Análisis del proceso de carga del Sistema de Data Warehousing de Enseñanza de la Facultad de Ingeniería[‡]

Lorena Etcheverry, Pablo Gatto, Salvador Tercia
CSI, Instituto de Computación, Facultad de Ingeniería
Universidad de la República
lorenae@fing.edu.uy, pgatto@fing.edu.uy, stercia@fing.edu.uy
Diciembre 2005

Resumen: En este documento se describe el proceso de ETL del Sistema de Data Warehousing de Enseñanza de la Facultad de Ingeniería. Se analiza dicho proceso en tres niveles de abstracción, mostrando el flujo de datos en un macro-nivel que identifica los subprocesos que intervienen, pasando por un nivel intermedio que detalla desde un punto de vista conceptual las actividades en ciertos subprocesos hasta llegar a un nivel de detalle que expresa las operaciones involucradas en dichos subprocesos. El objetivo principal de este trabajo es describir y agrupar las actividades de ETL para un posterior análisis de la propagación de propiedades de calidad en dicho proceso.

Palabras clave: ETL, Data Warehouses, proceso de carga.

[‡] Este trabajo fue parcialmente financiado por Comisión Sectorial de Investigación Científica, Universidad de la República, Montevideo, Uruguay

1 Introducción

Un Data Warehouse (DW) es un conjunto de datos orientados a temas, integrados, no volátiles e históricos, organizados de tal forma que sirven de apoyo a la toma de decisiones [3], dado que permiten analizar la información consolidada según diferentes puntos de vista. Dicho proceso de consolidación de información involucra actividades de extracción de diversas fuentes de datos, transformación de la información necesaria y finalmente su carga en el DW. Usualmente se denomina a este proceso ETL, del inglés *Extraction, Transformation and Loading*.

Las transformaciones aplicadas a los datos provenientes de las distintas fuentes son básicamente de limpieza y de estructuración. Las transformaciones de limpieza son necesarias para asegurar la calidad de los datos finalmente almacenados en el DW e incluye entre otros, la corrección de errores, eliminación de redundancia y resolución de inconsistencias, así como el asegurar las reglas de negocio definidas. Los cambios en la estructura se realizan para adecuar los esquemas a las funcionalidades de un DW, e incluyen la adecuación al modelo de datos del DW, cambios de formato, operaciones de agregación, etc.

El objetivo principal de este reporte es analizar el proceso de ETL de un sistema de Data Warehousing concreto con el fin de caracterizar y categorizar las actividades que en éste aparecen.

El proceso de ETL a estudiar corresponde al Sistema de Data Warehousing de Enseñanza de la Facultad de Ingeniería [2]. La elección del caso de estudio se basa principalmente en la disponibilidad de código y de datos.

Se analizará el proceso de ETL en tres niveles de abstracción diferentes:

- **Nivel alto o macro:** en este nivel se describe el proceso de carga completo, su descomposición en subprocesos y sus interdependencias, brindando una visión global del mismo.
- **Nivel medio:** en este nivel se describe tanto el flujo de datos como el flujo de control de los subprocesos más relevantes de los identificados en el nivel anterior. En cada uno de estos subprocesos se identifican actividades de carga; que involucran secuencias de operaciones; desde un punto de vista conceptual, expresando así la semántica del mismo.
- **Nivel bajo o de detalle:** en este nivel se describe el flujo de datos de cada uno de los subprocesos, indicando cómo se implementan las actividades presentadas, mostrando las operaciones asociadas. La implementación se basa en sentencias SQL.

El trabajo realizado se enmarca en un proyecto más general en el cual se estudian propiedades de calidad de datos – en particular *data freshness* y *data accuracy* - y la propagación de las mismas en un sistema de información multi-fuente. En actividades

previas de dicho proyecto se generó un modelo de la propagación de estas propiedades [4] el cual se pretende validar. Dicha validación se realizará comparando las estimaciones realizadas aplicando dicho modelo con la medición efectiva de las propiedades de calidad en el sistema de Data Warehousing de Enseñanza.
Es un objetivo a futuro la caracterización general de las actividades que intervienen típicamente en los procesos de ETL de sistemas de Data Warehousing y el posterior estudio del impacto de dichas actividades en las propiedades de calidad de datos.

El reporte se organiza de la siguiente forma: en la sección 2 se presenta una breve descripción del caso de estudio, las dimensiones y los cubos considerados. En la sección 3 se presenta el proceso de carga a alto nivel. En la sección 4 se presenta el proceso de carga a nivel medio, en la sección 5 se presenta el proceso de carga a bajo nivel y por último, en la sección 6, se presentan conclusiones.

2 Descripción del caso de estudio

La Facultad de Ingeniería cuenta con un Sistema de Data Warehousing con información referente a las actividades de enseñanza desarrolladas en dicha institución, el cual asiste a la misma en la gestión y la toma de decisiones en aspectos vinculados a la enseñanza [2].

El sistema fue diseñado con el objetivo de analizar las actividades de los estudiantes en la Facultad. Cuenta con información sobre las carreras dictadas y las actividades que los estudiantes desarrollan en el marco de las mismas, permitiendo por ejemplo el estudio de los estudiantes por materia o asignatura, el avance por materia, las inscripciones, así como también realizar el estudio por cursos, institutos o generaciones, el avance y el desempeño por carrera, plan o perfil, entre otros.

Actualmente el sistema cuenta con cuatro cubos de análisis los cuales permiten estudiar el desempeño de los estudiantes. En particular se analizará el proceso de carga de dos de éstos y las dimensiones correspondientes:

- **Cubo Actuación**: en este cubo se mide la actuación de los estudiantes en las asignaturas, es decir los resultados obtenidos por los estudiantes para todas las carreras en que se encuentran inscriptos, permitiendo clasificar a los estudiantes según diferentes aspectos relevantes (sexo, generación, lugar de origen, instituto preuniversitario). Permite responder por ejemplo consultas sobre un conjunto de estudiantes, la actuación en una asignatura, o la trayectoria de los estudiantes de una generación a través de una asignatura.
- **Cubo Duración**: en este cubo se mide el tiempo que le lleva a cada estudiante culminar la o las carreras a las cuales se encuentra inscripto. En caso de no haber culminado aún se estima una fecha probable de finalización en función del ritmo de avance hasta el momento. Permite por ejemplo obtener la duración estimada de cada carrera para cada generación.

Las figuras 2.1 y 2.2 presentan respectivamente los esquemas conceptuales[1] de cada uno de estos cubos.

Figura 2.1 – Relación dimensional actuación

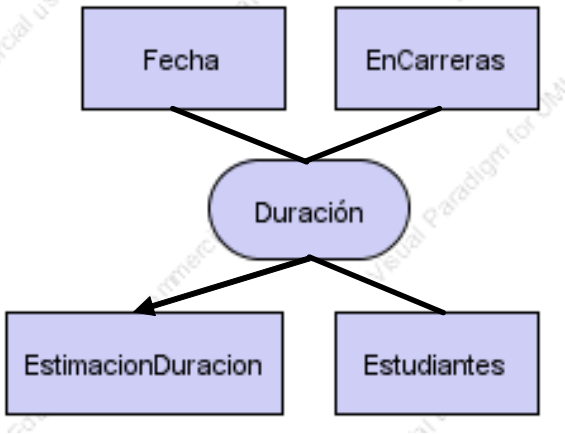

Figura 2.1 – Relación dimensional duración

En las figuras se observa qué dimensiones participan en cada uno de los cubos de interés. Estas son:

- Asignaturas: dimensión que modela la oferta curricular de la Facultad de Ingeniería. En la misma se dictan una serie de carreras. Cada carrera pertenece a

[1] Los esquemas conceptuales están expresados usando el modelo conceptual CMDM [1]

un plan de estudios y puede poseer uno o más ciclos. Cada ciclo se ordena en áreas temáticas o materias, las cuales a su vez agrupan asignaturas.

- EnCarreras: dimensión que modela las inscripciones a carreras de cada estudiante.

- Estudiantes: dimensión que modela la población estudiantil de la Facultad. Posee dos jerarquías de análisis: una en base al sexo del estudiante y otra en base a su procedencia geográfica.

- Fecha: dimensión que modela el tiempo

- Institutos: dimensión que modela los institutos de la Facultad, los cuales dictan las asignaturas

- InstitutosPreuniversitarios: dimensión que modela los institutos secundarios de los cuales provienen los estudiantes. Esta dimensión posee dos jerarquías: por tipo de acceso al instituto (público o privado) y por laicidad de la educación que imparten (laico o religioso)

- Lugares: dimensión que modela los lugares geográficos de procedencia de los estudiantes.

- Períodos: dimensión que modela los períodos de evaluación de la Facultad.

Los grupos de medidas correspondientes a los cubos Duración y Actuación respectivamente son:

- EstimacionDuracion, la cual es el resultado de cálculos realizados en cada carga y almacena la estimación de la duración en años de la carrera para un estudiante en determinado año.

- Resultados, la cual modela los resultados de actividades que pueden registrarse en las bases de bedelía. Entre ellos se encuentran los resultados de los cursos y los exámenes

3 Proceso de carga a nivel macro

La figura 3.1 representa el mapa general de la carga, identificando los procesos que cargan los cubos descritos en la sección anterior y el flujo de datos desde las tablas de la fuente de datos hasta las principales tablas del Data Warehouse:

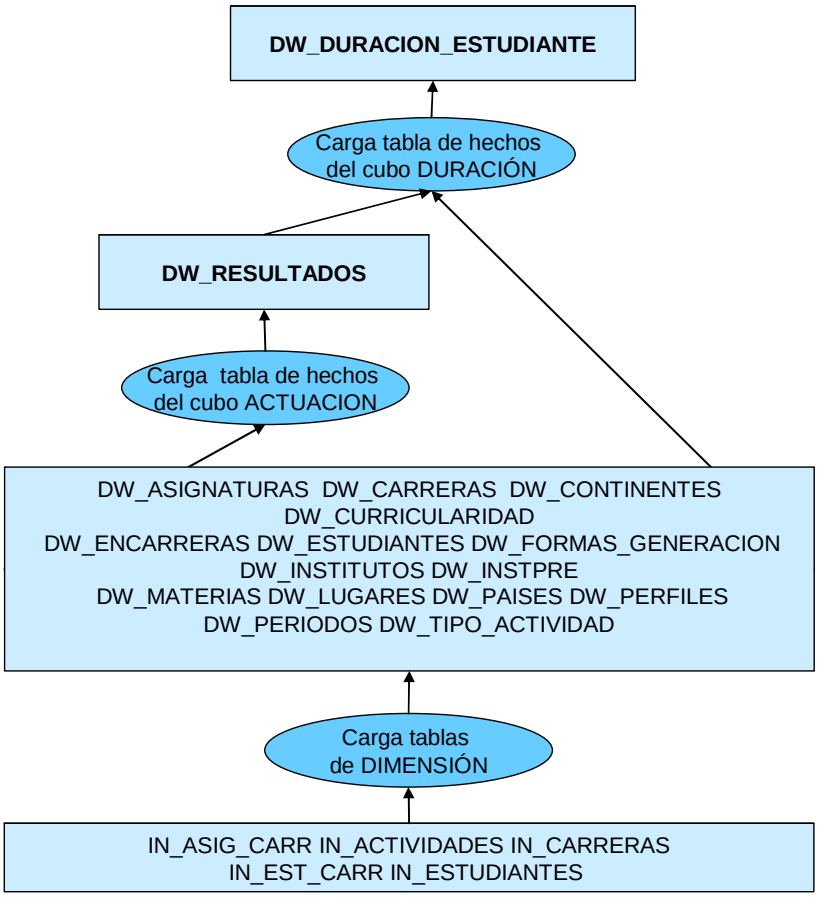

Figura 3.1 – Mapa general de la carga

A continuación se presenta, para cada cubo seleccionado, el proceso de carga a nivel macro de su fact table (o tabla de hechos) y de las dimensiones de análisis que intervienen. A este nivel se representa el flujo de datos vinculando los procesos de carga existentes con las tablas cargadas y las utilizadas para obtener y transformar los datos. Los rectángulos representan tablas, los óvalos procesos de carga y las aristas de este grafo dirigido modelan el flujo de datos, indicando las entradas y salidas de los mencionados procesos.

3.1 Dimensiones

Las figuras 3.2 y 3.3 representan el flujo de datos correspondiente al proceso de carga de las tablas de dimensión. El proceso se presenta en dos figuras únicamente por razones de claridad y espacio.

Se puede observar que existen dos fuentes de datos: la base de datos del sistema de Bedelías de la Facultad de Ingeniería (tablas nombradas con prefijo IN_) y un sistema legado de gestión desarrollado con anterioridad (tablas nombradas con prefijo BD_).

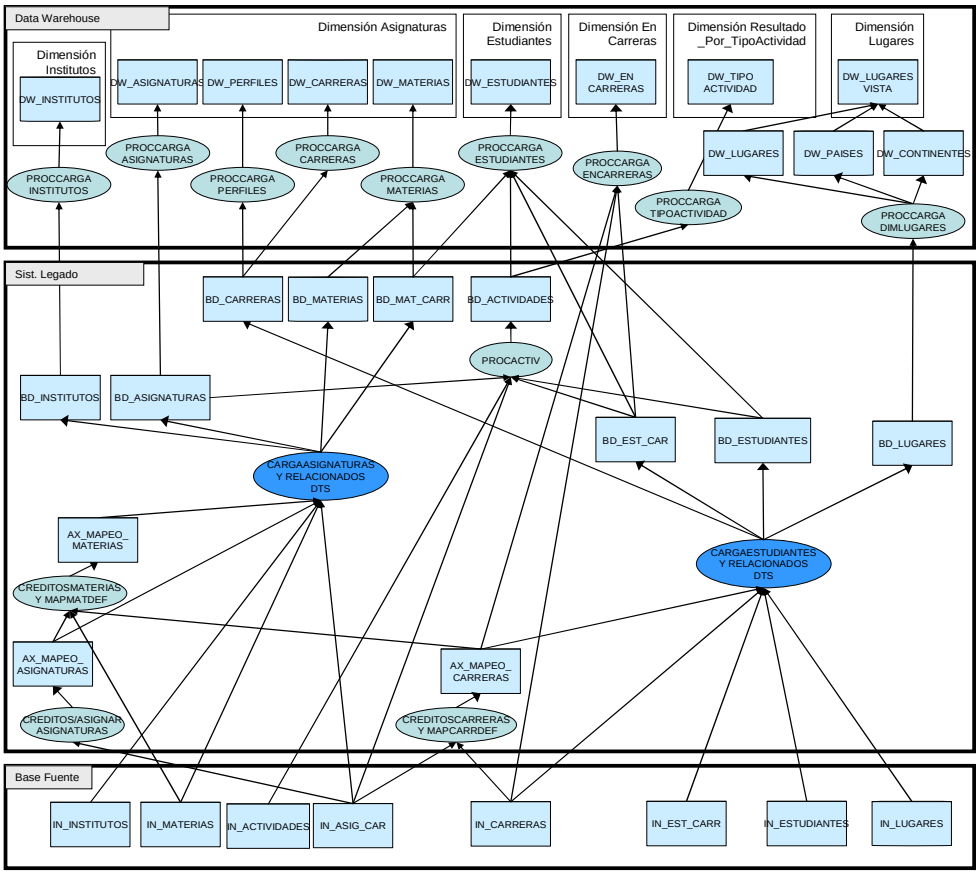

Figura 3.2 – Carga de tablas de dimensión (1)

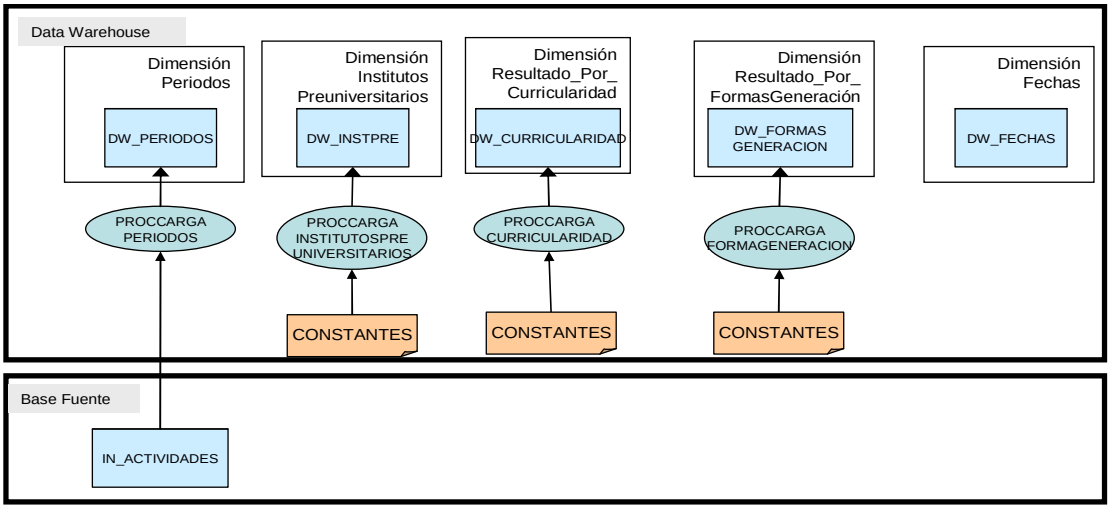

Figura 3.3 – Carga de tablas de dimensión (2)

3.2 Tabla de hechos del cubo Actuación

La figura 3.4 representa el flujo de datos correspondiente al proceso de carga de la fact table del cubo Actuación.

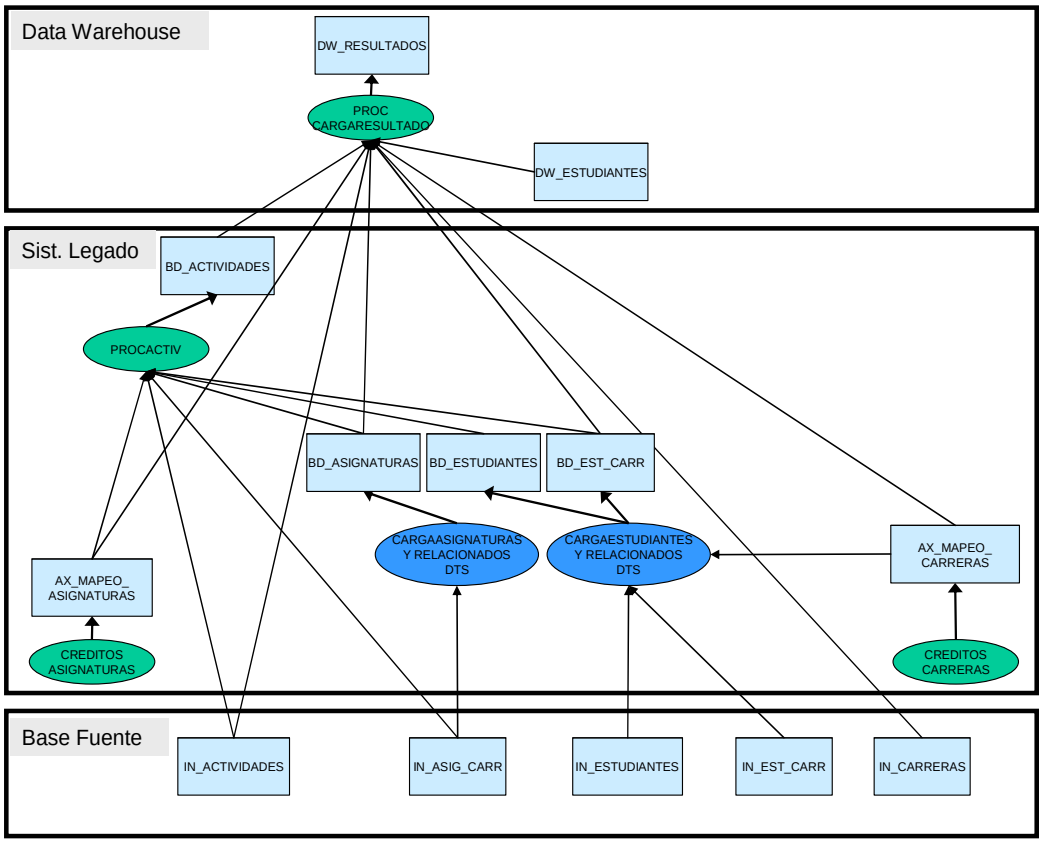

Figura 3.4 - Carga de la tabla de hechos del cubo Actuación

3.3 Tabla de hechos del cubo Duración

La figura 3.5 representa el flujo de datos correspondiente al proceso de carga de la Tabla de hechos del cubo Duración.

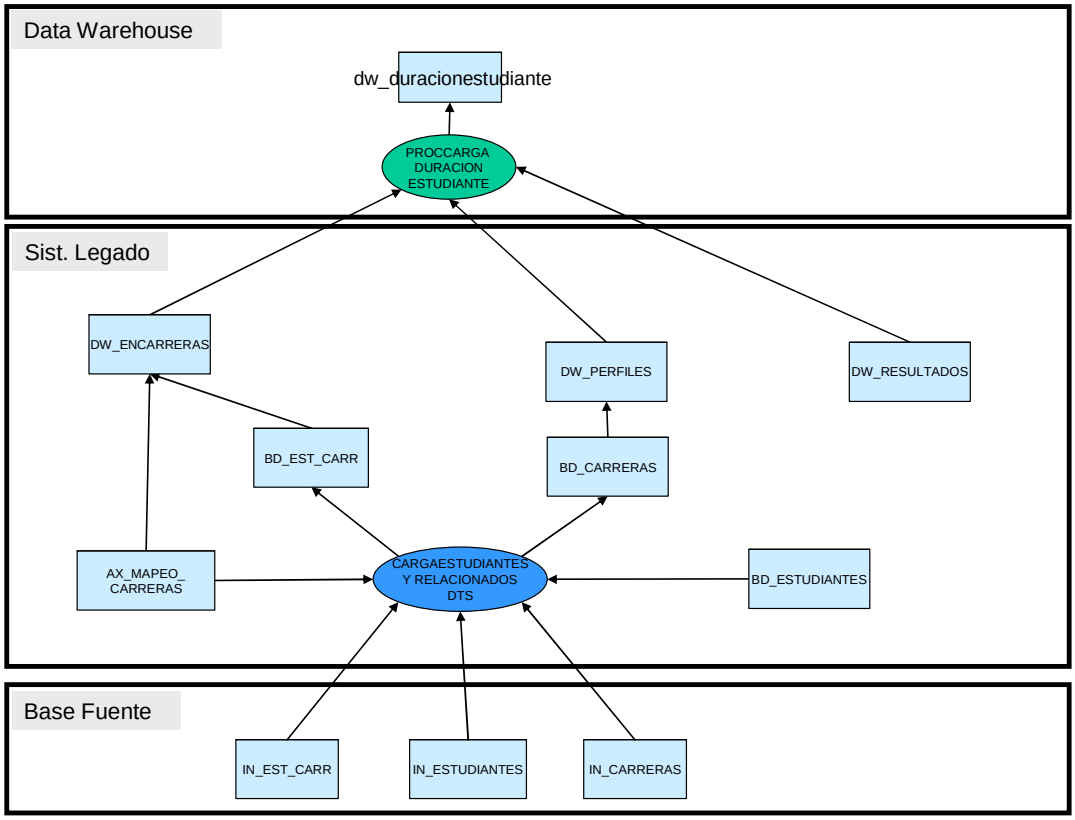

Figura 3.5 - Carga de la Tabla de hechos del cubo Duración[1]

4 Proceso de carga a nivel medio

A continuación se presentan los procesos de carga relevantes en un nivel de detalle medio, mostrando - para algunos procesos de las figuras anteriores – la relación existente entre los distintos tipos de actividades en las que se descompone cada proceso. Se representa el flujo de datos mostrando los tipos de actividades realizadas sobre los datos, desde su extracción de las tablas origen hasta su carga en las tablas destino. Estos tipos de actividades abstraen la semántica de un conjunto de operaciones sobre dichos datos.

Las actividades se clasifican según su semántica en los siguientes tipos:

- **Obtención de datos:** Extraer datos a partir de tablas origen.
- **Chequeo de existencia:** Verificar la existencia de un elemento previamente cargado en el Data Warehouse.
- **Transformación:** Ejecutar cualquier función que sea aplicada sobre los datos para transformarlos, por ejemplo:
  - **Cambio de Formatos:** Realizar una conversión en la presentación de los datos.
  - **Generación de códigos:** Generar identificadores, aplicando funciones sobre otros datos existentes.
  - **Cálculos:** Calcular un nuevo dato a partir de otros datos base.
- **Limpieza:** Efectuar cualquier actividad que procure identificar y corregir errores en los datos.
- **Inserción:** Cargar los datos en las tablas destino del Data Warehouse

Por simplicidad, en adelante se menciona a las actividades como sinónimo de sus tipos de actividad.

4.1 Dimensiones

Las figuras 4.1, 4.2 y 4.3 representan el flujo de datos del proceso de carga de las dimensiones. Estas corresponden a los procesos de carga de la capa Data Warehouse señalados en las figuras 3.2 y 3.3. Los procesos se presentan en tres figuras simplemente por razones de claridad y espacio.

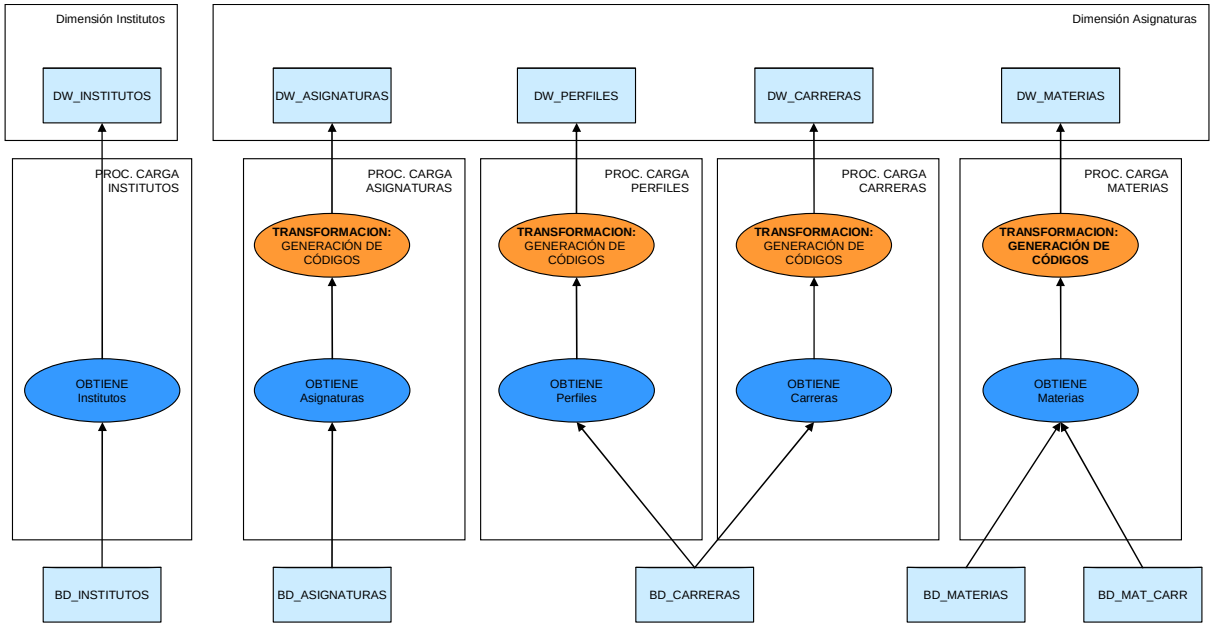

Figura 4.1 – Carga de tablas de dimension

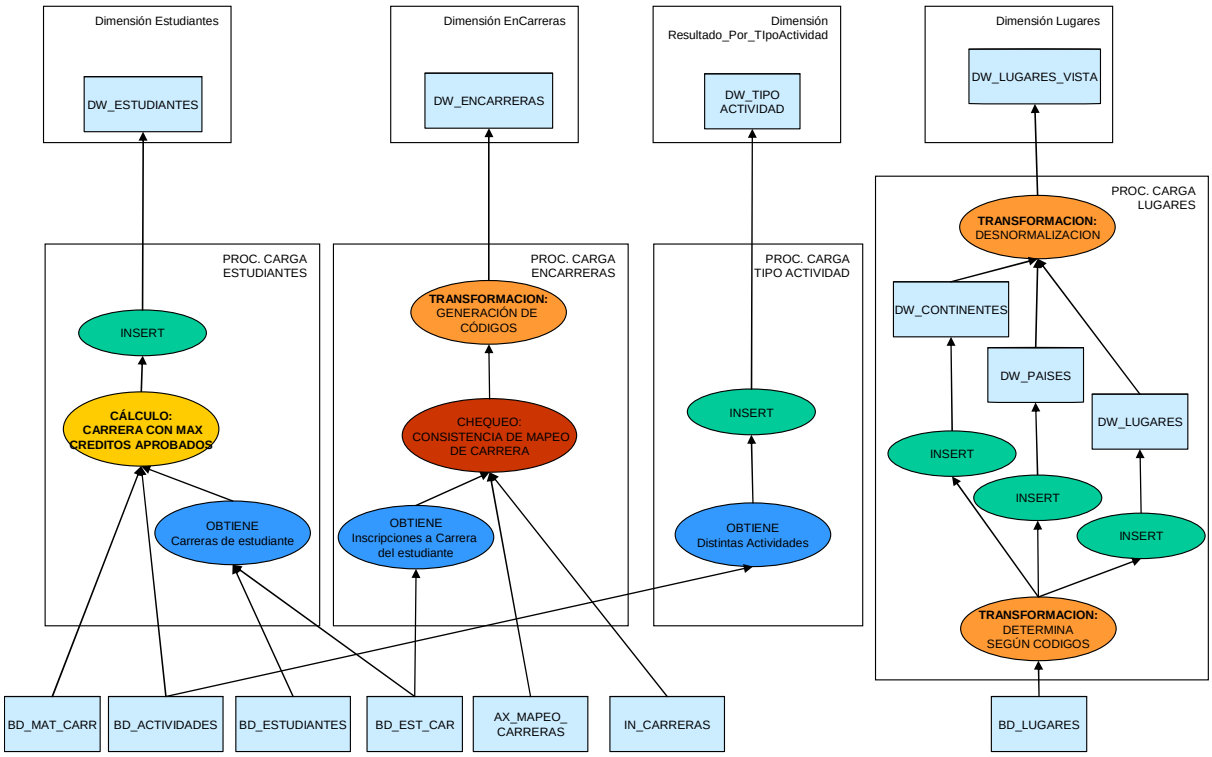

Figura 4.2 – Carga de tablas de dimensión (2)

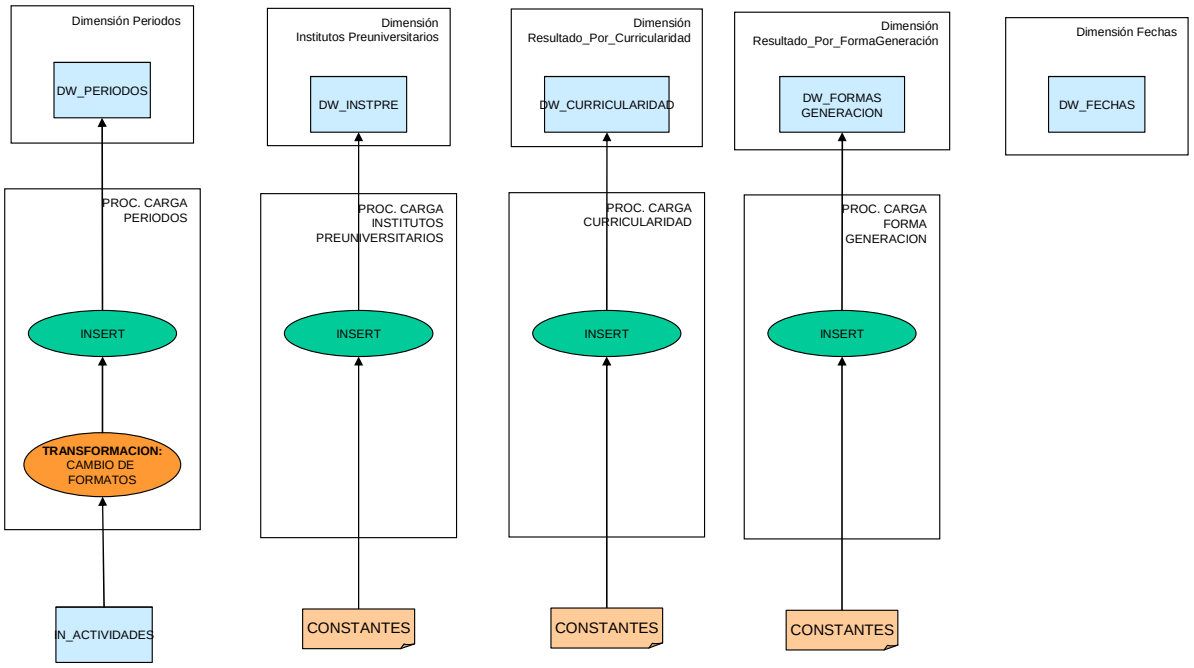

Figura 4.3 – Carga de tablas de dimensión (3)

A continuación se presenta el flujo de control asociado a la carga de las dimensiones de las figuras anteriores. Dicho flujo de control se representa mediante el pseudo código de los procesos de carga involucrados y debido a la similitud existente entre los mismos se incluyen sólo algunos ejemplos:

Dimensión Periodos

- Mientras existan actividades en IN_ACTIVIDADES
  - Obtener actividad
    - Realiza el cambio de formato de datos varios
    - Calcula el período para esa actividad
  - Ver si existe ese periodo en DW_PERIODOS
    - Si no existe se da de alta

Dimensión Estudiantes

- Elimina contenido de DW_ESTUDIANTES
- Mientras existan estudiantes en BD_ESTUDIANTES
  - Busca los datos de inscripciones a carrera de ese estudiante en BD_EST_CARR
    - Para cada carrera a la que se inscribió
      - Obtener los créditos en esta carrera de BD_ACTIVIDADES y BD_MAT_CARR
    - Se asigna al estudiante la carrera en la que al momento haya acumulado más créditos
    - Se inserta el estudiante en DW_ESTUDIANTES

Dimensión Tipo-actividad

- Elimina contenido de DW_TIPOACTIVIDAD
- Para cada tipo de actividad distinta en BD_ACTIVIDADES
  - Inserta la actividad en DW_TIPO_ACTIVIDAD

Dimensión Lugares

- Elimina contenido de DW_LUGARES, DW_PAISES, DW_CONTINENTES
- Para cada lugar de BD_LUGARES
  - Inserta la descripción del lugar dependiendo del código de lugar
  - Inserta en DW_LUGARES, DW_PAISES, DW_CONTINENTES dependiendo del código de lugar (selecciona por rango en duro)

4.2 Tabla de hechos del cubo Actuación

La figura 4.4 representa el flujo de datos del proceso de carga de la Tabla de hechos del cubo actuación, correspondiente al proceso ProcCargaResultado de la figura 3.4.

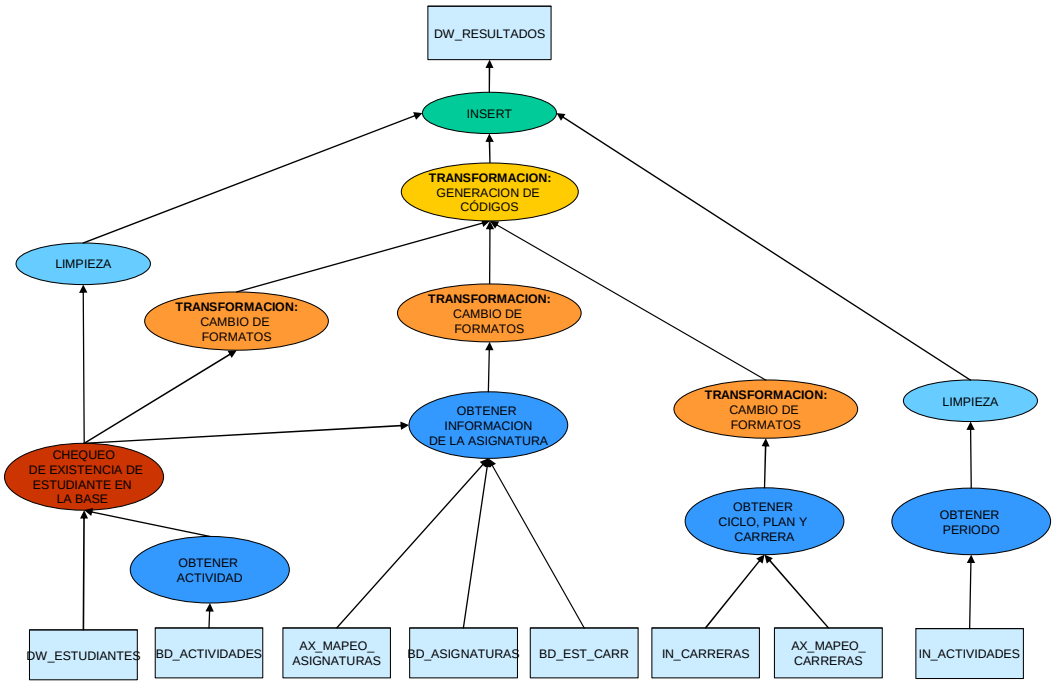

Figura 4.4 – Carga de la tabla DW_RESULTADOS

A continuación se representa mediante pseudo código el flujo de control asociado a la figura anterior.

- Mientras existan actividades
  - Obtener actividad
  - Si el estudiante de la actividad existe
    - Obtener mas información de la asignatura
    - Obtener ciclo, plan y carrera
    - Obtener información del período
      - Si no hay información del período
        - » Asignar valores por defecto
    - Realizar transformaciones de formatos
    - Generar códigos
    - Insertar actividad en la tabla DW_RESULTADOS
  - Si el estudiante no existe
    - Registrar error en log

4.3 Tabla de hechos del cubo Duración

La figura 4.4 representa el flujo de datos del proceso de carga de la Tabla de hechos del cubo duración estudiante, correspondiente al proceso ProcCargaDuracionEstudiante de la figura 3.5.

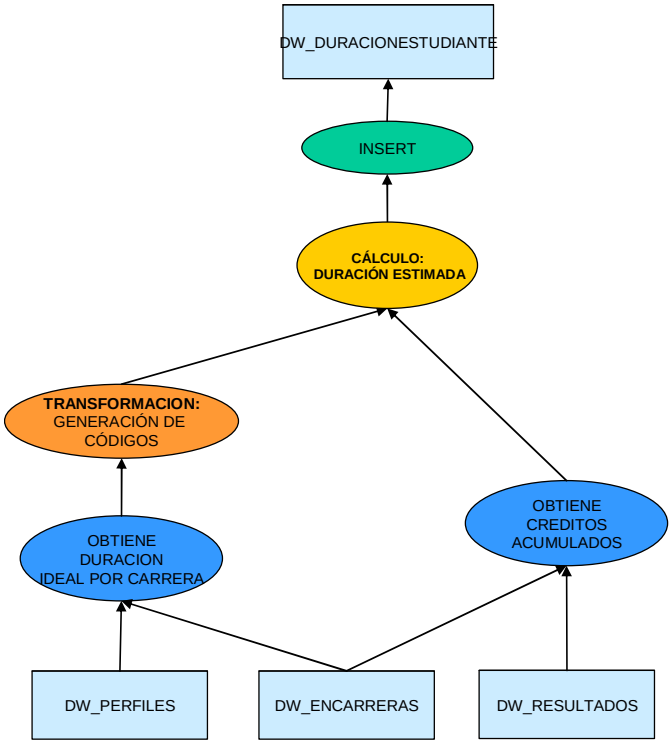

Figura 4.1 – Carga de la tabla DW_DURACIONESTUDIANTE

A continuación se representa mediante pseudo código el flujo de control asociado a la figura anterior.

- Mientras existan Estudiantes y la Carrera a la cual esta inscripto en DW_ENCARRERAS
  - Para cada Estudiante-Carrera
    - Obtiene la duración ideal para esa carrera a partir de DW_PERFILES
    - Calcula los créditos acumulados en el primer año de la carrera para dicho estudiante
    - Si no acumulo créditos
      - Calcula una estimación de los créditos acumulados esperados
      - Si no existe el Estudiante-Carrera en DW_DURACIONESTUDIANTE
        - » Inserta Estudiante-Carrera y sus créditos estimados en DW_DURACIONESTUDIANTE
    - Para cada año de la carrera del estudiante, hasta el año actual o el ultimo de carrera
      - Suma los créditos acumulados por asignaturas aprobadas asociadas al Estudiante para la Carrera en DW_RESULTADO
      - Si no acumulo créditos en el año

- Si tuvo actividades en el año registradas en DW_RESULTADO
  - Incrementa la ultima estimación utilizada
- Si no tuvo actividades en el año registradas en DW_RESULTADO
  - Resetea la estimación

- Si acumulo créditos en el año
  - Calcula los créditos hasta ese año sumando los nuevos créditos a los calculados de años previos
  - Recalcula la Estimación en función de los créditos calculados hasta ese año.
- Si no existe el Estudiante-Carrera en DW_DURACIONESTUDIANTE
  - Inserta Estudiante-Carrera y sus créditos estimados en DW_DURACIONESTUDIANTE

5 Proceso de carga a nivel bajo

En esta sección se presentan en un bajo nivel de detalle los procesos de carga ya identificados. Para cada figura de nivel medio estudiada se baja a un nivel operacional, representando el flujo de datos que muestra concretamente las operaciones realizadas sobre los datos obtenidos de las tablas origen hasta su carga en las tablas destino.

En este nivel los tipos de actividad se corresponden con operaciones relacionales tales como proyección, selección, join e insert. En el caso de las actividades de transformación y limpieza, la correspondencia es con un grupo de funciones y procedimientos, por lo cual se mantienen los nombres de actividad que los agrupan. En el caso de la actividad de transformación incluye una larga lista de funciones tales como cast, upper, substring, concatenación, así como operaciones aritméticas y booleanas; en cuanto a la limpieza, es muy similar al caso anterior, algunos ejemplos son in/not in, is NULL, ltrim/rtrim, cast, así como operaciones de comparación, aritméticas o booleanas.

5.1 Dimensiones

Las figuras 5.1 y 5.2 representan el flujo de datos del proceso de carga de las dimensiones, correspondiente a los procesos descritos en las figuras 4.1 y 4.2. Por razones de espacio se separan en dos figuras.

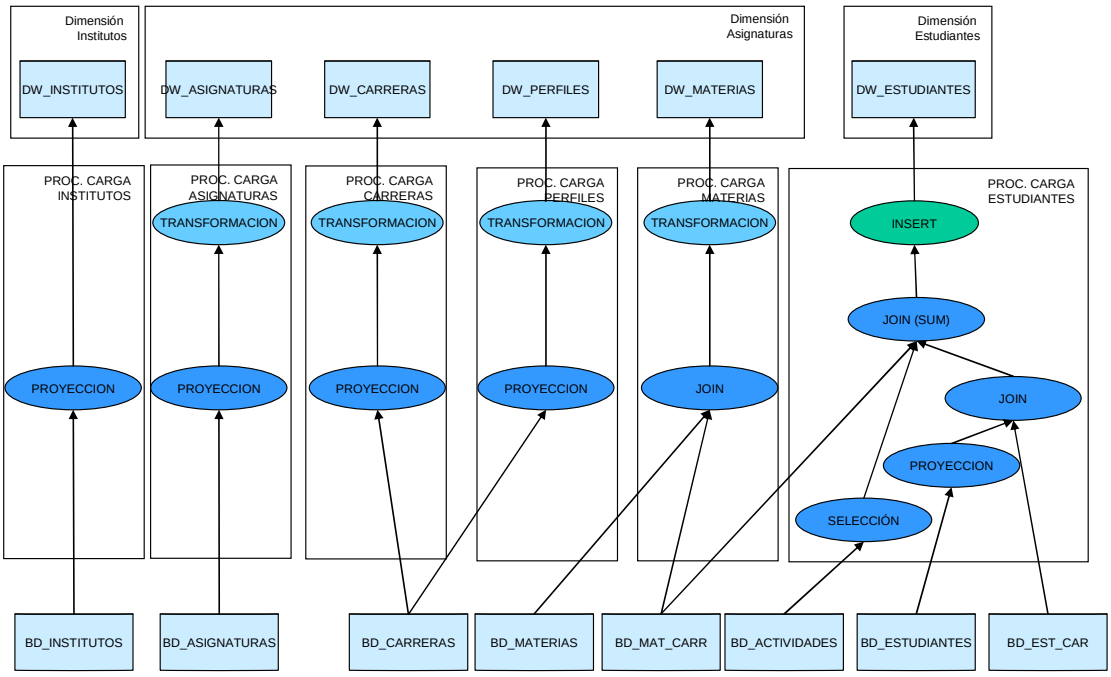

Figura 5.1 – Carga de tablas de dimensión (1)

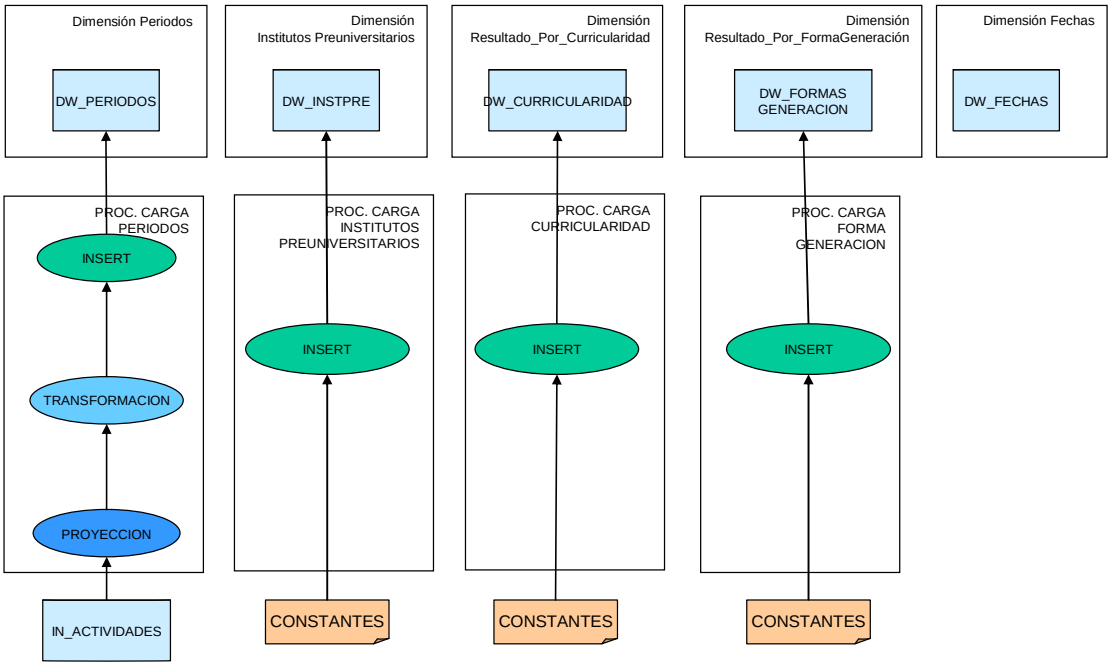

Figura 5.1 – Carga de tablas de dimensión (2)

5.2 Tabla de hechos del cubo Actuación

La figura 5.3 representa el flujo de datos del proceso de carga de la Tabla de hechos del cubo actuación, correspondiente al proceso de la figura 4.3.

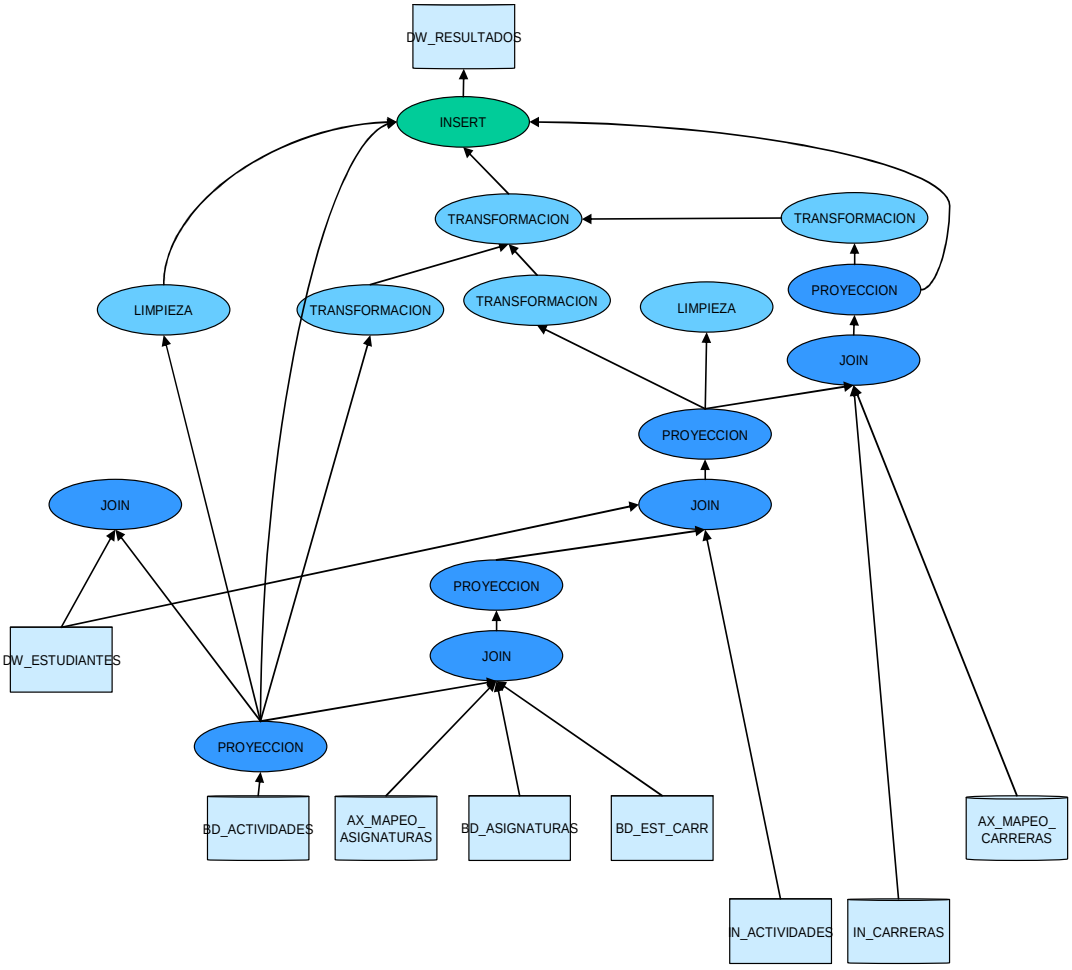

Figura 5.3 – Carga de la tabla DW_RESULTADOS

5.3 Tabla de hechos del cubo Duración

La figura 5.4 representa el flujo de datos del proceso de carga de la Tabla de hechos del cubo actuación, correspondiente al proceso de la figura 4.4.

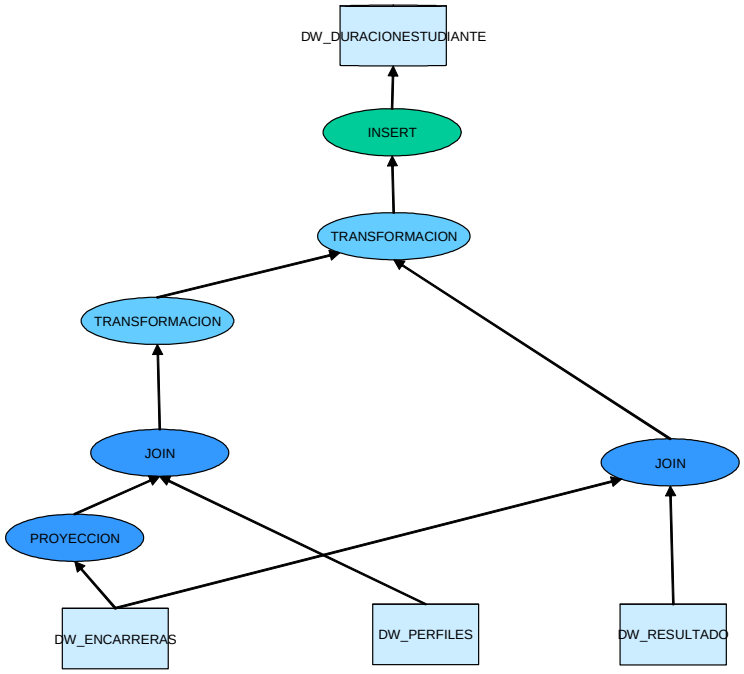

Figura 5.4 - Carga de la tabla DW_DURACIONESTUDIANTE

6 Conclusiones

El trabajo realizado permitió interiorizarse de las particularidades de un proceso de carga de Data Warehouse y específicamente en el correspondiente al Data Warehouse de Enseñanza de la Facultad de Ingeniería.
El proceso de carga de un Data Warehouse puede involucrar muchas fuentes de datos y/o transformaciones complejas a partir de los datos fuente, por lo tanto es importante planificar y documentar cuidadosamente los procesos de ETL, resultando a su vez muy conveniente el uso de una herramienta apropiada. En el caso particular, la falta de documentación y planificación inicial dificultó la tarea.
Se realizó un estudio top-down en tres niveles de granularidad, un nivel general de procedimientos de carga y su interdependencia: un nivel correspondiente a las actividades de carga realizadas desde un punto de vista conceptual y un tercer nivel donde se expresaron dichas actividades en términos de las operaciones realizadas sobre los datos. Este enfoque permitió lograr una mayor comprensión del proceso de carga y el flujo de datos asociado en los distintos niveles, facilitando la identificación de las actividades involucradas y la agrupación de operaciones típicas.
Este estudio es un primer paso hacia el objetivo general de analizar la calidad de la información en un sistema de Data Warehousing. Los resultados obtenidos, en

particular las actividades identificadas y el flujo de datos correspondiente, constituyen un caso de estudio para el análisis de la propagación de las propiedades de calidad en este ambiente.

Referencias

[1] CARPANI, Fernando. CMDM: A conceptual multidimensional model for Data Warehouse". Tesis de maestría. PEDECIBA Área Informática, Montevideo, 2000(RT 00-11) ISSN: 0797-6410

[2] ETCHEVERRY, Lorena; MARRERO, Pablo. Sistema de Data Warehouse de Enseñanza en la Facultad de Ingeniería. Marotta, Adriana (tutor). Trabajo de grado. Facultad de Ingeniería. Instituto de Computación, Montevideo 2003.

[3] INMON, W.H. Building the Data Warehouse 2nd Edition. New York: Wiley, 1996, 401 p. ISBN: 0471-14161-5

[4] TERCIA, Salvador; GATTO, Pablo. Propagación de valores de correctitud a través de operadores de álgebra relacional. Reporte interno. Facultad de Ingeniería. Instituto de Computación, Montevideo, 2005.