspirantes aceptados del sexo Femenino
	FGB_ACF_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes aceptados del sexo Femenino.
	FGB_ACF_EDAD_12_15	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 12 a 15 años.
	FGB_ACF_EDAD_16_19	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 16 a 19 años.
	FGB_AC_MASCULINO	INT(4)	Total de aspirantes aceptados del sexo Masculino.
	FGB_ACM_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes aceptados del sexo Masculino.
	FGB_ACM_EDAD_12_15	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 12 a 15 años.
	FGB_ACM_EDAD_16_19	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 16 a 19 años.

	FGB_NO_ACEPTADO	INT(4)	Número de aspirantes NO aceptados.
	FGB_NOAC_FEMENINO	INT(4)	Total de aspirantes NO aceptados del sexo Femenino
	FGB_NOACF_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes NO aceptados del sexo Femenino.
	FGB_NOACF_EDAD_12_15	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 12 a 15 años.
	FGB_NOACF_EDAD_16_19	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 16 a 19 años.
	FGB_NOAC_MASCULINO	INT(4)	Total de aspirantes NO aceptados del sexo Masculino.
	FGB_NOACM_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes NO aceptados del sexo Masculino.
	FGB_NOACM_EDAD_12_15	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 12 a 15 años.
	FGB_NOACM_EDAD_16_19	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 16 a 19 años.

Tabla 6. 15 Descripción de los atributos que conforman las estadísticas generales de los aspirantes a LICENCIATURA.

HECHO	**INFORMACIÓN DE LOS ASPIRANTES PARA LICENCIATURA.**		
FACT_GENERAL_LIC	**ATRIBUTO**	**TIPO DE DATO**	**DESCRIPCIÓN**
	FGL_TIM	INT(5)	FK, llave foránea para conocer el periodo o año de los registros.
	FGL _TOTAL_ASPIRANTES	INT(5)	Almacena el total de aspirantes para licenciatura en un determinado tiempo.
	FGL _FEMENINO	INT(4)	Total de aspirantes del sexo Femenino.
	FGL _FPROM_EDAD	INT(4)	Promedio de edad de los aspirantes del sexo Femenino.
	FGL _MASCULINO	INT(4)	Total de aspirantes del sexo Masculino.
	FGL_MPROM_EDAD	INT(4)	Promedio de edad de los aspirantes del sexo Masculino.
	FGL_ACEPTADO	INT(4)	Número de aspirantes aceptados.
	FGL_AC_FEMENINO	INT(4)	Total de aspirantes aceptados del sexo Femenino
	FGL_ACF_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes aceptados del sexo Femenino.
	FGL_ACF_EDAD_16_19	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 16 a 19 años.
	FGL_ACF_EDAD_20_23	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 20 a 23 años.
	FGL_ACF_EDAD_24_27	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 24 a 27 años.
	FGL_ACF_EDAD_28_32	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 28 a 32 años.
	FGL_ACF_EDAD_33-60	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 33 a 60 años.
	FGL_AC_MASCULINO	INT(4)	Total de aspirantes aceptados del sexo Masculino.
	FGL_ACM_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes aceptados del sexo Masculino.
	FGL_ACM_EDAD_16_19	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 16 a 19 años.
	FGL_ACM_EDAD_20_23	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 20 a 23 años.

	FGL_ACM_EDAD_24_27	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 24 a 27 años.
	FGL_ACM_EDAD_28_32	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 28 a 32 años.
	FGL_ACM_EDAD_33-60	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 33 a 60 años.
	FGL_NO_ACEPTADO	INT(4)	Número de aspirantes NO aceptados.
	FGL_NOAC_FEMENINO	INT(4)	Total de aspirantes NO aceptados del sexo Femenino
	FGL_NOACF_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes NO aceptados del sexo Femenino.
	FGL_NOACF_EDAD_16_19	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 16 a 19 años.
	FGL_NOACF_EDAD_20_23	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 20 a 23 años.
	FGL_NOACF_EDAD_24_27	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 24 a 27 años.
	FGL_NOACF_EDAD_28_32	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 28 a 32 años.
	FGL_NOACF_EDAD_33_60	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 33 a 60 años.
	FGL_NOAC_MASCULINO	INT(4)	Total de aspirantes NO aceptados del sexo Masculino.
	FGL_NOACM_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes NO aceptados del sexo Masculino.
	FGL_NOACM_EDAD_16_19	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 16 a 19 años.
	FGL_NOACM_EDAD_20_23	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 20 a 23 años.
	FGL_NOACM_EDAD_24_27	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 24 a 27 años.
	FGL_NOACM_EDAD_28_32	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 28 a 32 años.
	FGL_NOACM_EDAD_33_60	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 33 a 60 años.

Tabla 6. 16 Descripción de los atributos que conforman las estadísticas generales de los aspirantes a LICENCIATURA por FACULTAD.

HECHO	**INFORMACIÓN DE LOS ASPIRANTES PARA LICENCIATURA POR FACULTAD.**		
FACT_FF_FACULTAD	**ATRIBUTO**	**TIPO DE DATO**	**DESCRIPCIÓN**
	FF_FAC	VARCHAR(2)	FK, llave foránea que contiene la clave de la facultad a la que pertenece determinado número de aspirantes.
	FF_TIM	INT(5)	FK, llave foránea para conocer el periodo o año de los registros.
	FF _TOTAL_ASPIRANTES	INT(5)	Almacena el total de aspirantes que pretenden ingresar por facultad en un determinado tiempo.
	FF _FEMENINO	INT(4)	Cantidad de aspirantes de sexo Femenino.
	FF _FPROM_EDAD	INT(4)	Promedio de edad de los aspirantes del sexo Femenino.
	FF _MASCULINO	INT(4)	Cantidad de aspirantes de sexo Masculino.
	FF_MPROM_EDAD	INT(4)	Promedio de edad de los aspirantes del sexo Masculino.
	FF_ACEPTADO	INT(4)	Número de aspirantes aceptados.
	FF_AC_FEMENINO	INT(4)	Total de aspirantes aceptados del sexo Femenino
	FF_ACF_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes aceptados del sexo Femenino.
	FF_ACF_EDAD_16_19	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 16 a 19 años.
	FF_ACF_EDAD_20_23	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 20 a 23 años.
	FF_ACF_EDAD_24_27	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 24 a 27 años.
	FF_ACF_EDAD_28_32	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 28 a 32 años.

	FF_ACF_EDAD_33-60	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 33 a 60 años.
	FF_AC_MASCULINO	INT(4)	Total de aspirantes aceptados del sexo Masculino.
	FF_ACM_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes aceptados del sexo Masculino.
	FF_ACM_EDAD_16_19	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 16 a 19 años.
	FF_ACM_EDAD_20_23	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 20 a 23 años.
	FF_ACM_EDAD_24_27	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 24 a 27 años.
	FF_ACM_EDAD_28_32	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 28 a 32 años.
	FF_ACM_EDAD_33-60	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 33 a 60 años.
	FF_NO_ACEPTADO	INT(4)	Número de aspirantes NO aceptados.
	FF_NOAC_FEMENINO	INT(4)	Total de aspirantes NO aceptados del sexo Femenino
	FF_NOACF_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes NO aceptados del sexo Femenino.
	FF_NOACF_EDAD_16_19	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 16 a 19 años.
	FF_NOACF_EDAD_20_23	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 20 a 23 años.
	FF_NOACF_EDAD_24_27	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 24 a 27 años.
	FF_NOACF_EDAD_28_32	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 28 a 32 años.
	FF_NOACF_EDAD_33_60	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 33 a 60 años.

	FF_NOAC_MASCULINO	INT(4)	Total de aspirantes NO aceptados del sexo Masculino.
	FF_NOACM_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes NO aceptados del sexo Masculino.
	FF_NOACM_EDAD_16_19	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 16 a 19 años.
	FF_NOACM_EDAD_20_23	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 20 a 23 años.
	FF_NOACM_EDAD_24_27	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 24 a 27 años.
	FF_NOACM_EDAD_28_32	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 28 a 32 años.
	FF_NOACM_EDAD_33_60	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 33 a 60 años.

Tabla 6. 17 Descripción de los atributos que conforman las estadísticas generales de los aspirantes a LICENCIATURA por CARRERA.

HECHO	**INFORMACIÓN DE LOS ASPIRANTES PARA LICENCIATURA POR CARRERA.**		
FACT_FC_CARRERA	**ATRIBUTO**	**TIPO DE DATO**	**DESCRIPCIÓN**
	FC _FAC	VARCHAR(2)	FK, llave foránea que contiene la clave de la facultad a la que pertenece determinado número de aspirantes.
	FC_CARR	VARCHAR(14)	FK, llave foránea de la carrera, en un periodo determinado, a la que pertenecen un número de aspirantes.
	FC _TIM	INT(5)	FK, llave foránea para conocer el periodo o año de los registros.
	FC _TOTAL_ASPIRANTES	INT(5)	Almacena el total de aspirantes que pretenden ingresar por carrera en un determinado tiempo.
	FC _FEMENINO	INT(4)	Cantidad de aspirantes de sexo Femenino.
	FC _FPROM_EDAD	INT(4)	Promedio de edad de los aspirantes del sexo Femenino.
	FC _MASCULINO	INT(4)	Cantidad de aspirantes de sexo Masculino.
	FC_MPROM_EDAD	INT(4)	Promedio de edad de los aspirantes del sexo Masculino.
	FC_ACEPTADO	INT(4)	Número de aspirantes aceptados.
	FC_AC_FEMENINO	INT(4)	Total de aspirantes aceptados del sexo Femenino
	FC_ACF_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes aceptados del sexo Femenino.
	FC_ACF_EDAD_16_19	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 16 a 19 años.
	FC_ACF_EDAD_20_23	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 20 a 23 años.
	FC_ACF_EDAD_24_27	INT(4)	Rango de edad para los aspirantes aceptados del

			sexo Femenino de 24 a 27 años.
	FC_ACF_EDAD_28_32	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 28 a 32 años.
	FC_ACF_EDAD_33-60	INT(4)	Rango de edad para los aspirantes aceptados del sexo Femenino de 33 a 60 años.
	FC_AC_MASCULINO	INT(4)	Total de aspirantes aceptados del sexo Masculino.
	FC_ACM_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes aceptados del sexo Masculino.
	FC_ACM_EDAD_16_19	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 16 a 19 años.
	FC_ACM_EDAD_20_23	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 20 a 23 años.
	FC_ACM_EDAD_24_27	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 24 a 27 años.
	FC_ACM_EDAD_28_32	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 28 a 32 años.
	FC_ACM_EDAD_33-60	INT(4)	Rango de edad para los aspirantes aceptados del sexo Masculino de 33 a 60 años.
	FC_NO_ACEPTADO	INT(4)	Número de aspirantes NO aceptados.
	FC_NOAC_FEMENINO	INT(4)	Total de aspirantes NO aceptados del sexo Femenino
	FC_NOACF_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes NO aceptados del sexo Femenino.
	FC_NOACF_EDAD_16_19	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 16 a 19 años.
	FC_NOACF_EDAD_20_23	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 20 a 23 años.
	FC_NOACF_EDAD_24_27	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 24 a 27 años.
	FC_NOACF_EDAD_28_32	INT(4)	Rango de edad para los aspirantes NO aceptados

			del sexo Femenino de 28 a 32 años.
	FC_NOACF_EDAD_33_60	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Femenino de 33 a 60 años.
	FC_NOAC_MASCULINO	INT(4)	Total de aspirantes NO aceptados del sexo Masculino.
	FC_NOACM_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes NO aceptados del sexo Masculino.
	FC_NOACM_EDAD_16_19	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 16 a 19 años.
	FC_NOACM_EDAD_20_23	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 20 a 23 años.
	FC_NOACM_EDAD_24_27	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 24 a 27 años.
	FC_NOACM_EDAD_28_32	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 28 a 32 años.
	FC_NOACM_EDAD_33_60	INT(4)	Rango de edad para los aspirantes NO aceptados del sexo Masculino de 33 a 60 años.

Tabla 6. 18 Descripción de los atributos que conforman las estadísticas generales de los aspirantes a LICENCIATURA por INSTITUCIÓN de origen.

HECHO	**INFORMACIÓN DE LOS ASPIRANTES PARA LICENCIATURA POR INSTITUCIÓN DE ORIGEN.**		
FACT_FCI_INSTITUCION	**ATRIBUTO**	**TIPO DE DATO**	**DESCRIPCIÓN**
	FCI _FAC	VARCHAR(2)	FK, llave foránea que contiene la clave de la facultad a la que pertenece determinado número de aspirantes.
	FCI_CARR	VARCHAR(14)	FK, llave foránea de la carrera, en un periodo determinado, a la que pertenecen un número de aspirantes.
	FCI _TIM	INT(5)	FK, llave foránea para conocer el periodo o año de los registros.
	FCI_INST	VARCHAR(3)	FK, llave foránea que apunta a la institución de origen de los aspirantes
	FCI _TOTAL_ASPIRANTES	INT(5)	Almacena el total de aspirantes que pretenden ingresar por institución de origen en un determinado tiempo.
	FCI _FEMENINO	INT(4)	Cantidad de aspirantes de sexo Femenino.
	FCI _FPROM_EDAD	INT(4)	Promedio de edad de los aspirantes del sexo Femenino.
	FCI _MASCULINO	INT(4)	Cantidad de aspirantes de sexo Masculino.
	FCI_MPROM_EDAD	INT(4)	Promedio de edad de los aspirantes del sexo Masculino.
	FCI_ACEPTADO	INT(4)	Número de aspirantes aceptados.
	FCI_AC_FEMENINO	INT(4)	Total de aspirantes aceptados del sexo Femenino
	FCI_ACF_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes aceptados del sexo Femenino.
	FCI_AC_MASCULINO	INT(4)	Total de aspirantes aceptados del sexo

			Masculino.
	FCI_ACM_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes aceptados del sexo Masculino.
	FCI_NO_ACEPTADO	INT(4)	Número de aspirantes NO aceptados.
	FCI_NOAC_FEMENINO	INT(4)	Total de aspirantes NO aceptados del sexo Femenino
	FCI_NOACF_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes NO aceptados del sexo Femenino.
	FCI_NOAC_MASCULINO	INT(4)	Total de aspirantes NO aceptados del sexo Masculino.
	FCI_NOACM_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes NO aceptados del sexo Masculino.

Tabla 6. 19 Descripción de los atributos que conforman las estadísticas generales de los aspirantes a LICENCIATURA por ESTADO según institución.

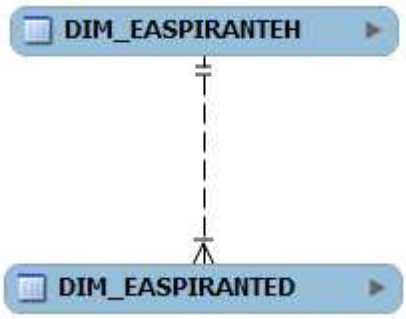
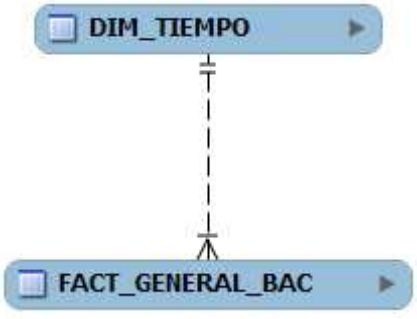
HECHO	**INFORMACIÓN DE LOS ASPIRANTES PARA LICENCIATURA POR ESTADO SEGÚN INSTITUCIÓN.**		
FACT_FCIE_ESTADO	**ATRIBUTO**	**TIPO DE DATO**	**DESCRIPCIÓN**
	FCIE _FAC	VARCHAR(2)	FK, llave foránea que contiene la clave de la facultad a la que pertenece determinado número de aspirantes.
	FCIE_CARR	VARCHAR(14)	FK, llave foránea de la carrera, en un periodo determinado, a la que pertenecen un número de aspirantes.
	FCIE _TIM	INT(5)	FK, llave foránea para conocer el periodo o año de los registros.
	FCIE_INST	VARCHAR(3)	FK, llave foránea que apunta a la institución de origen de los aspirantes
	FCIE_EST	INT(2)	FK, llave foránea que apunta al estado de la institución.
	FCIE _TOTAL_ASPIRANTES	INT(5)	Almacena el total de aspirantes que pretenden ingresar por estado según institución.
	FCIE _FEMENINO	INT(4)	Cantidad de aspirantes de sexo Femenino.
	FCIE _FPROM_EDAD	INT(4)	Promedio de edad de los aspirantes del sexo Femenino.
	FCIE _MASCULINO	INT(4)	Cantidad de aspirantes de sexo Masculino.
	FCIE_MPROM_EDAD	INT(4)	Promedio de edad de los aspirantes del sexo Masculino.
	FCIE_ACEPTADO	INT(4)	Número de aspirantes aceptados.
	FCIE_AC_FEMENINO	INT(4)	Total de aspirantes aceptados del sexo Femenino
	FCIE_ACF_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes aceptados del sexo Femenino.

	FCIE_AC_MASCULINO	INT(4)	Total de aspirantes aceptados del sexo Masculino.
	FCIE_ACM_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes aceptados del sexo Masculino.
	FCIE_NO_ACEPTADO	INT(4)	Número de aspirantes NO aceptados.
	FCIE_NOAC_FEMENINO	INT(4)	Total de aspirantes NO aceptados del sexo Femenino
	FCIE_NOACF_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes NO aceptados del sexo Femenino.
	FCIE_NOAC_MASCULINO	INT(4)	Total de aspirantes NO aceptados del sexo Masculino.
	FCIE_NOACM_PROM_EDAD	INT(4)	Promedio de edad de los aspirantes NO aceptados del sexo Masculino.

6.4.3 Diseño y modelo de los Data Mart

Con las tablas de hecho y las dimensiones descritas anteriormente, se procede a la elaboración de los diferentes Data Mart. Se presenta nuevamente una descripción sobre las tablas de dimensión y la tabla de hecho que compone a cada uno de los Data Mart elaborado, además se muestra un diagrama general y un diagrama donde se encuentran los atributos que conforman a los Data Mart.

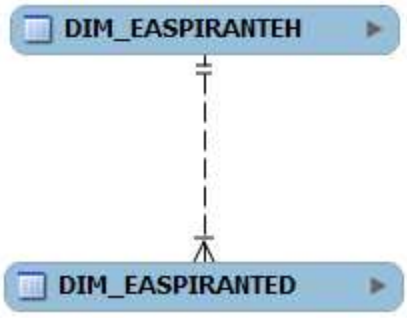
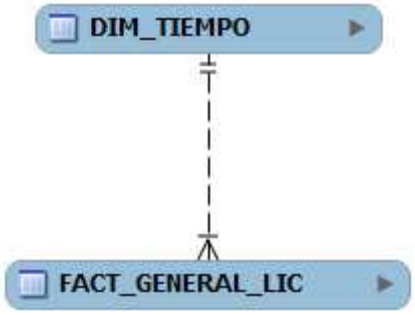
- **Modelo General para Preparatoria**

DIMENSIONES	DESCRIPCIÓN
	Son la materia prima de las tablas de hecho, no tienen una relación directa con la tabla de hecho, pero se utilizan obtener los datos de los diferentes atributos que conforman la tabla FACT_GENERAL_BAC.
	Diagrama general del Data Mart aspirantes para preparatoria. Almacena datos como, total de aspirantes, aspirantes por sexo, aspirantes aceptados y no aceptados, aspirantes por grupo de edad, etc. La figura 6.7 muestra los atributos que conforman el Data Mart: FACT_GENERAL_LIC.

Data Mart: Aspirantes para preparatoria.

Figura 6. 7 Data Mart: Aspirantes para nivel preparatoria (FACT_GENERAL_BAC)

❖ **Modelo general para Licenciatura**

DIMENSIONES	DESCRIPCIÓN
	Son la materia prima de las tablas de hecho, no tienen una relación directa con la tabla de hecho, pero se utilizan obtener los datos de los diferentes atributos que conforman la tabla FACT_GENERAL_LIC.
	Diagrama general del Data Mart aspirantes para licenciatura. Almacena datos como, total de aspirantes, aspirantes por sexo, aspirantes aceptados y no aceptados, aspirantes por grupo de edad, etc. La figura 6.8 muestra los atributos que conforman el Data Mart: FACT_GENERAL_LIC.

Data Mart: Aspirantes para licenciatura.

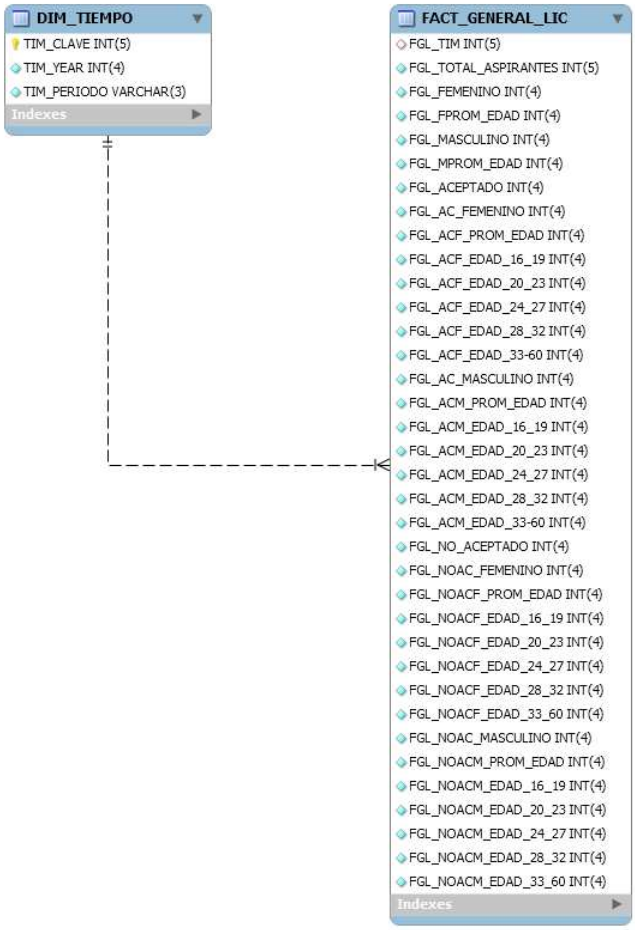


Figura 6. 8 Data Mart: Aspirantes para nivel licenciatura (FACT_GENERAL_LIC)

❖ **Data Mart: Aspirantes por FACULTAD**

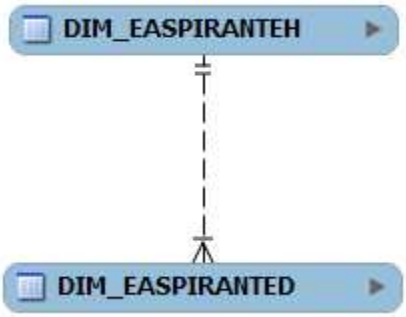
DIMENSIONES	DESCRIPCIÓN
	Son la materia prima de las tablas de hecho, no tienen una relación directa con la tabla de hecho, pero se utilizan obtener los datos de los diferentes atributos que conforman la tabla FACT_FF_FACULTAD.

Diagrama general del Data Mart aspirantes por facultad a nivel licenciatura. Almacena datos como, total de aspirantes por facultad, aspirantes por sexo de una determinada facultad, aspirantes aceptados y no aceptados a dicha facultad, aspirantes por grupo de edad, etc.

La figura 6.9 muestra los atributos que conforman el Data Mart FACT_FF_FACULTAD descrito en el diagrama anterior.

Data Mart: Aspirantes para licenciatura por Facultad.

Figura 6. 9 Data Mart: Aspirantes por facultad de nivel licenciatura (FACT_FF_FACULTAD)

❖ **Data Mart: Aspirantes por CARRERA**

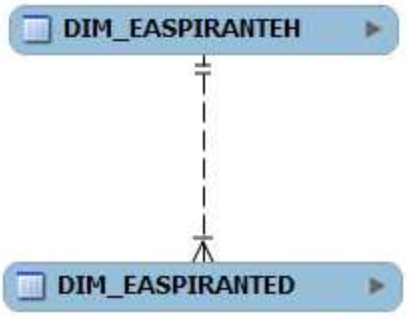
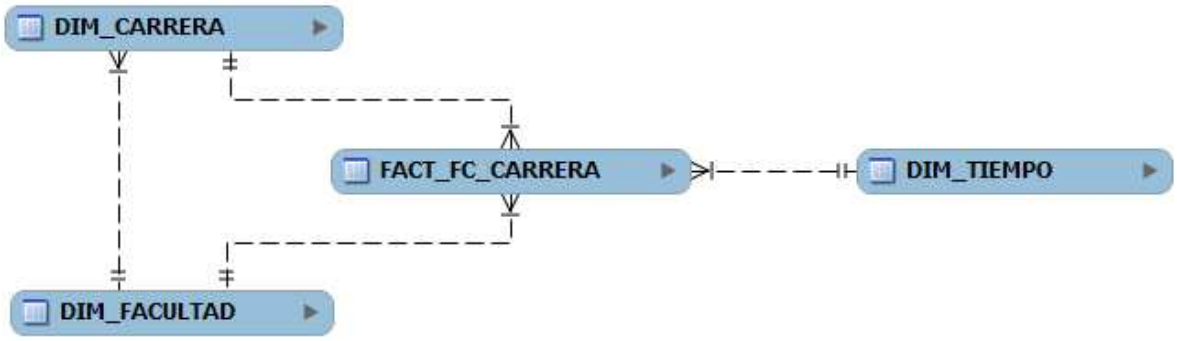
DIMENSIONES	DESCRIPCIÓN
	Son la materia prima de las tablas de hecho, no tienen una relación directa con la tabla de hecho, pero se utilizan obtener los datos de los diferentes atributos que conforman la tabla FACT_FC_CARRERA.

Diagrama general del Data Mart aspirantes por carrera a nivel licenciatura. Almacena datos como, total de aspirantes por carrera, aspirantes por sexo de cada una de las carreras, aspirantes aceptados y no aceptados, aspirantes por grupo de edad, etc.



La figura 6.10 muestra los atributos que conforman el Data Mart descrito en el diagrama anterior.

Data Mart: Aspirantes para licenciatura por Carrera.

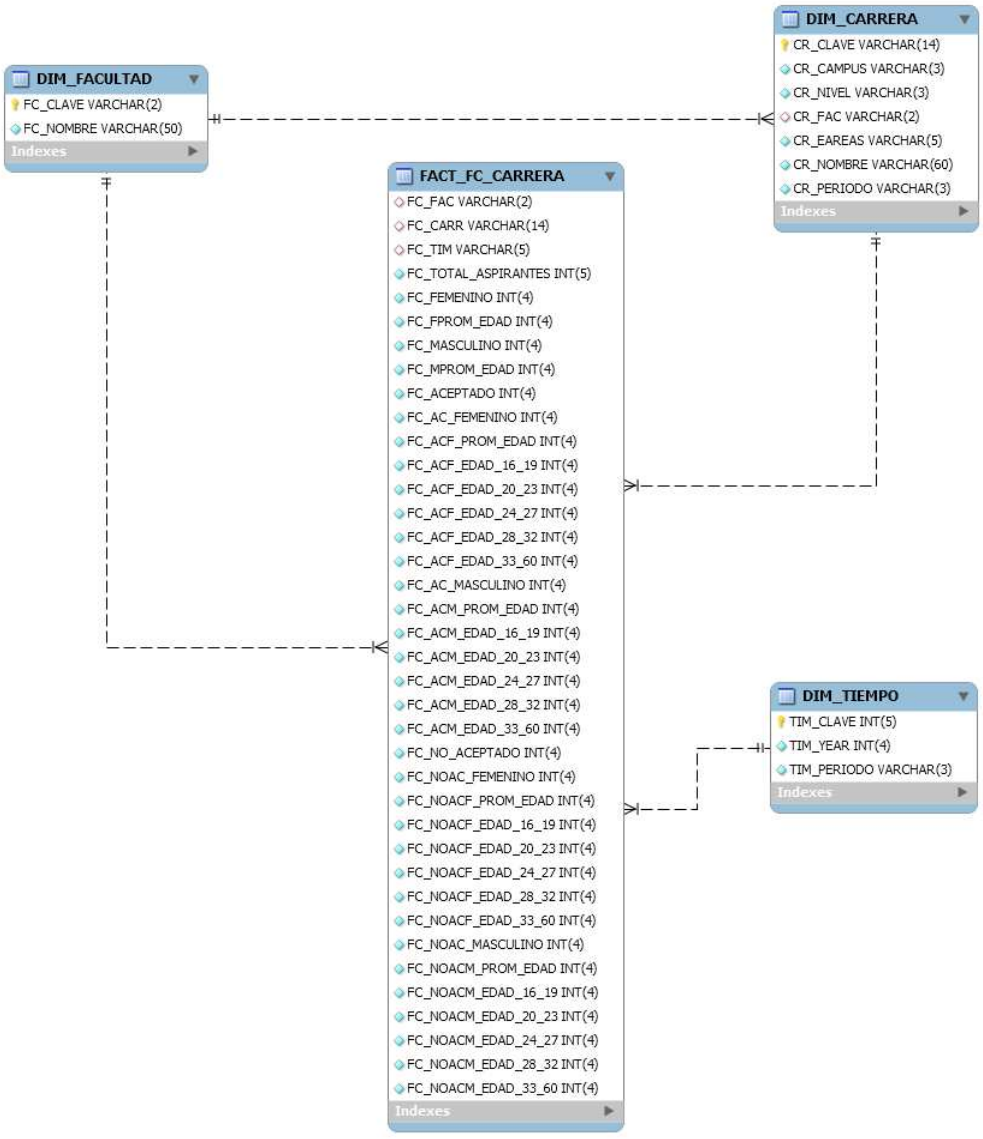


Figura 6. 10 Data Mart: Aspirantes por carrera de nivel licenciatura (FACT_FC_CARRERA)

❖ **Data Mart: Aspirantes por INSTITUCIÓN**

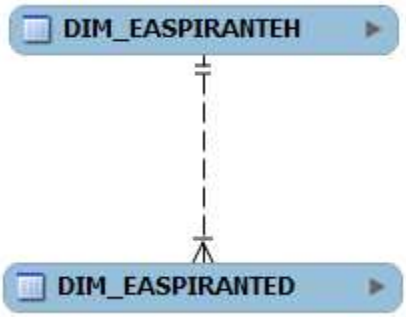
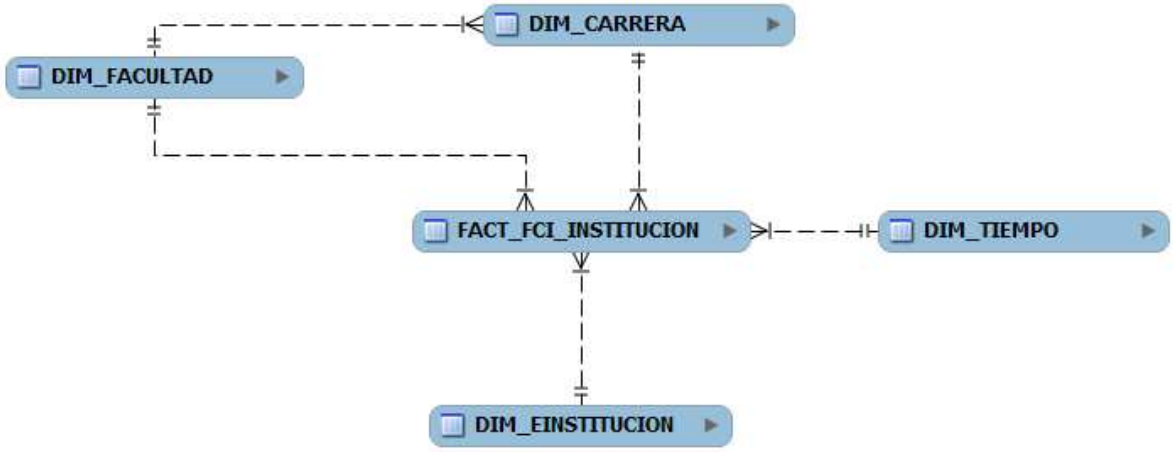
DIMENSIONES	DESCRIPCIÓN
	Son la materia prima de las tablas de hecho, no tienen una relación directa con la tabla de hecho, pero se utilizan obtener los datos de los diferentes atributos que conforman la tabla FACT_FCI_INSTITUCION.

Diagrama general del Data Mart aspirantes por institución de origen a nivel licenciatura. Almacena datos como, total de aspirantes por institución de origen, aspirantes por sexo, aspirantes aceptados y no aceptados, aspirantes por grupo de edad, etc.



La figura 6.11 muestra los atributos que conforman el Data Mart descrito en el diagrama anterior.

Data Mart: Aspirantes para licenciatura por Institución de origen.

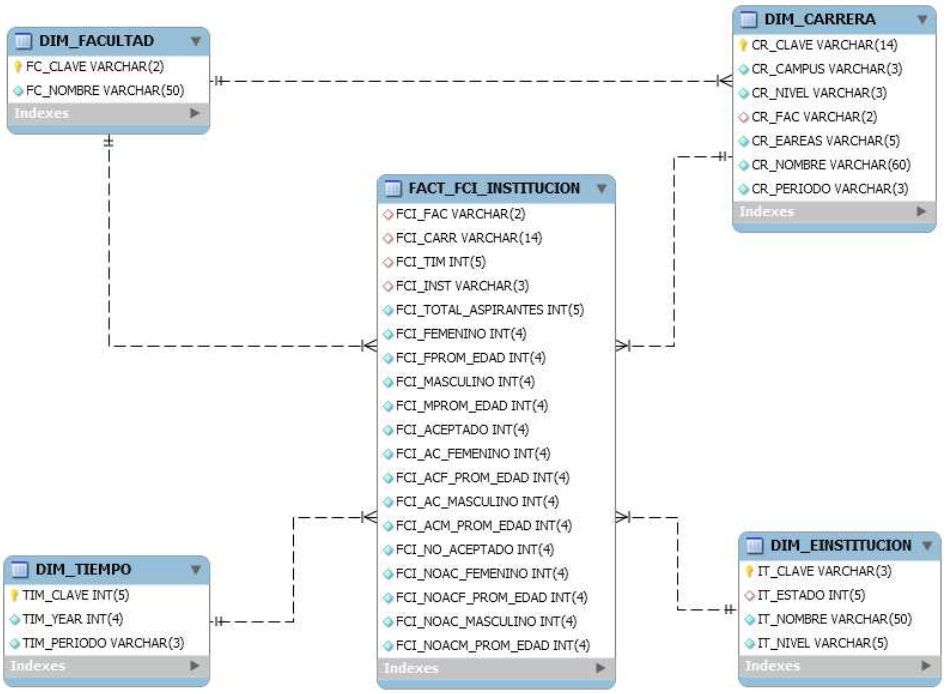


Figura 6. 11Data Mart: Aspirantes por institución de origen a nivel licenciatura (FACT_FCI_INSTITUCION)

❖ **Data Mart: Aspirantes por ESTADO**

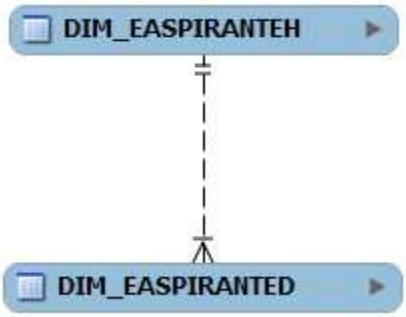
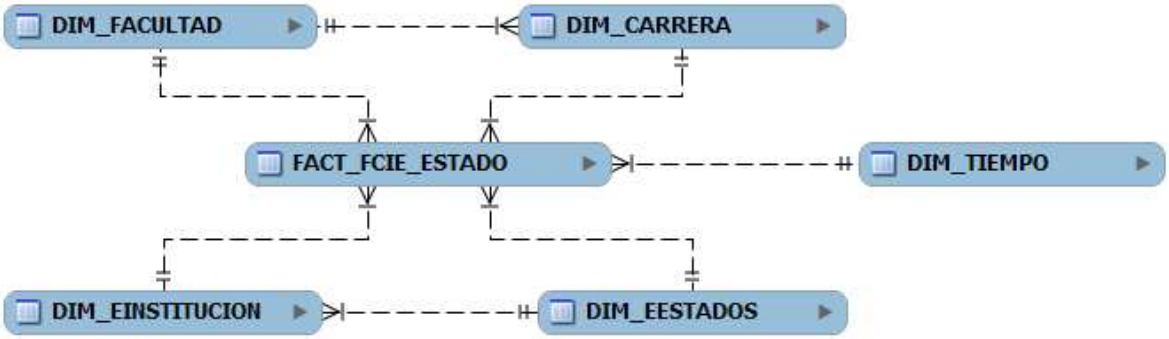
DIMENSIONES	DESCRIPCIÓN
	Son la materia prima de las tablas de hecho, no tienen una relación directa con la tabla de hecho, pero se utilizan obtener los datos de los diferentes atributos que conforman la tabla FACT_FCIE_ESTADO.

Diagrama general del Data Mart aspirantes por estado de nacimiento a nivel licenciatura. Almacena datos como, total de aspirantes por estado de nacimiento, aspirantes por sexo, aspirantes aceptados y no aceptados, aspirantes por grupo de edad, etc



La figura 6.12 muestra los atributos que conforman el Data Mart descrito en el diagrama anterior.

Data Mart: Aspirantes para licenciatura por Estado según Institución de origen.

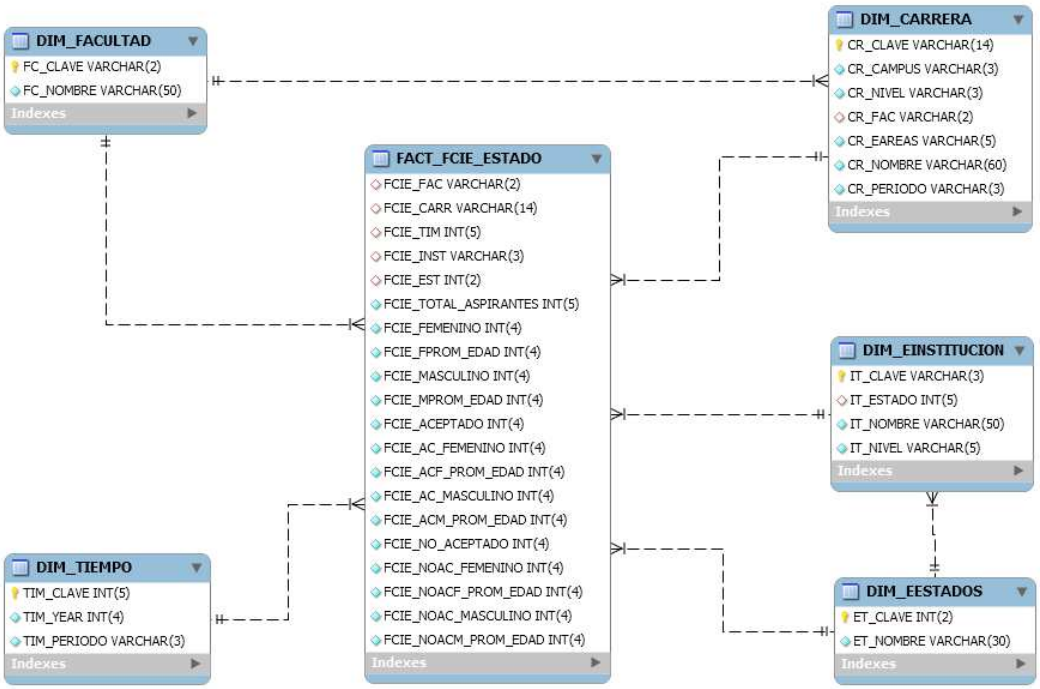


Figura 6. 12 Data Mart: Aspirantes por estado según institución de nivel licenciatura (FACT_FCIE_ESTADO)

En seguida se muestra el diagrama general de los Data Mart(Figura 6.13).

Diagrama general de los diferentes Data Marts.

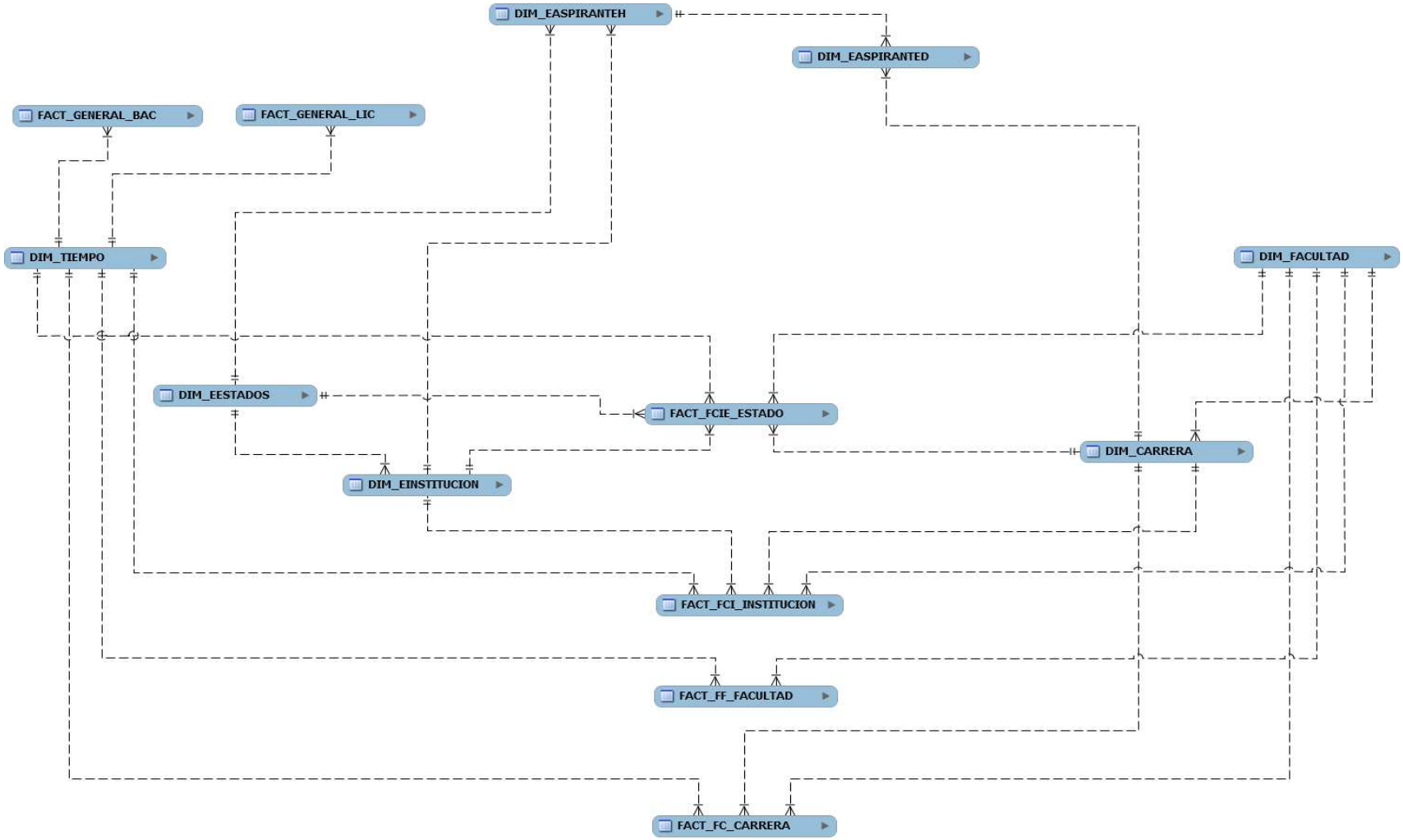


Figura 6. 13 Modelo en Constelación de los Data Mart.

6.5 Migración de los datos a utilizar (dimensiones)

Para realizar las diferentes transformaciones se tomo en cuenta las inconsistencias encontradas en la etapa de búsqueda de inconsistencias y su tratamiento. En general los filtros aplicados permiten almacenar los datos en el Data Mart de la forma más consistente e integra posible. El proceso que se describe en los apartados siguientes se realizo únicamente para llenar las dimensiones de los Data Mart, en cada caso se explica de manera general el procedimiento realizado.

La tabla 6.20 muestra la nomenclatura utilizada a lo largo del procedimiento.

6.5.1 Nomenclatura

NOMBRE	ICONO	DESCRIPCIÓN
TRANS_[NOMBRE DE LA TABLA]		Indica que se realizará una transformación de un conjunto de datos.
TIF_[NOMBRE DE LA TABLA]		Indica que la tabla señalada es una tabla de entrada, es decir, tabla de origen de los datos. Permite leer los datos de una tabla.
FRF_[NOMBRE DE LA TABLA] o FRF_[VALOR DE UN ATRIBUTO DE UNA TABLA]		Se trata de un filtro que se aplicará a los datos de entrada.
EOF_[NOMBRE DE LA TABLA]___		Indica que es un archivo de salida, obtiene los datos que no cumplen con las restricciones de un filtro. Después del nombre de la tabla pueden tener una palabra clave para saber qué tipo de datos almacena, ejemplo: NULL
TOP_[NOMBRE DE LA TABLA]		Son las tablas de salida. Almacenan aquellos datos que han cumplido con todos los requisitos de integridad propuestos por los filtros.

JSF_[NOMBRE DE LA TABLA] o JSF_[VALOR DE UN ATRIBUTO DE UNA TABLA]		Indica que se ejecutará código Java Script.
RVF_[NOMBRE DE LA TABLA] o RVF_[VALOR DE UN ATRIBUTO DE UNA TABLA]		Indica que los atributos de entrada cambiarán de orden de salida o serán eliminados por no ser necesarios en la tabla destino.
FOP_[NOMBRE DE LA TABLA]		Se trata de un archivo de texto que contiene datos temporales.
UNION_[COMBINA EL NOMBRE DE DOS TABLAS]		Indica que se realizará la unión de dos tablas.

Tabla 6. 20 Nomenclatura utilizada para la migración de datos.

❖ Proceso ETL para DIM_EESTADOS

Para migrar los datos de la vista DM_EESTADOS a la dimensión DIM_EESTADOS fue necesario construir un filtro para no almacenar valores nulos en la dimensión. Los datos contenidos en la vista DM_EESTADOS no presentaban inconsistencia alguna por tal motivo y porque estos datos permanecen constantes solo se aplico dicho filtro.

El filtro consiste en leer los valores de entrada y, por medio de una condición, seleccionar aquellos que cumplen con la misma y el resto se almacena en un archivo para su posterior análisis. En este caso, como ya se mencionó, no existen valores nulos, por tanto, todos los valores fueron almacenados en la dimensión TOP_DM_EESTADOS.

Asimismo, esta transformación no será necesario que se repita en cada semestre pues los Estados de la República Mexicana no cambian frecuentemente; al igual que en la base de datos origen, esta dimensión será considerada como un catalogo.

La figura 6.14 muestra cómo se acomodaron los diferentes pasos antes de migrar los datos.

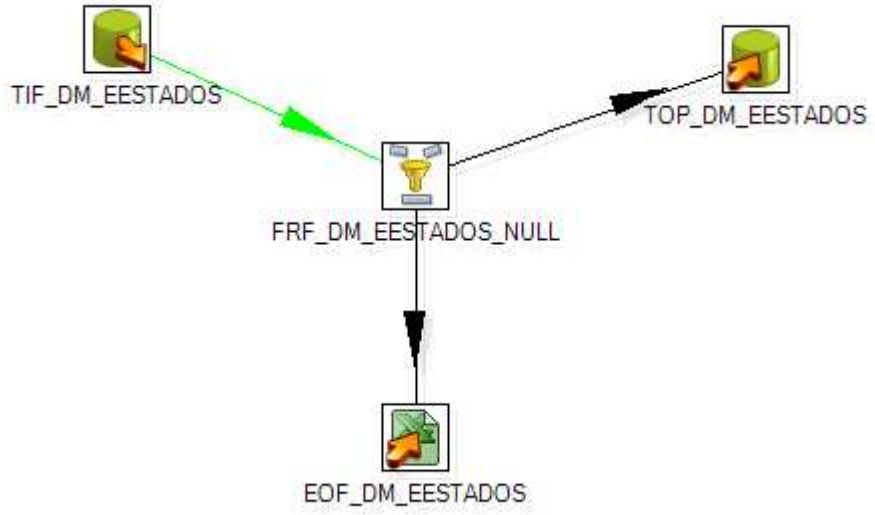


Figura 6. 14 Tratamiento de los datos pertenecientes a los Estados.

❖ Proceso ETL para DIM_FACULTAD

La tabla origen (TIF_DM_EESCUELA) es considerada como un catalogo porque sus datos no cambian constantemente o no se agregan nuevos registros con regularidad. Además no se encontraron inconsistencias en sus datos; por lo anterior, solo se aplico un filtro para buscar valores nulos.

La figura 6.15 muestra el modelo utilizado en esta transformación.

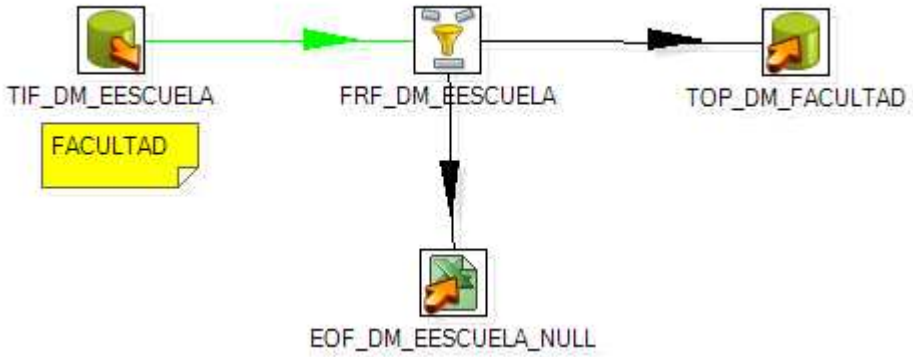


Figura 6. 15 Tratamiento de los datos pertenecientes a las Facultades de la UAQ.

❖ Proceso ETL para DIM_TIEMPO

El mecanismo para obtener la dimensión de tiempo se muestra la figura 6.16, donde la tabla de entrada hace referencia a las carreras que se ofertan por semestre, logrando así, registrar en la dimensión DIM_TIEMPO cada uno de los semestres del año en curso y anteriores.

La consulta ingresada en la tabla de entrada se describe a continuación:

```
SELECT DISTINCT (CR_PERIODO) AS TIM_PERIODO,
SUBSTR('20'||CR_PERIODO,1,4) AS TIM_YEAR
FROM BASE_DE_DATOS.DM_ECARRERA
ORDER BY CR_PERIODO
```

Como la consulta lo indica, solo es necesario leer aquellos valores que permitan conocer los semestres transcurridos hasta ahora y, como no existe un campo que contenga el año, se agregaron dos dígitos al periodo para adquirir el año.

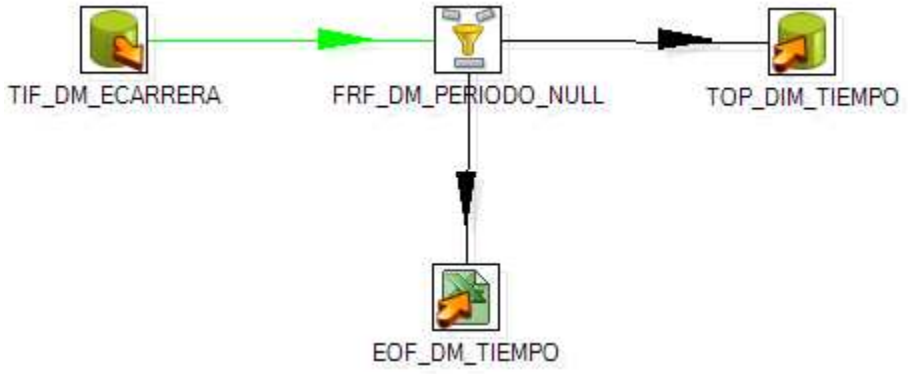


Figura 6. 16 Obtención de la dimensión Tiempo.

❖ Proceso ETL para DIM_EINSTITUCION

Al igual que el apartado 6.6.3, las instituciones son consideradas como un catalogo y su tratamiento consistió en leer posibles valores nulos.

La siguiente figura permite visualizar las etapas realizadas para migrar los datos.

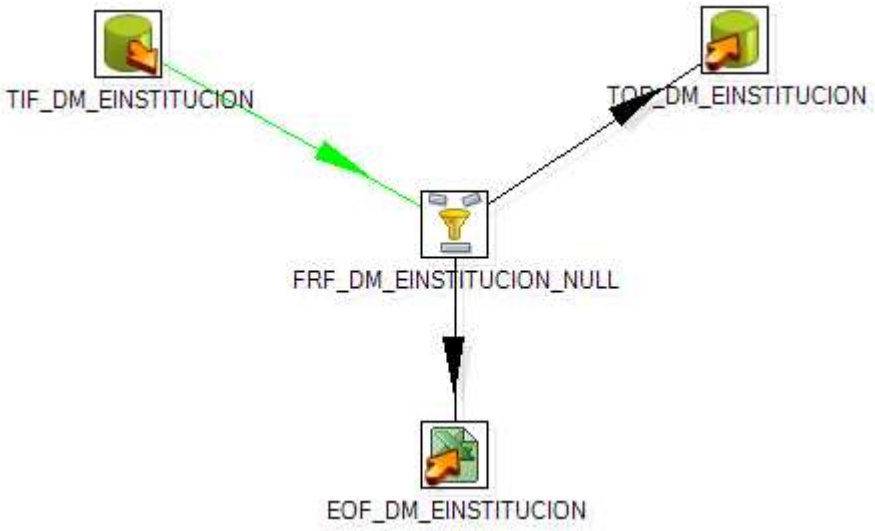


Figura 6. 17 Tratamiento de los datos relacionados con las Instituciones de origen de los Aspirantes.

❖ **Proceso ETL para DIM_CARRERA**

La vista DM_ECARRERA permite conocer de varias formas el nivel de la carrera, la expuesta en este apartado permite obtener los datos sin posibilidad de error, ya que se realizaron diversas pruebas para encontrar inconsistencias en estos datos.

```
SELECT
 CR_CARRERA||'-'||CR_PERIODO AS CR_CLAVE, CR_CAMPUS,
CR_NIVEL, ? AS CR_?, CR_EAREAS, CR_NOMBRE, CR_PERIODO,
CR_?,
CR_CARRERA, CR_?, CR_?, ?, ?
FROM BASE_DE_DATOS.DM_ECARRERA, BASE_DE_DATOS.?
WHERE CR_?=?
```

Por cuestiones de seguridad en la vista DM_ECARRERA, expuesta en el apartado 6.4.1 no hace mención de tres atributos y una tabla extra que se utilizaron para la carga de la dimensión DIM_CARRERA. Con esta consulta se obtienen los datos para la dimensión mencionada. Su tratamiento se describe enseguida:

Se utilizaron tres filtros:

1. Búsqueda de valores nulos: La figura 6.18 muestra las condiciones que los datos de entrada deben de cumplir para poder seguir con el proceso.

2. Seleccionar de acuerdo a Nivel de la carrera: Para conocer el nivel de la carrera se tomo en cuenta un digito que indica a qué nivel corresponde dicha carrera, por tanto, solo es necesario indicar al filtro hacia donde debe enviar los valores. La figura 6.19 muestra aquellas que terminan con '1'. Se opto por este método debido a que el atributo CR_NIVEL presentaba algunas inconsistencias que no podrían ser tratadas en la base de datos origen porque existían varios registros que dependen de ellos.

   La terminación 1 indica que su nivel es TECNICO, el 2 indica que su nivel es LICENCIATURA y la terminación 3 indica que su nivel es PREPARATORIA.

   Existen algunos casos especiales, pero no serán mencionados en este documento.

3. Remplazar registros de Nivel de la carrera que no correspondan con la terminación mencionada: Se programaron tres Java Script para lograr cambiar los registros a la hora de ingresar a la dimensión. La figura 6.20 muestra un fragmento de código relacionado con el nivel TECNICO.

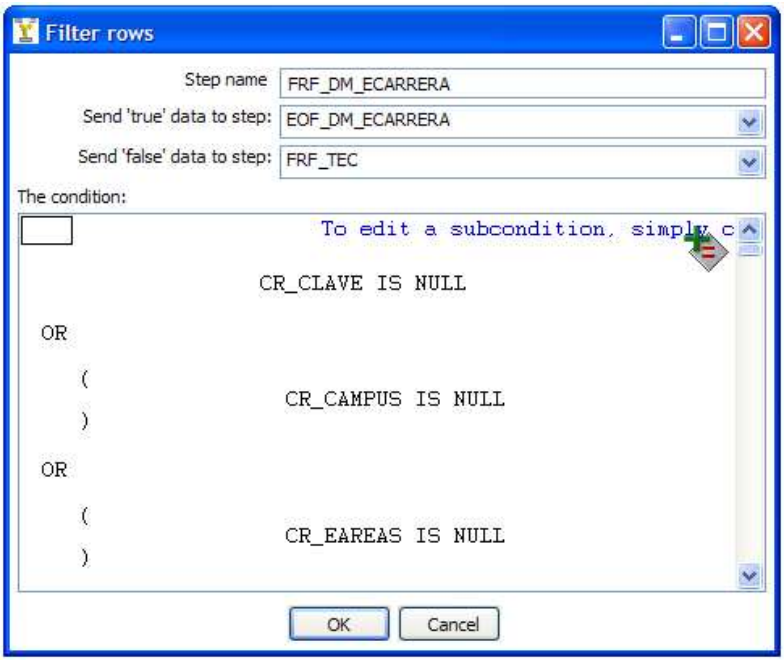


Figura 6. 18 Filtro para valores nulos.

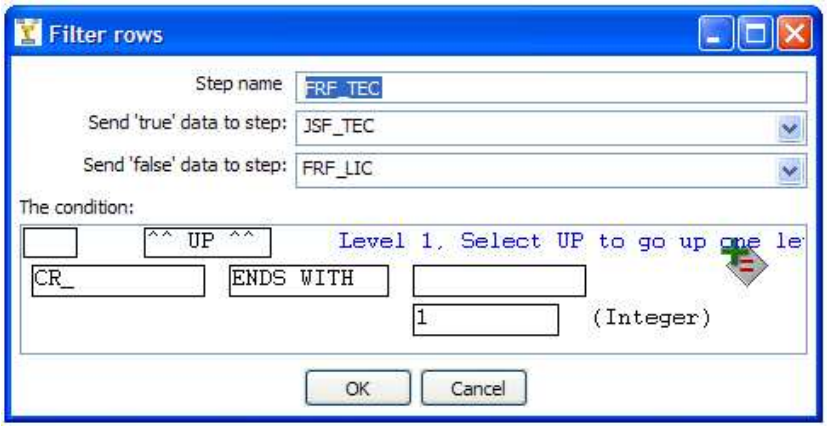


Figura 6. 19 Filtro para seleccionar el Nivel de la carrera.

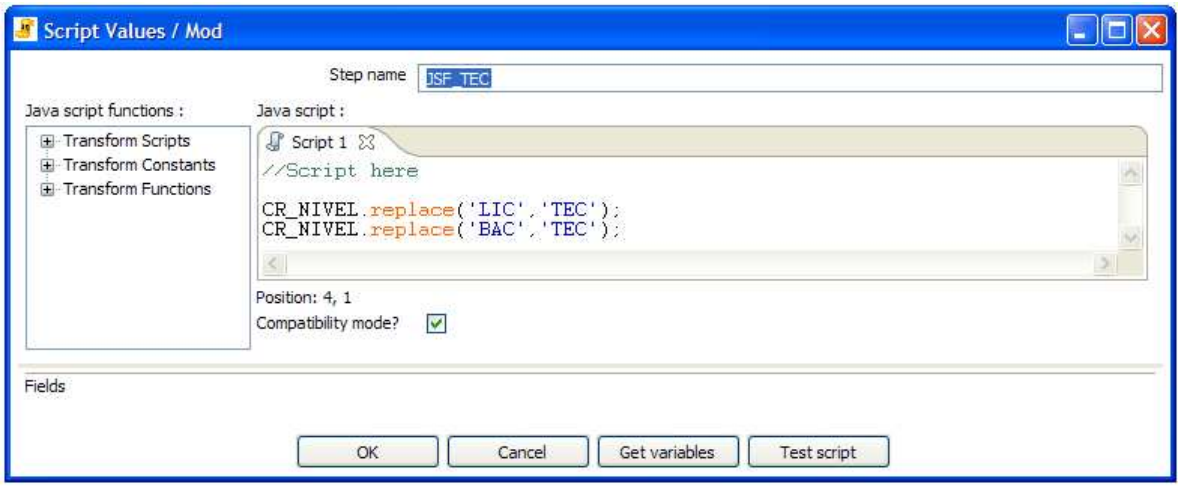


Figura 6. 20 Código Java Script para modificar el nivel de la carrera.

4. Una vez filtrados los datos de entrada se eliminan aquellos que no son necesarios en la dimensión DIM_CARRERA;

5. Finalmente se cargan los datos a la dimensión. La figura 6.21 muestra todo el procedimiento realizado para leer, transformar y cargar los datos a la dimensión DIM_CARRERA.

Proceso de Extracción-Transformación-Carga para DIM_CARRERA.

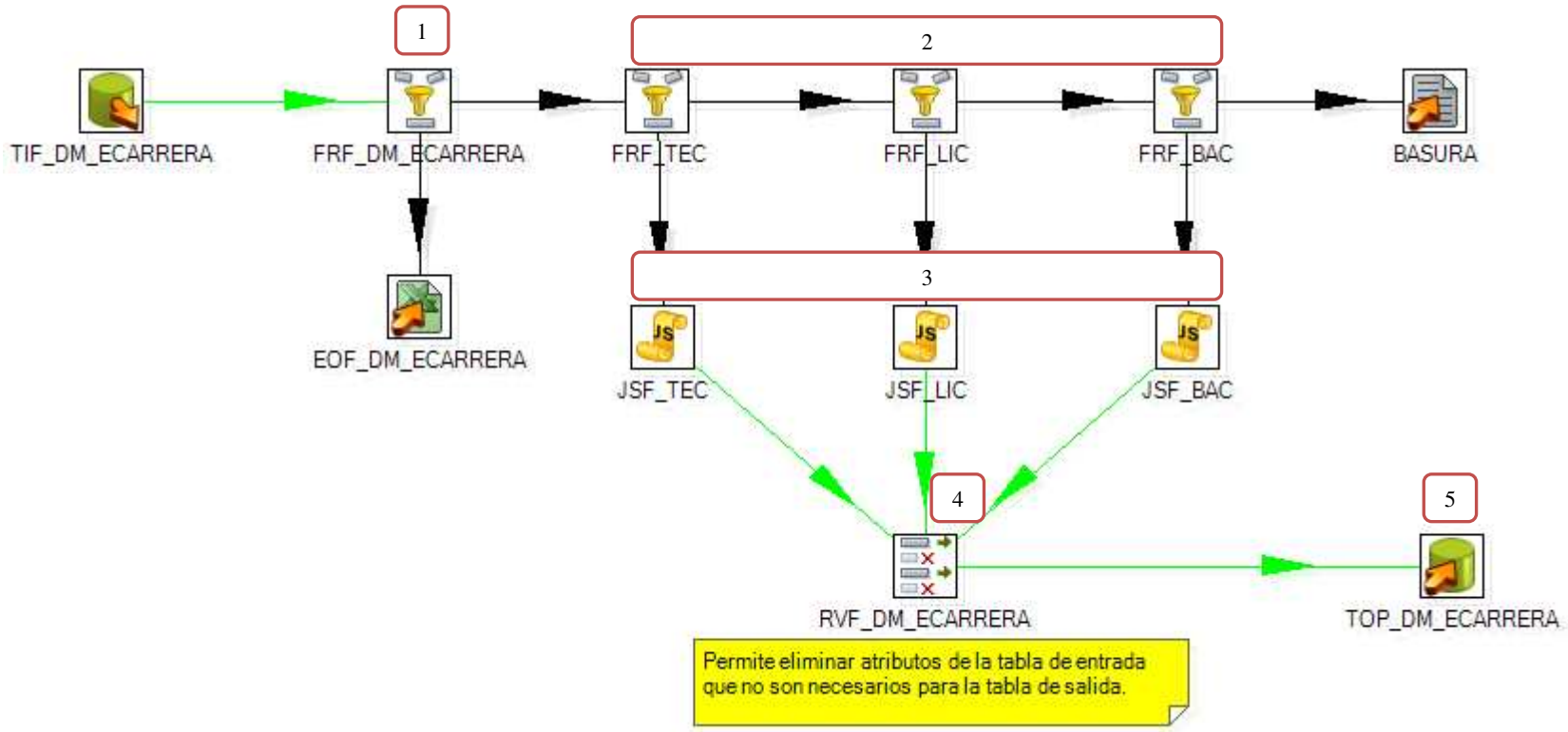


Figura 6. 21 Tratamiento de los datos relacionados con las Carreras ofertada por la UAQ.

❖ **Proceso ETL para DIM_EASPIRANTEH**

1. Extracción de los datos de entrada: La siguiente consulta permite obtener los datos de entrada para su posterior tratamiento. La consulta requería de otros atributos y tablas que no fueron mencionadas en la sección 6.4, por tanto, y por los motivos ya planteados no se muestran en la consulta.

```
SELECT
  AH_ASPIRANTE, AH_SEXO, AH_PERIODO, AH_FECNACIMIE,
AH_FEC
, TRUNC (MONTHS_BETWEEN (AH_FEC, AH_FECNACIMIE)/12,0) AS
AH_EDAD
, AH_EDONACIMIE, AH_CLAVESECUN, AH_CLAVEPREPA
     FROM BASE_DE_DATOS.DM_EASPIRANTEH
     WHERE AH_ASPIRANTE IN (
          SELECT DISTINCT (AD_ASPIRANTE) FROM
          BASE_DE_DATOS.DM_EASPIRANTED WHERE AD_?
          IN …
ORDER BY AH_PERIODO
```

2. Filtro para valores nulos: La figura 6.22 muestra el paso utilizado para no permitir el ingreso de valores nulos si no son necesarios.

3. Filtro para edad: En la vista DM_EASPIRANTEH no existe un atributo que guarde el valor de la edad, este dato se obtiene con la función TRUNC que recibe como parámetro la fecha en que se registro y fecha de nacimiento del aspirante.

   Algunos valores obtenidos con la operación anterior resultaban ilógicos y/o no concuerdan con el nivel académico deseado, para estos casos se aplicaron una serie de filtros en los cuales se asigna un valor por default dependiendo del nivel de la carrera a la que se aspira. La figura 6.23 muestra el procedimiento para el caso de la preparatoria.

4. Modificar edad: Los valores mencionados en el paso anterior pasan por una serie de códigos Java Script que modifican este valor antes de cargarse a la dimensión DIM_EASPIRANTEH. La figura 6.24 muestra el código utilizado para lograrlo.

5. Seleccionar atributos para la carga de datos en la dimensión: Algunos de los datos son innecesarios para la dimensión DIM_EASPIRANTEH solo fueron utilizados para la realización de algunos cálculos, por ello, serán eliminados de los valores de entrada; al finalizar este paso los datos son cargados en la dimensión. La figura 6.25 muestra el proceso descrito es este apartado.

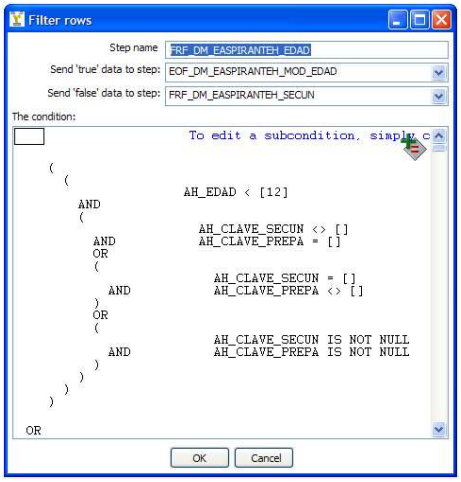


Figura 6. 22 Filtro para delimitar la edad

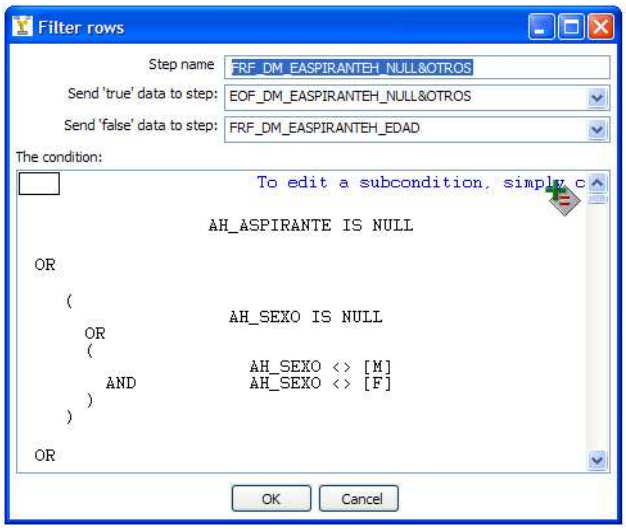


Figura 6. 23 Filtro de valores nulos (DM_EASPIRANTEH)

Java script :

Script 1

```
//Script here
if(AH_EDAD.getBigNumber()<=12 && AH_CLAVESECUN.getString()!=null && AH_CLAVEPREPA.getString()==null)
{
 AH_EDAD.setValue(15);
}
if(AH_EDAD.getBigNumber()<=12 && AH_CLAVESECUN.getString()==null && AH_CLAVEPREPA.getString()!=null)
{
 AH_EDAD.setValue(18);
}

if(AH_EDAD.getBigNumber()>=60 && AH_CLAVESECUN.getString()!=null && AH_CLAVEPREPA.getString()==null)
{
 AH_EDAD.setValue(15);
}
if(AH_EDAD.getBigNumber()>=60 && AH_CLAVESECUN.getString()==null && AH_CLAVEPREPA.getString()!=null)
{
 AH_EDAD.setValue(18);
}
```

Figura 6. 24 Java Script para modificar la edad según el nivel de la carrera elegida.

Proceso de Extracción-Transformación-Carga para DIM_EASPIRANTEH.

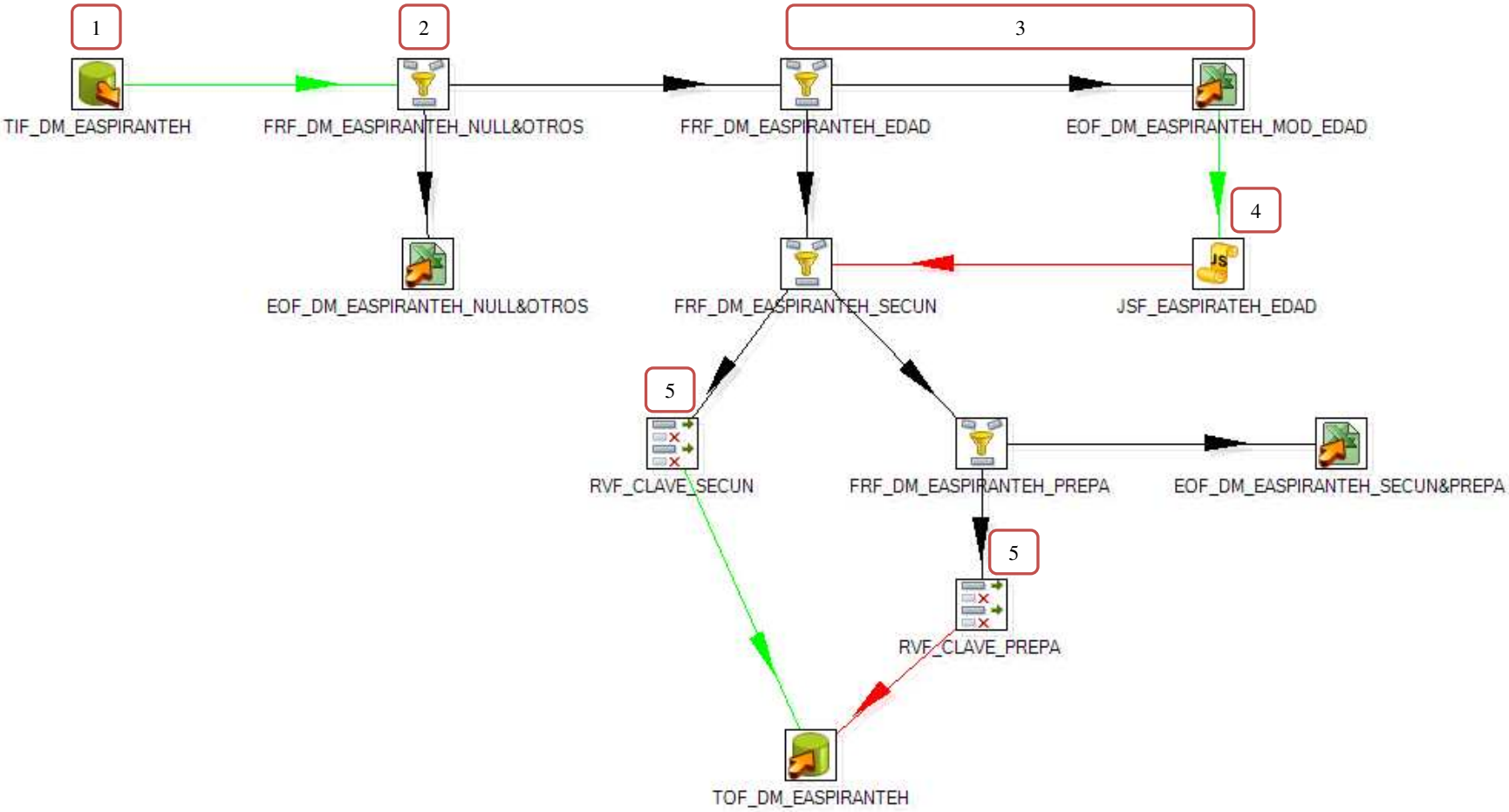


Figura 6. 25 Tratamiento de los datos relacionados con los Aspirantes (Encabezado).

❖ Proceso ETL para DIM_EASPIRANTED

Para la migración de los datos a la dimensión DIM_EASPIRANTED primero se crea un archivo con todas las claves de los aspirantes para poder compararlas con los detalles de los aspirantes, se realiza este procedimiento porque solo se deben migrar aquellas filas que contengan un encabezado, de lo contrario existiría una inconsistencia en los datos, ya que, algunos detalles no contarían con un encabezado para conocer a que aspirante le corresponde dichos datos.

La figura 6.26 muestra los dos pasos que se realizaron para la obtención de los datos. Estas claves se tomaron de la dimensión DIM_EASPIRANTEH, por tanto, no fue necesario aplicar ningún filtro.

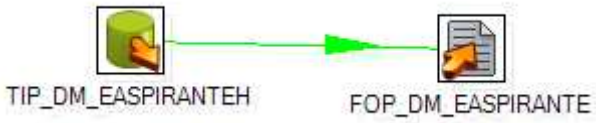


Figura 6. 26 Archivo que contiene las claves de los aspirantes a comparar.

1. Filtro para valores nulos: La figura 6.27 muestra el paso utilizado para no permitir el ingreso de valores nulos si no son necesarios.

2. Lectura de claves de los aspirantes: El archivo generado con la dimensión DIM_EASPIRANTEH será la base para conocer las filas de registros de la vista DM_EASPIRANTED que serán cargadas a la dimensión DIM_EASPIRANTED.

3. Comparación de claves del encabezado y detalles de los aspirantes: se realiza una unión entre las claves del archivo descrito anteriormente y los datos de entrada de la vista DM_EASPIRANTED, aquellas claves, de dicha vista, que no se encuentren en el archivo no serán cargadas en la dimensión, ver figura 6.28 y 6.29.

4. Selección y carga de datos: Por último, se cargan los datos considerados consistentes para la dimensión. La figura 6.30 muestra todos los pasos descritos en esta etapa.

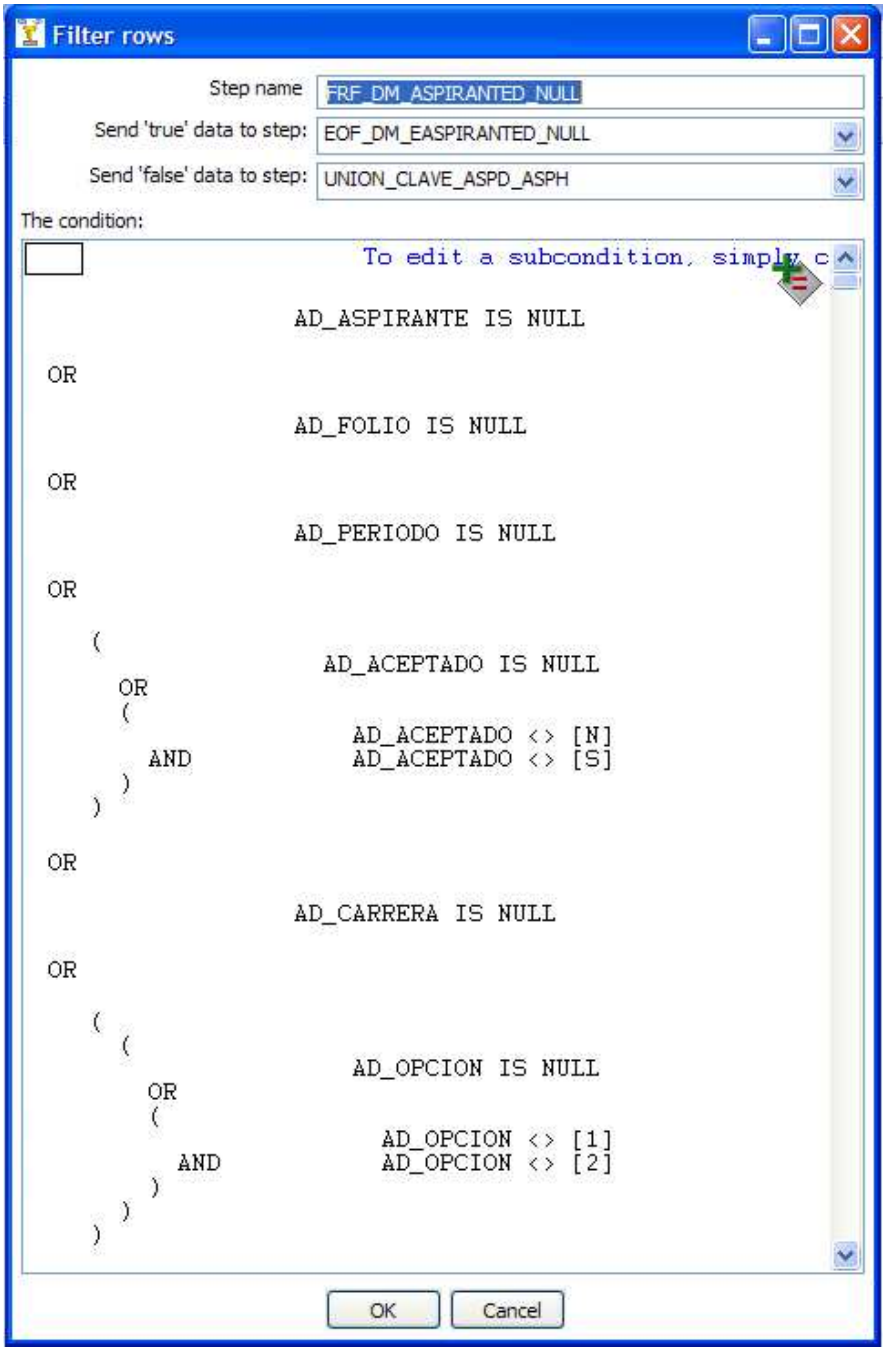


Figura 6. 27 Filtro de valores nulos y otros.

Figura 6. 28 Unión de AH_ASPIRANTE y AD_ASPIRANTE

El apartado 6.6 se realizo utilizando la herramienta de PENTAHO BI llamada kettle.

- **Modelo general para el tratamiento de los datos**

La figura 6.31 indica la serie de pasos a realizarse para migrar los registros de la Base de Datos Origen a las diferentes dimensiones de los Data Mart. Basta con ejecutar la aplicación para que el proceso inicie automáticamente hasta finalizarlo. La figura 6.32 muestra la ejecución de la trabajo.

Proceso de Extracción-Transformación-Carga para DIM_EASPIRANTED.

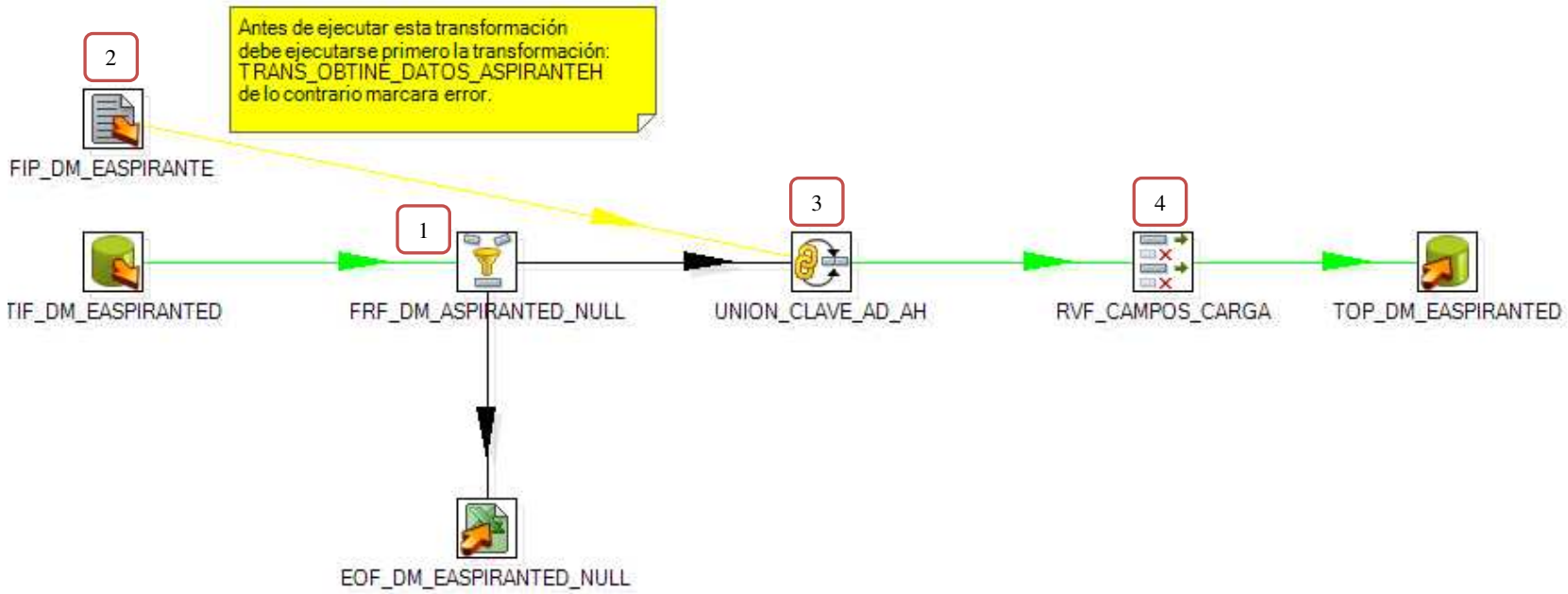


Figura 6. 29 Tratamiento de los datos relacionados con los Aspirantes (Detalles).

Modelo general para el proceso ETL de las dimensiones:

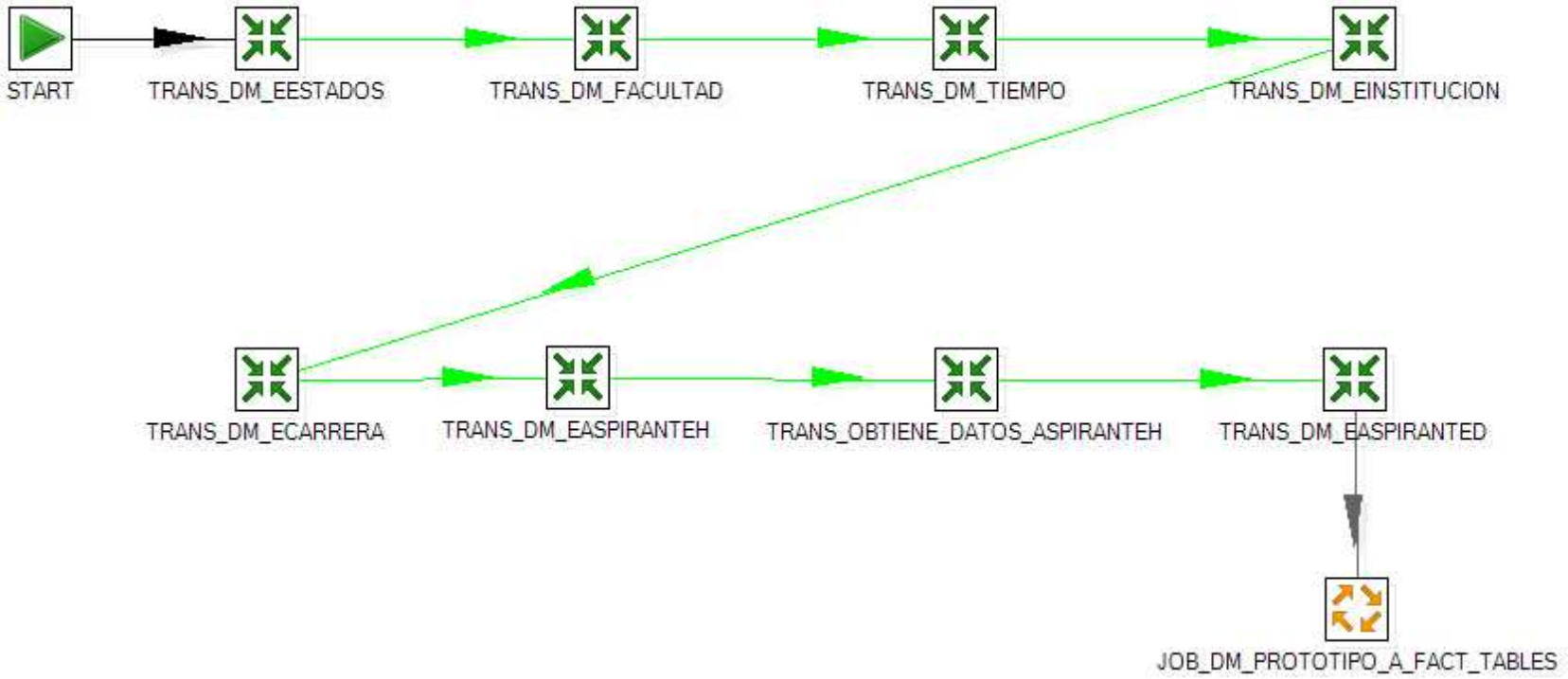


Figura 6. 30 Modelo general para el proceso ETL de las dimensiones de los diferentes Data Mart.

Ejecución de la transformación DIM_EASPIRANTEH:

Execution Results

Execution History | Logging | Step Metrics | Performance Graph

	Nombre paso	Numero Copia	Leído	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo	Velocidad (r/s)
1	FRF_DM_EASPIRANTEH_NULL&OTROS	0	66820	66820	0	0	0	0	0	Finalizado	85.4	782.5
2	FRF_DM_EASPIRANTEH_EDAD	0	66820	66820	0	0	0	0	0	Finalizado	104.7	638.2
3	EOF_DM_EASPIRANTEH_MOD_EDAD	0	4	4	0	5	0	0	0	Finalizado	105.4	0.0
4	JSF_EASPIRATEH_EDAD	0	4	4	0	0	0	0	0	Finalizado	105.1	0.0
5	EOF_DM_EASPIRANTEH_NULL&OTROS	0	0	0	0	0	0	0	0	Finalizado	85.8	0.0
6	TIF_DM_EASPIRANTEH	0	0	66820	66820	0	0	0	0	Finalizado	85.4	782.5
7	FRF_DM_EASPIRANTEH_SECUN	0	66820	66820	0	0	0	0	0	Finalizado	136.3	490.3
8	CLAVE_SECUN	0	21114	21114	0	0	0	0	0	Finalizado	136.3	154.9
9	TOF_DM_EASPIRANTEH	0	66807	66807	0	66807	0	0	0	Finalizado	219.8	303.9
10	CLAVE_PREPA	0	45693	45693	0	0	0	0	0	Finalizado	196.0	233.0
11	FRF_DM_EASPIRANTEH_PREPA	0	45706	45706	0	0	0	0	0	Finalizado	172.3	265.3
12	EOF_DM_EASPIRANTEH_SECUN&PREPA	0	13	13	0	14	0	0	0	Finalizado	172.3	0.0
13	JSF_EASPIRANTEH_CLAVE_INST	0	0	0	0	0	0	0	0	Finalizado	0.0	0.0

Figura 6. 31 Ejecución de la transformación DIM_EASPIRANTEH

6.6 Migración de los datos a utilizar (hechos)

Los datos almacenos en las dimensiones de los Data Mart, lo siguiente es insertar datos en las tablas de hechos, los cuales, en su mayoría, son valores que resultan de sumatorias, promedios, redondeos, etc., de los datos contenidos en las dimensiones.

En las siguientes páginas se hará una descripción general sobre este proceso. La nomenclatura utilizada en este apartado será la misma que en la sección 6.6.1.

La siguiente figura muestra el proceso utilizado para insertar algunos valores en la tabla de hecho FACT_GENERAL_LIC, tales como, total de aspirantes a nivel licenciatura, total de aspirantes del sexo Femenino, total de aspirantes del sexo Masculino, promedio de edad de aspirantes aceptados y no aceptados.

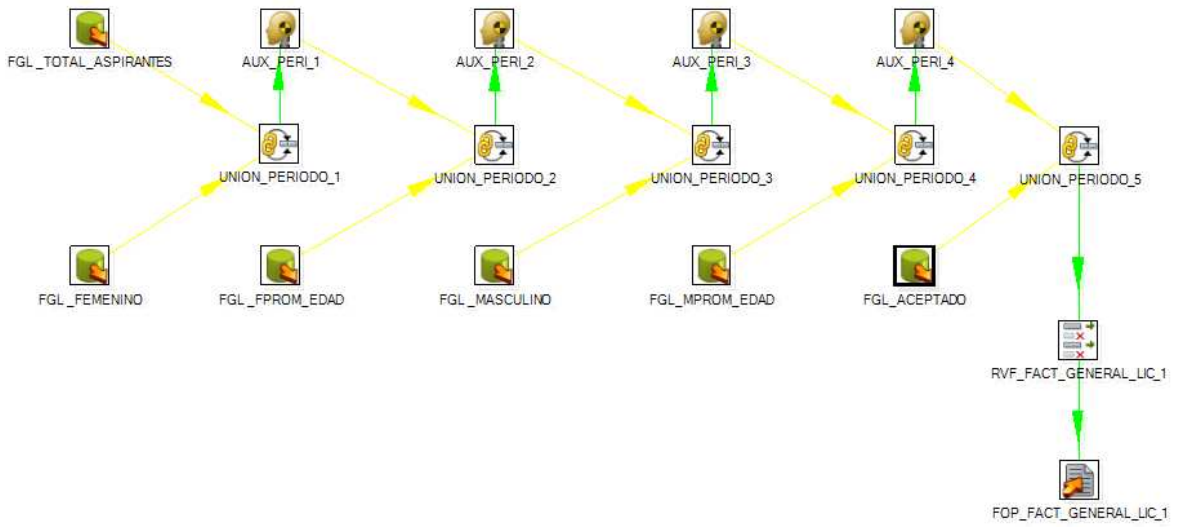


Figura 6. 32 Proceso para inserción de datos en FACT_GENERAL_LIC

En cada uno de las tablas de entrada se tiene una consulta para obtener un valor determinado, dichas consultas están delimitadas dependiendo del nivel de detalle deseado, enseguida se describen algunas de estas consultas.

Consulta: Total de aspirantes para licenciatura (nivel general).

```
SELECT AH_PERIODO,
COUNT(AH_ASPIRANTE) AS FGL_TOTAL_ASPIRANTES
FROM DIM_EASPIRANTEH
     INNER JOIN DIM_TIEMPO
     ON AH_PERIODO = TIM_PERIODO
WHERE AH_ASPIRANTE
  IN( SELECT AH_ASPIRANTE
     FROM DIM_EASPIRANTEH, DIM_EASPIRANTED
  WHERE AD_ASPIRANTE = AH_ASPIRANTE
  AND AD_CARR_PERI
     IN(SELECT CR_CLAVE
          FROM DIM_CARRERA
     WHERE CR_NIVEL = ?))
GROUP BY TIM_PERIODO
```

Al ejecutar la consulta se obtiene el total de aspirantes por periodo (semestre) de nivel licenciatura.

Consulta: Total de aspirantes del sexo femenino para licenciatura (nivel general).

```
SELECT AH_PERIODO,
COUNT(AH_ASPIRANTE) AS FGL_FEMENINO
FROM DIM_EASPIRANTEH
     INNER JOIN DIM_TIEMPO
     ON AH_PERIODO = TIM_PERIODO
WHERE AH_ASPIRANTE
  IN( SELECT AH_ASPIRANTE
     FROM DIM_EASPIRANTEH, DIM_EASPIRANTED
  WHERE AD_ASPIRANTE = AH_ASPIRANTE
  AND AD_CARR_PERI
     IN(SELECT CR_CLAVE
          FROM DIM_CARRERA
     WHERE CR_NIVEL = ?))
     AND AH_SEXO = 'F'
GROUP BY TIM_PERIODO
```

Con la consulta anterior se calcula el número de aspirantes del sexo femenino por periodo de nivel licenciatura.

Consulta: Total de aspirantes aceptados para licenciatura (nivel general).

```
SELECT AH_PERIODO,
COUNT(AH_ASPIRANTE) AS FGL_ACEPTADO
FROM DIM_EASPIRANTEH
     INNER JOIN DIM_TIEMPO
     ON AH_PERIODO = TIM_PERIODO
WHERE AH_ASPIRANTE
  IN( SELECT AH_ASPIRANTE
     FROM DIM_EASPIRANTEH, DIM_EASPIRANTED
  WHERE AD_ASPIRANTE = AH_ASPIRANTE
  AND AD_ACEPTADO = 'S'
  AND AD_CARR_PERI
     IN(SELECT CR_CLAVE
          FROM DIM_CARRERA
     WHERE CR_NIVEL = ?))
GROUP BY TIM_PERIODO
```

En la consulta se aplica un filtro que permite obtener solo aquellos aspirantes que, tras haber realizado el examen de admisión, fueron aceptados en alguna de las licenciaturas.

La consulta obtiene los datos por periodo.

Consulta: Total de aspirantes no aceptados del sexo Masculino para licenciatura (nivel general).

```
SELECT AH_PERIODO,
COUNT(AH_ASPIRANTE) AS FGL_NOAC_MASCULINO
FROM DIM_EASPIRANTEH
     INNER JOIN DIM_TIEMPO
     ON AH_PERIODO = TIM_PERIODO
WHERE AH_ASPIRANTE
  IN( SELECT AH_ASPIRANTE
     FROM DIM_EASPIRANTEH, DIM_EASPIRANTED
  WHERE AD_ASPIRANTE = AH_ASPIRANTE
  AND AD_ACEPTADO = 'N'
  AND AD_CARR_PERI
     IN(SELECT CR_CLAVE
          FROM DIM_CARRERA
     WHERE CR_NIVEL = ?))
AND AH_SEXO = 'M'
GROUP BY TIM PERIODO
```

Por último, esta consulta permite conocer el número de aspirantes no aceptados del sexo masculino por periodo escolar.

Para obtener los datos que serán insertados en los demás atributos de las tablas de hecho se desarrollaron consultas similares a las ya descritas, por tanto, no se considera necesario la explicación de cada una. La única diferencia entre cada una de ellas es el nivel de detalle que se desea lograr.

CAPÍTULO 7. CONCLUSIONES

El proceso de convertir datos en información e información en conocimiento para cualquier tipo de organización, llamase de gobierno, privada, educativa, etc., es un proceso iterativo que requiere el apoyo de toda la organización.

En general, uno de los factores que más incide en la necesidad de implementar un Data Warehouse o Data Mart es la dificultad que se tiene para brindar, de forma eficiente, información consistente e integra, a consecuencia de que los modelos relacionales, de cada organización, fueron diseñados para soportar las operaciones rutinarias del negocio (sistemas operacionales), por ejemplo, altas de usuarios, control de inventario, etc., y no para ofrecer información que permita la toma de decisiones con mayor fundamento y no por meras especulaciones.

Al contar con un repositorio dedicado especialmente para aquellos datos que faciliten la obtención de información relevante de la institución permite ahorrar tiempo en su procesamiento y no afecta a los sistemas transaccionales de uso rutinario.

Business Intelligence no es una tecnología que se encuentre solo al alcance de las grandes compañías que están dispuestas a invertir fuertes sumas de dinero para su implementación, sino que, con la utilización de software libre también pueden acceder las pequeñas y medianas empresas, porque lo más importante de Business Intelligence consiste en la capacidad de poder generar conocimiento a partir de información confiable y a su vez, crear información con base a datos íntegros y consistentes; no el costo de la infraestructura para poder implementarla.

A lo largo del Trabajo de Investigación se plantea la forma en cómo se elaboraron los diferentes Data Marts, cuyas características se ajustan a las necesidades que la Institución tenía en ese momento. Al contar con los pasos descritos y probados se pueden utilizar como prototipo o guía para implementar otros Data Mart en cualquier departamento de la propia institución e incluso en distintas instituciones de educación, ya que se menciona que información se puede obtener a partir de los Data Marts, sin embargo, como cada institución educativa presenta particularidades que las hacen distintas de otras, los procesos ETL deban ser cambiados según las necesidades que se presenten.

Tener todos los datos consistentes y ordenados en los Data Marts brinda una fuente confiable y estandarizada para el desarrollo de futuros Data Marts o para la ampliación del alcance de los existentes, facilitando el desarrollo de estos.

7.1 Recomendaciones

- Identificar los datos inconsistentes (sucios) en la base de datos origen y buscar la mejor manera de tratarlos.
- Tener especial cuidado en la fase de análisis para evitar una reestrucutración en los procesos.
- Realizar varias pruebas a los datos que se obtienen de los Data Mart para verificar que después de estas pruebas se llega al mismo resultado siempre y cuando se apliquen los mismos filtros.
- Verificar que las consultas realmente estén cumpliendo solo con su cometido, pues puede ocurrir que se delimite la consulta más de lo debido o que se permita la entrada de datos que no son necesarios para el repositorio.

Referencias bibliográficas

[1] **Abril**, Diego O. y **Pérez**, José N., "Estado actual de las tecnologías de bodega de datos y OLAP aplicadas a bases de datos espaciales", Revista Ingeniería e Investigación, Vol. 27 No.1, Abril 2007, pp. 58-67.

[2] **Becker**, Bob, "Kimball University: The Subsystems of ETL Revisited", Kimball Group Octubre 2007, pp. 1-5

[3] **Boyno**, Edward A., "Extraction, Transformation and Loading in a Data Warehouse Course", EDSIG, 2003, pp. 1-7.

[4] **Chaudhuri, S. y Dayal, U.,** "An overview of data warehousing and OLAP technology", ACM SIGMOD Record, 1997, pp. 7-10

[5] **Codd**, E. F., **Codd**, S. B. y **Salley**, C. T., "Providing OLAP to user-analysts: An IT mandate." , E. F. Codd and Associates, 1993., pp. 2-5.

[6] **Colin**, J. White, March 1999, "The IBM Business Intelligence Software Solution", Internet, Disponible en http://www.sigmod.org/disc/disc99/disc/ibm/bisolution.pdf; accesado el 10/01/2008.

[7] **Estrada**, Aaron I., "Modelo de implementación de proyectos de Data Mining como una herramienta estratégica dentro de las empresas mexicanas", Tesis de Maestría, Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Monterrey, Abril 2003.

[8] **Fayyad**, Usama, **Piatetsky**, Gregory y **Smyth**, Padhraic. "From Data Mining to Knowledge Discovery in Databases". American Association for Artificial Intelligence, 1996. http://www.aaai.org/AITopics/assets/PDF/AIMag17-03-2-article.pdf, consultada el 09/12/2008

[9] **Hancock**, John C. y Toren Roger, *Practical Business Intelligence with SQL Server 2005*, Printed in the United States of America: Addison Wesley Professional, Agosto 2006.

[10] **Kaplan**, R S y **Norton**, D P, "The balanced scorecard: measures that drive performance", Harvard Business Review, Febrero 1992, pp. 71-80.

[11] **Kmonk**, J., "Viador information portal provides Web data access and reporting for the IRS". DM Review. 1999., pp. 18

[12] **Kimball**, Ralph, "The 38 Subsystems of ETL", Kimball Group, 4 de Diciembre 2004, pp. 1-5

[13] **Kimball**, Ralph and **Ross**, Margy. *The Data Warehouse Toolkit,* Second Edition, John Wiley and Sons, Inc., 2002, pp. 2-22, 35-90, 380-196.

[14] **Inmon**, W. H., *Building the Data Warehouse*. 4th edition, John Wiley & Sons, 1997.

[15] **Inmon**, W.H., "Tech Topic: What is a Data Warehouse?", Prism Solutions., Volumen 1. 1995.

[16] **Imhoff**, Claudia, **Galemmno**, Nicholas and **Geiger**, Jonathan, *Mastering Data Warehouse Design,* Ed. Kathryn Herman, Indianapolis, Indiana, por Wiley Publishing, 2003.

[17] **Javed**, Asad y **Rafique,** Sardar S., "Data Warehouse Maintenance", Tesis de Maestría, Lulea University of Technology, 2006.

[18] **Johnson**, Amy, "Data Warehousing", Computerworld, No. 33, 6 Diciembre 1999, pp. 74.

[19] **Rifaieh**, Rami y **Benharkat**, Nabila A., "Query-based Data Warehousing Tool", ACM, 8 de Noviembre 2002, 35-36.

[20] **Rosas**, Leopoldo, "Zombi: una arquitectura para el análisis de información que integra procesamiento en línea con minería de datos", Tesis de Maestría, Universidad de las Américas Puebla, Mayo 2005.

[21] **Rubio**, Gerardo, "Inteligencia de Negocios: Los cinco estilos de BI", SG: Software Guru, conocimiento en práctica, Enero 2006, pp. 24-26.

[22] **S. Gardner**, Stephen R., "Building the Data Warehouse", Communications of the ACM, Septiembre 1998/Vol.41, No. 9., pp.59

[23] **Santos**, Ricardo J. y **Bernardino**, Jorge, "Real-Time Data Warehouse Loading Methodology", ACM, Septiembre 2008, pp. 49.

[24] **Tamayo**, Marysol y **Moreno**, Francisco J., "Análisis del modelo de almacenamiento MOLAP frente al modelo de almacenamiento ROLAP", Revista de Ingeniería e Investigación, Diciembre de 2006, Vol. 26 No. 3, pp. 137.

[25] **T. Brown**, Tony, "Data Warehouse Implementation with the SAS System", SAS Institute Inc., Dallas, TX, pp. 1-11

[26] **Vallejos**, Sofia J., "Minería de Datos", Universidad Nacional del Nordeste 2006., pp 2-5

[27] **Wang**, John, *Encyclopedia of Data Warehousing and Mining*, Published in the United States of America by Idea Group Reference, 2006.

[28] **W. Staudt**, M., **Vaduva**, A. and **Vetterli** T., "Metadata Management and Data Warehousing", World Scientific Connecting Great Minds,The Department of Information Technology (IFI) at the University of Zurich, Mayo 1999.

[29] **W. Powell**, Gavin, *Oracle Data Warehouse Tuning for 10g*, El SEVIER, 2005, p. 5.

[30] **W. Kimball**, Ralph and **Caserta**, Joe, *The Data Warehousing ETL Toolking Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*, 2004, p. 352.

[31] **Wrembel**, Robert, **Koncilida**, Christian, *Data Warehouses and OLAP: Concepts, Architectures and Solutions*, 2007.

Direcciones de Internet:

[32] http://kettle.pentaho.org/, consultada el 25 de Abril del 2008.

[33] http://es.talend.com/index.php, consultada el 03 de Marzo del 2008.

[34] http://www.debian.org/ , consultada el 06 de Marzo del 2008

Glosario

Bases de datos relacionales: Es una base de datos en donde todos los datos visibles al usuario están organizados estrictamente como tablas de valores, y en donde todas las operaciones de la base de datos operan sobre estas tablas. Estas bases de datos son percibidas por los usuarios como una colección de relaciones normalizadas de diversos grados que varían con el tiempo.

CRM: Customer Relationship Management, La administración de la relación con los clientes, CRM, es parte de una estrategia de negocio centrada en el cliente. Una parte fundamental de su idea es, precisamente, la de recopilar la mayor cantidad de información posible sobre los clientes, para poder dar valor a la oferta. La empresa debe trabajar para conocer las necesidades de los mismos y así poder adelantar una oferta y mejorar la calidad en la atención.

Drag & drop: (Arrastrar y soltar) Técnica que permite relacionar dos programas, dos ventanas o dos partes de un mismo programa. Por ejemplo, es posible llevar un ícono que representa un archivo sobre el ícono de la papelera de reciclaje y el archivo pasará a la misma.

Drill-down: Usado comúnmente en los sistemas de información gerencial o de análisis de información, es la habilidad para poder navegar de lo general a lo particular en la información presentada. Por ejemplo, en un informe de ventas en una compañía, se debe poder "taladrar" en los datos de cada región mundial para obtener los datos por país, y en el total de un país para obtener la información de las ciudades dentro del país.

Dril-Up: Es el efecto contrario a drill -down. Significa ver menos nivel de detalle, sobre la jerarquía significa generalizar o sumarizar, es decir, subir en el árbol jerárquico.

ERP: (Enterprise Resource Planning). Sistema o Software administrativo que integra todas las áreas de una empresa (Como contabilidad, compras, o inventarios), mediante procesos transparentes y en tiempo real en bases de datos relacionales y centralizadas.

Feedback: Es un término anglosajón que se traduce por “retroalimentación”, donde se desarrolla el saber escuchar, procesar la información recibida y externar una respuesta a su entorno, más puede llegar a ser positivo o negativo.

Granularidad: Se refiere a la especificidad a la que se define un nivel de detalle en una tabla, es decir, si hablamos de una jerarquía la granularidad empieza por la parte más alta de la jerarquía, siendo la granularidad mínima, el nivel más bajo. En Data Warehouse, no solo existe granularidad para las tablas de hechos (Fact's), También existe granularidad en las dimensiones.

Humanware: Son los elementos humanos de aplicaciones específicas según el ámbito de la empresa o institución donde se instalen los computadores.

Inconsistencia: Una base de datos está inconsistente si dos datos que deberían ser iguales no lo son. Por ejemplo, un empleado aparece en una tabla como activo y en otra como jubilado.

Integridad: Mantener la integridad de una base de datos es asegurarse de que los datos que contiene son correctos, evitando datos inconsistentes o erróneos de cualquier otro tipo. Por ejemplo, que un empleado aparezca como perteneciente a un departamento que no existe en la tabla correspondiente.

Integridad referencial: Se refiere a las claves foráneas. Recordemos que una clave foránea es un atributo de una relación, cuyos valores se corresponden con los de una clave primaria en otra o en la misma relación. Este mecanismo se usa para establecer interrelaciones. Consiste en que si un atributo o conjunto de atributos se define como una clave foránea, sus valores deben existir en la tabla en que ese atributo es clave principal.

Modelo relacional: Propuesto por Codd a finales de los sesenta introduce la teoría de las relaciones en el campo de las BD. En este modelo los datos se estructuran en tablas manteniendo la independencia de esta estructura lógica respecto al modo de almacenamiento u otras características físicas. Las tablas se manejan mediante operaciones de la teoría de conjuntos y el álgebra relacional.

Normalización: Según el modelo relacional, las tablas deben definirse siguiendo una serie de reglas precisas para asegurarse de que no se producirán anomalías en la actualización de la base de datos. Para ello, es habitual que se necesite descomponer las tablas iníciales en otras más simplificadas que no presenten dichos problemas. Este proceso es lo que se conoce como normalización y es un método formalizado con diferentes niveles, a cada uno de los cuales se le llama forma normal.

Redundancia: Se llama redundancia al hecho de que los mismos datos estén almacenados más de una vez en la base de datos. Las redundancias además de suponer un consumo de recursos de almacenamiento pueden llevar a situaciones en las que un dato se actualice en una de sus ubicaciones y en otra no y se pierda la integridad de la BD, por tanto deben evitarse.

Redundancia controlada: En ocasiones, es necesario introducir voluntariamente redundancia en la BD por consideraciones de rendimiento. En estos casos los administradores del sistema repiten conscientemente algunos datos y, a la vez, preparan al sistema para mantener automáticamente las distintas copias y que no se pierda la integridad.

Repositorio: Base de datos central en herramientas de ayuda al desarrollo. El repositorio amplía el concepto de diccionario de datos para incluir toda la información que se va generando a lo largo del ciclo de vida del sistema.

SQL (Structured Query Language): Lenguaje de consulta estructurado, es un lenguaje declarativo de acceso a base de datos relacionales que permite especificar diversos tipos de operaciones en éstas. Una de sus características es el manejo del álgebra y el cálculo relacional permitiendo efectuar consultas con el fin de recuperar -de una forma sencilla- información de interés de una base de datos, así como también hacer cambios sobre ella.

Triggers: Es una clase que implementa una interfaz que dispone de un método run(...), en el cual se implementa la tarea que debe ser llevada a cabo, y un método error(...), en el cual se implementa la tarea que debe ser llevada a cabo en caso de que se produzca algún problema, en la mayoría de casos debería deshacer las acciones llevadas a cabo en el método run().

Anexos

ANEXO A: HERRAMIENTAS DE BUSINESS INTELLIGENCE

COGNOS INC.

Con sus productos se construyen cubos multidimensionales para analizar los resultados y tendencias de los procesos claves (*Cognos PowerPlay*); se generan reportes ad-hoc y reportes operativos (*Cognos ReportNet*); se construyen Tableros Balanceados de Indicadores estratégicos (*Cognos Metrics Manager*); y se preparan presupuestos, proyecciones y simulaciones financieras (*Cognos Finance, Cognos Planning*); todo lo anterior funcionando en ambiente Web, gráficas y seguridad, También se integra un software de notificación, el cual alerta a las personas claves vía e-mail o celular, la variación de sus indicadores de desempaño (*Cognos NoticeCast*). Con la solución end-to-end de Cognos se obtienen resultados tangibles desde la construcción del primer data mart hasta la elaboración del Data warehouse Corporativo, incluyendo los más exigentes procesos de extracción y transformación de datos (*Cognos DecisionStream*). *IBM DataStage and QualityStage* le permite integrar estrechamente información empresarial, a pesar de tener muchas fuentes o destinos y plazos breves de tiempo. DataStage entrega tres capacidades claves que son necesarias para el éxito en la integración de datos empresariales: la conectividad más amplia para acceder fácil y rápidamente a cualquier sistema fuente o destino, herramientas avanzadas de desarrollo y mantenimiento que aceleran la implementación y simplifican la administración, y una plataforma escalable que pueda manejar fácilmente los volúmenes masivos actuales de datos corporativos. Analysis es una de las principales capacidades de IBM Cognos 8 Business Intelligence, un único producto que proporciona completas capacidades de BI en una arquitectura probada. *Analysis* permite la exploración guiada y el análisis de información relacionado con todas las dimensiones de su negocio, con independencia de donde se encuentren almacenados los datos. Analice y genere informes a partir de fuentes OLAP (*online analytical processing*) y fuentes de datos relacionales basadas en dimensiones.

Requerimientos:

- Sistemas Operativos: UNIX: IBM AIX, HP-UX, Sun Solaris, Compaq Tru 64; Windows NT, Win95, Win98, Win2000, Windows XP.
- Bases de Datos Soportados: Oracle, Informix, DB2, SQL Server, Sybase, etc.
- Acceso vía ODBC o conectividad directa.

Oracle Business Intelligence Suite

Es un miembro de la familia de productos Oracle Fusion Middleware es la plataforma más completa para la inteligencia de negocios (BI) disponible en la actualidad, cubriendo un amplio espectro de necesidades de inteligencia de negocios, incluidos los tableros interactivos, el análisis ad-hoc, alertas e inteligencia proactivas, publicación e informes avanzados, análisis predictivo en tiempo real, análisis de tecnología móvil, y mucho más.

OBI Suite - Oracle Business Intelligence Suite

Productos:

- Oracle BI Suite Enterprise Edition (EE) Oracle BI Suite Enterprise Edition (EE) es una plataforma de BI completa e innovadora de próxima generación que ofrece la mejor base para crear soluciones BI empresariales desde fuentes de datos heterogéneas para la distribución de datos, ya sean con bases de datos Oracle.

- Oracle BI EE está diseñado para un uso extensivo, con nuevos niveles de uso y alcance con el fin de brindar a un público más amplio conocimientos integrales y puntuales. Oracle BI EE también representa la base tecnológica para la inteligencia de negocios en las aplicaciones Oracle Fusion.

- Oracle BI Standard Edition (SE) Oracle BI Standard Edition (SE), que incluye Oracle Discoverer, está optimizado para trabajar con los datos y las aplicaciones Oracle, y ofrecer análisis e inteligencia avanzados al menor costo total.

- Oracle BI Publisher (inglés) Oracle BI Publisher, también denominado Oracle XML Publisher, ofrece la solución más eficiente y escalable para informes y publicaciones, disponible para entornos complejos y distribuidos. Disponible con Oracle BI Suite EE o como solución independiente, Oracle BI Publisher brinda una arquitectura central para generar y proporcionar información a los empleados, clientes y socios comerciales, tanto de manera segura como en el formato adecuado.

- ODI -Oracle Data Integrator ofrece la mejor alternativa para la integración de información en tiempo real, con una oferta adecuada a las necesidades de cada organización. La plataforma de integración de Oracle ODI pionera en el concepto de herramientas para extracción, transformación y carga de información (ETL) de cuarta generación, incluye claros beneficios como:

interfase completamente gráfica, adaptabilidad a la problemática de cada instalación, tiempo de desarrollo reducido, control de datos integrado, procesamiento en batch y/o tiempo real, compatibilidad con todas las fuentes de información incluyendo soluciones ERP y CRM, además de una instalación e implementación muy rápida sobre cualquier plataforma.

La plataforma Pentaho Open Source Business Intelligence

La plataforma Open Source Pentaho Business Intelligence cubre muy amplias necesidades de Análisis de los Datos y de los Informes empresariales. Las soluciones de Pentaho están escritas en Java y tienen un ambiente de implementación también basado en Java. Eso hace que Pentaho es una solución muy flexible para cubrir una amplia gama de necesidades empresariales, tanto las típicas como las sofisticadas y especificas al negocio.

Los módulos de la plataforma Pentaho BI son:

- Reporting - un modulo de los informes ofrece la solución adecuada a las necesidades de los usuarios. Pentaho Reporting es una solución basada en el proyecto JFreeReport y permite generar informes ágil y de gran capacidad. Pentaho Reporting permite la distribución de los resultados del análisis en múltiples formatos, todos los informes incluyen la opción de imprimir o exportar a formato PDF, XLS, HTML y texto. Los reportes Pentaho permiten también programación de tareas y ejecución automática de informes con una determinada periodicidad.

- Análisis - Pentaho Análisis suministra a los usuarios un sistema avanzado de análisis de información. Con uso de las tablas dinámicas (pivot tables, crosstabs), generadas por Mondrian y JPivot, el usuario puede navegar por los datos, ajustando la visión de los datos, los filtros de visualización, añadiendo o quitando los campos de agregación. Los datos pueden ser representados en una forma de SVG o Flash, los dashboards widgets, o también integrados con los sistemas de minería de datos y los portales web (portlets). Además, con el Microsoft Excel Analysis Services, se puede analizar los datos dinámicos en Microsoft Excel (usando la conexión a OLAP server Mondrian).

- Dashboards - todos los componentes del modulo Pentaho Reporting y Pentaho Análisis pueden formar parte de un Dashboard. En Pentaho Dashboards es muy fácil incorporar una gran variedad en tipos de gráficos, tablas y velocímetros (dashboard widgets) e integrarlos con los

Portlets JSP, en donde podrá visualizar informes, gráficos y análisis OLAP.

- Data Mining - análisis en Pentaho se realiza con una herramienta WeKa.

- Integración de Datos - se realiza con una herramienta Kettle ETL (Pentaho Data Integration) que permite implementar los procesos ETL. Últimamente Pentaho lanzó una nueva versión PDI 3.0 que marcó un gran paso adelante en OSBI ETL y que hizo Pentaho Data Integration una alternativa interesante para las herramientas comerciales.

Artus

Es una familia de productos de Business Intelligence con la cual puede construir Sistemas de Información Ejecutiva (EIS) y Sistemas de Soporte de Decisiones (DSS), para la comunicación y la administración de la empresa.

La comunicación se logra con Artus al medir y analizar los principales indicadores de desempeño de su empresa (KPIs) y poner ese conocimiento en las máquinas de los tomadores de decisiones, quienes podrán accesar en línea los datos críticos que manejan el negocio. La administración se logra en el momento en que pueden tomarse decisiones basadas en los indicadores mostrados.

Con Artus puede construir pequeños almacenes de datos multidimensionales con la información relevante para apoyar una decisión y la estructura que permita la explotación desde una perspectiva de negocios. Sin embargo, *su principal función no es la construcción*, sino la explotación de los almacenes.

La base de datos multidimensional con información sumarizada y agrupada es nuestro punto de partida, en ella existen dimensiones e indicadores de negocio determinados mediante un análisis previo. *Artus* es una herramienta OLAP (On-Line Analytical Processing) que emplea la base de datos multidimensional para analizar el negocio desde cualquier ángulo concebible como ventas por vendedor, margen de utilidad por línea de producto por cliente, etc. La ventaja es que la información se organiza en dimensiones que permiten su manejo en la forma como lo hace en la vida real, de la misma forma se plasma al usuario la misma información que es un resumen de la empresa, en términos de su perspectiva. Esto es la forma como una persona identifica y soluciona los problemas cotidianos, teniendo primeramente información sumarizada, por ejemplo: reporte mensual de ventas, una vez que lo visualiza, si el reporte está en un nivel calificado como 'normal' o

'aceptable' para el cumplimiento de los objetivos, se canaliza y termina siendo archivado. Sin embargo, en caso de observar cifras que no cumplan con los objetivos, que lo excedan o simplemente por conocer la distribución de las ventas, lo que comúnmente se hace es investigar el detalle de las ventas por sucursal, por zona, por vendedor o por la dimensión que aplique a la empresa.

Precisamente Artus hace esto, extrae información relevante de fuentes de datos dispersas y los pone en una base de datos común y multidimensional, con una estructura muy distinta a la operacional, posteriormente, extrae el verdadero valor de la base de datos a través de consultas cotidianas y predefinidas, las cuales poseen información agrupada que constantemente es monitoreada. El tomador de decisiones monitorea, pero también puede conocer el detalle que explica el por qué de cierto indicador. Esa función es también realizada con Artus valiéndose del análisis OLAP. El resultado es conocimiento global y detallado de la organización para fundamentar decisiones que mejoren el funcionamiento de la empresa.

Artus ofrece, además, otras herramientas para el análisis de información en la línea de Business Intelligence, como las proyecciones, tendencias, alertas ejecutivas, reporteo, comparaciones, etc. Todas ellas con la misma idea de comunicar y corregir los esfuerzos para alcanzar la visión corporativa, con el enfoque que los administradores requieren.

SAS

SAS ofrece una suite completa de soluciones de gestión de datos y software analítico de reconocido prestigio. Tanto si se necesita un software hecho a medida o una solución que responda a una tarea específica para fortalecer la infraestructura de inteligencia SAS ofrece la solución deseada.

Las soluciones SAS ofrecen el conocimiento analítico para la toma de desiciones, entre ellas están:

- SAS: Entrerprise Miner: Muestra pautas y tendencias explica resultados conocidos e identifica factores que permiten asegurar efectos deseados.

- SAS/OR: Proporciona un potente conjuntos de técnicas de análisis de optimización, gestión de proyectos, simulación y decisión, identificando las acciones que producirán mejores resultados.

- SAS-STAT: Proporciona herramientas para realizar análisis tradicionales hasta la generación de modelos predictivos adaptados a las necesidades.

- SAS-ETS: ofrece una amplia selección de series temporales, técnicas de previsión y econométricas que permiten planificar, predecir y simular procesos de negocio para mejorar la planificación estratégica táctica.

Enhydra: Talend Open Data Solutions

Consiste en extraer los datos de orígenes distintos, como por ejemplo base de datos, ficheros, aplicaciones, servicios Web, emails, etc, aplicando las transformaciones (join, lookup, deduplicación, cálculos, etc.) a estos datos y en transmitir los resultados a los sistemas target.

Las soluciones Open Source para la integración de datos de Talend cubren las siguientes necesidades:

- *Integración operacional de datos*
  En la mayoría de las organizaciones, se responde a la integración operacional de los datos realizando programas para cada necesidad específica. La migración y sincronización/duplicación de datos son las aplicaciones más comunes para la Integración de datos operacionales.

- *Migración de datos*
  Al actualizar a una nueva versión de una base de datos o de una aplicación, o al cambiar a un nuevo sistema, por ejemplo, los datos necesitan ser preservados en este nuevo sistema. El propósito de la migración de datos es transferir datos existentes a un nuevo ambiente. Necesita transformar los datos a un formato conveniente para el nuevo sistema, mientras que se preserva la información presente en el viejo.

- *Sincronización de datos*
  Existen muchos casos en los sistemas de información en que los datos se administran de forma independiente por múltiples aplicaciones o bases de datos. Sin embargo, es necesario mantener la coherencia entre dichos sistemas. La necesidad de la sincronización de datos puede ser permanente (sincronización entre los diferentes sistemas) o temporal (por ejemplo, durante una migración). La sincronización de datos incluye todos los procesos que mantienen sincronizados los datos entre las aplicaciones y las bases de datos.

- *ETL para Business Intelligence y el Data Warehousing*
  Los procesos de ETL (extracción, transformación y carga) son los componentes más críticos y de valor añadido de una infraestructura de

Business Intelligence (BI). Mientras que es normalmente invisible al usuario de la plataforma de BI, los procesos de ETL recuperan los datos de todos los sistemas operacionales y los pre-procesan para las herramientas de análisis y de reporting. La exactitud de toda la plataforma BI depende de los procesos de ETL.

Toda la información del software mencionado en este anexo fue tomada de los sitios oficiales de cada uno de ellos.

ANEXO B: INSTALACIÓN DEL SOFTWARE:

- ❖ Instalación de Linux Debian Etch 4

Esta instalación contiene varios pasos, los cuales no se mencionaran por completo en este documento, pues el asistente de instalación de Debian incluye una descripción clara en cada uno de los pasos. Por ello, solo describirán en términos generales.

Esta es una pequeña guía de los pasos por los que pasará durante el proceso de instalación:

1. Realice una copia de seguridad de los datos o documentación existentes en el disco duro donde planea realizar la instalación.

2. Reúna información sobre su sistema, así como toda la documentación que necesite antes de iniciar la instalación.

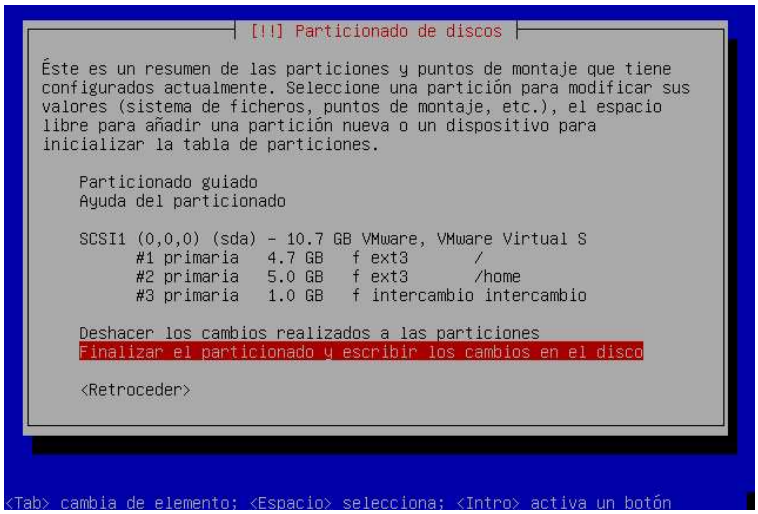
3. Cree un espacio particionable para Debian en su disco duro.

4. Descargue el programa del instalador de la página oficial de Debian.
5. Arranque el sistema de instalación. De ser necesario modifique el orden de booteo del BIOS, iniciando con la unidad de CD o DVD.

6. Elija el idioma para la instalación.

7. Active la conexión de red, si está disponible. Basta con tener conectado el cable de red al modem y a la maquina.

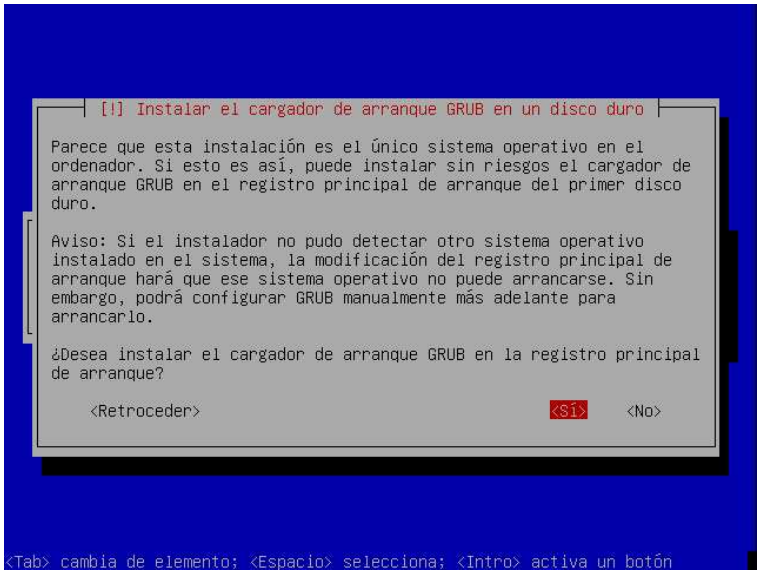
8. Cree y monte las particiones en las que instalará Debian.



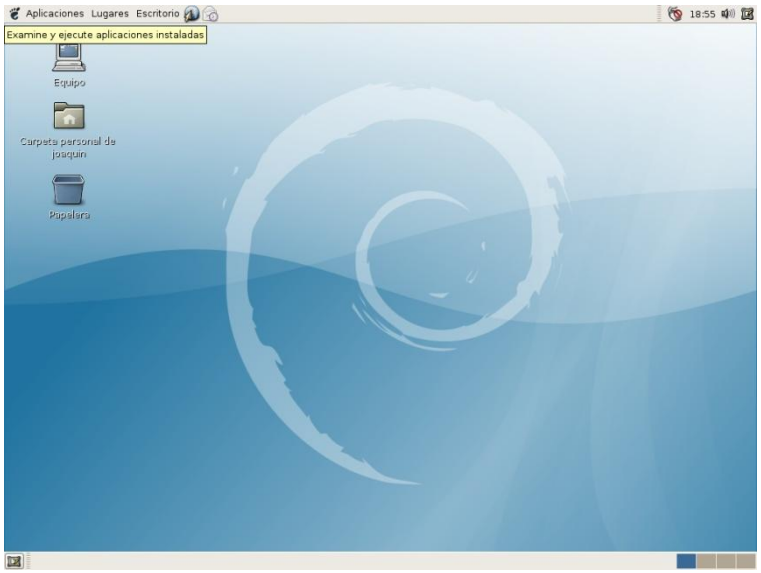
9. Configuración de usuarios y contraseñas.

10. Espere a la descarga/instalación/configuración automática del *sistema base*. Dependiendo de las características de la maquina (memoria, procesador, etc.) puede demorar varios minutos.

11. Instale un *gestor de arranque,* GRUB, que pueda iniciar Debian GNU/Linux y/o su sistema existente.



12. Inicie por primera vez el sistema que acaba de instalar.



❖ Instalación de JDK

1. Después de descargar el paquete jdk es necesario darle permisos de ejecución como sigue:

```
chmod –x jdk-1.5.bin
```

2. Al ejecutarlo pregunta o siguiente: *¿Do you agree to the above license terms? [yes or not],* se teclea *yes* y se presiona la tecla *enter*.

3. La instalación continúa de forma automática.

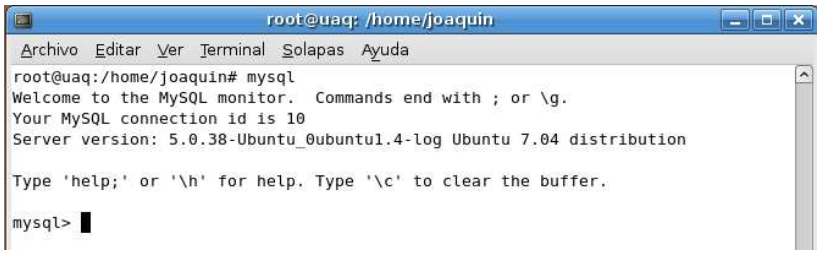
❖ Instalación de MySQL Server 5.0 y MySQL Client 5.0

1. Se abre una terminal y se teclea el siguiente comando:

```
apt-get install mysql-server mysql-client
```

2. Después de cierto tiempo aparece lanza la siguiente pregunta: *¿Desea continuar [S/n]?* tecleamos la letra S para continuar.

3. A continuación se ejecuta el comando MySQL para abrir el administrador de la base de datos:



¿Para qué se utilizo?
Es el sistema de gestión de base de datos que se utilizará para el almacenamiento de los datos (Data Mart) generados por KETTLE. Dichos datos ya habrán pasado por el proceso ETL (Extracción, Transformación y Carga).

❖ Instalación de Kettle

1. Descargar el archivo de página de Pentaho.

2. Después de esto, simplemente se descomprime el archivo *Kettle3.0.0.zip* en el directorio deseado. Para este caso se creó un directorio con el nombre "*Kettle*".

3. Por último es necesario hacer los scripts ejecutables:

```
cd Kettle
chmod +x *.sh
```

4. Para ejecutar la aplicación basta con teclear en consola: sh spoon.sh

```
sh spoon.sh
```

5. La siguiente figura muestra la pantalla de inicio del software:

¿Para qué se utilizo?
Permitió migrar los datos de Oracle a MySQL, para hacerlo cuenta con varios objetos (filtros, actualizaciones, alertas, errores, datos de entrada-salida), que se pueden entrelazar para obtener solo los datos que realmente se necesiten y con la mayor integridad y consistencia posible.