# PONTIFICIA UNIVERSIDAD CATOLICA DEL ECUADOR

## SEDE AMBATO

DEPARTAMENTO DE INVESTIGACIÓN, POSTGRADOS Y AUTOEVALUACIÓN

Tema:

"DESARROLLO DE UN SISTEMA DE APOYO A LA TOMA DE DECISIONES QUE GESTIONE LA INFORMACIÓN DE LA PRUEBA DE APTITUD ACADÉMICA DE LA UNIVERSIDAD TÉCNICA DE AMBATO"

Tesis de grado previo a la obtención del título de Magister en GERENCIA INFORMÁTICA CON MENCIÓN EN DESARROLLO DE SOFTWARE Y REDES

Autor:

LUIS VICENTE SÁNCHEZ ALVAREZ

Asesor:

ING., Msc. JANIO JADÁN



Ambato - Ecuador

Julio, 2008

# PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR
# SEDE AMBATO

## DEPARTAMENTO DE INVESTIGACIÓN, POSTGRADO Y AUTOEVALUACIÓN

## Hoja de Aprobación

**Tema:**

**"DESARROLLO DE UN SISTEMA DE APOYO A LA TOMA DE DECISIONES QUE GESTIONE LA INFORMACIÓN DE LA PRUEBA DE APTITUD ACADÉMICA DE LA UNIVERSIDAD TÉCNICA DE AMBATO"**

**Autor:**

**LUIS VICENTE SÁNCHEZ ALVAREZ**

Janio Jadán Ing. Msc. ..... f. .....
DIRECTOR DE TESIS

Galo López Ing. Msc. ..... f. .....
CALIFICADOR

Washington Medina Ing. Msc. ..... f. .....
CALIFICADOR

Telmo Viteri Ing. Msc. ..... f. .....
DIRECTOR DE LA UNIDAD ACADEMICA

Pablo Poveda Ab. ..... f. .....
SECRETARIO GENERAL DE LA PUCESA

# DECLARACIÓN DE AUTENTICIDAD Y RESPONSABILIDAD

Yo, Luis Vicente Sánchez Alvarez portador de la cédula de ciudadanía Nº-180265778-1 declaro que los resultados obtenidos en la investigación que presento como informe final, previo la obtención del título de MAGISTER EN GERENCIA INFORMÁTICA CON MENCIÓN EN DESARROLLO DE SOFTWARE Y REDES son absolutamente originales, auténticos y personales.

En tal virtud, declaro que el contenido, las conclusiones y los efectos legales y académicos que se desprenden del trabajo propuesto de investigación y luego de la redacción de este documento son y serán de mi sola y exclusiva responsabilidad legal y académica.

Luis Vicente Sánchez Alvarez

CI. 180265778-1

## AGRADECIMIENTO

A la Pontificia Universidad Católica del Ecuador Sede Ambato y en ella a todo el personal Administrativo y Docente, a todas aquellas personas que compartieron sus conocimientos conmigo. De manera especial al Ing. Msc. Janio Jadán, por su apoyo, asesoría, ideas y sobre todo paciencia.

## DEDICATORIA

A quienes siempre llevo en mi mente y corazón.

# RESUMEN

La evolución y globalización de los mercados, así como la alta competencia entre las diferentes organizaciones obliga a éstas a contar con tecnologías de información para proporcionar a directivos, ejecutivos y analistas clara y oportuna información; sencilla de comprender y manipular, relativa a la organización en general o algún área de ella. Es por ello que se necesita contar con un modelo de infraestructura para el soporte de toma de decisiones, como el Data Warehouse. A través de ésta se puede convertir la información que posee una organización en conocimiento fundamental para una mejor toma de decisiones. La tesis propuesta tiene como finalidad el desarrollo de un prototipo de un sistema de apoyo a la toma de decisiones que gestione la información de la Prueba de Aptitud Académica de la Universidad Técnica de Ambato, y ha sido dividida en cuatro capítulos que incluyen definiciones de temas relacionados como el Data Warehouse, Data Mart, sus características, metodologías de desarrollo y técnicas de modelado. En el penúltimo capítulo, se desarrolla un prototipo motivo de nuestro estudio. Para ello bajo los requerimientos establecidos se analizan las fuentes de información, se realiza el modelo multidimensional del Data Mart, se desarrolla el proceso de extracción, transformación y carga de la información, a continuación se construye un cubo de información y se finaliza con la presentación de informes. Para todas estas tareas se usan herramientas de software proporcionadas por Microsoft como son SSIS, SSAS y Excel.

# ABSTRACT

Market's evolution and globalization, fierce competences among organization has been the most important reason to improve information technologies in order to have managers, executives and analyzer with clear and fast information, easy to understand and handle, about the organization general status or in a specific area. That's why we have to get a strong infrastructure model to take decision, such as Data Warehouse with this; we are able to transform information into basic knowledge in order to make decisions. This thesis wants to develop a support system prototype to make decision making easier to handle the UTA's academic exam. It has been divided in four chapters which include Data Warehouse, Data Mart definitions, characteristics, develop methodologies and modeling techniques. Before the last chapter, there's a prototype for us to study. Under the established requirements we analyze the information sources, we make a multidimensional Data Mart model, and we develop the extraction transformation and information load process. Later, we build an information square and we end by presenting informs. To do all this tasks, we use Microsoft Software such as: SSIS, SSAS and Excel.

# TABLA DE CONTENIDOS

# FIGURAS

# TABLAS

# CAPÍTULO I

## 1. PROYECTO DE INVESTIGACIÓN

### 1.1. ANTECEDENTES

Estar bien informados a veces resulta una meta difícil de alcanzar, sobre todo cuando las organizaciones crecen y se desarrollan al ritmo de los tiempos actuales. Los desafíos de los distintos sectores económicos tienen en general como puntos comunes clientes cada vez más exigentes, cambios cada vez más rápidos y una competencia día a día más fuerte.

Para hacer frente a estos desafíos e ir más allá de la reactividad, es necesario anticipar. Anticipar los cambios, anticipar las nuevas necesidades de los usuarios y anticipar respecto a la competencia. Para que esta anticipación sea la más adecuada hay que disponer de informaciones pertinentes. Todas las organizaciones disponen de datos que provienen de sus sistemas internos o bien del exterior. El problema de estas organizaciones es alcanzar los objetivos definidos por los desafíos de su sector sacando provecho de los datos accesibles.

La organización actual se encuentra sumergida bajo gran cantidad de datos. Esta sobreabundancia tiene como consecuencia directa un rechazo por saturación. Sin embargo, los datos representan una mina de informaciones. Son una ventaja de la que se debe sacar partido. Para ello, resulta fundamental implementar una nueva informática de decisión para obtener una mejor comprensión del valor de las

informaciones disponibles, definir indicadores de negocio para facilitar la toma de decisiones operativas y conservar la memoria de la organización.

No ajena a la situación antes mencionada la Universidad Técnica de Ambato que posee varios sistemas de información (entre ellos el de la Prueba de Aptitud Académica), requiere del nuevo papel de la informática que es definir e integrar una arquitectura que sirva de fundación a las aplicaciones de ayuda a la decisión. Esta arquitectura global se ve reflejada en el denominado Data Warehouse.

## 1.2. PLANTEAMIENTO DEL PROBLEMA

El Data Warehouse ha aparecido estos últimos años tras la convergencia entre las nuevas necesidades de informaciones en las empresas y la capacidad de integrar e implementar tecnologías aptas para responder a ello. Resulta el centro de atención de las grandes instituciones, porque provee un ambiente para que hagan un mejor uso de la información que está siendo administrada por diversas aplicaciones operacionales.

Cabe destacar que pese a la importancia que resulta para toda organización, disponer de un sistema que ayude al soporte de toma de decisiones, en la actualidad la Universidad Técnica de Ambato, un buen punto de partida constituye la gestión de la información de la Prueba de Aptitud Académica.

## 1.3. PROBLEMATIZACIÓN

- Inexistencia de un Sistema de apoyo a la toma de decisiones que gestione la información de la prueba de aptitud académica.
- Escasa comprensión de los fundamentos del Data Warehouse como

herramienta de apoyo a la toma de decisiones.

- Desconocimiento de metodologías que faciliten el desarrollo de sistemas de apoyo a la toma de decisiones.
- Falta de reportes que permitan conocer aspectos puntuales como por ejemplo: el número de postulantes inscritos para la Prueba de Aptitud Académica de la Universidad Técnica de Ambato según sus colegios, género, procedencia, etc.

## 1.4. DELIMITACIÓN

El presente proyecto de tesis se ejecutará en su totalidad en la Universidad Técnica de Ambato en la Dirección de Sistemas Informáticos y Redes de Comunicación (DISIR). Se apoyará en toda su plataforma tecnológica (especialmente de sus sistemas de información); está previsto realizarlo en un período de tiempo comprendido entre: octubre de 2007 a julio de 2008.

Va a constar de los siguientes aspectos: estudio del Data Warehouse como herramienta de apoyo a la toma de decisiones, que permita obtener los conocimientos necesarios para desarrollar una solución de calidad; conocer varias metodologías para el desarrollo de este tipo de sistemas; desarrollar el prototipo de un sistema de apoyo a la toma de decisiones que gestione la información de la Prueba de Aptitud Académica de la Universidad Técnica de Ambato.

Este proyecto es solo una parte de uno más ambicioso: un Data Warehouse General para la Universidad Técnica de Ambato que deberá incluir otras áreas operativas como: Departamento Financiero, Dirección de Recursos Humanos, entre otras.

## 1.5. MARCO TEÓRICO

Uno de los principales objetivos y a la vez, uno de los principales problemas en las empresas de cualquier tipo ha sido desde siempre el de mantener organizado y actualizado (al día, al minuto y quizá al segundo) el conjunto de datos que le sirve como fuente de información imprescindible para llevar a cabo sus negocios de forma satisfactoria.

La importancia del problema era, y es, tan grande que se ha constituido en tema de estudio teórico bajo la denominación de sistemas de información. El propósito de un sistema de información es producir un modelo lo más exacto y eficiente posible de los recursos significativos para una empresa.

Si se analiza la situación actual de muchos sistemas de información nos encontramos con una proliferación de ficheros (definidos como un conjunto de instrucciones o datos coherentes almacenados en un soporte magnético) específicos cada uno de ellos de una determinada aplicación o conjunto de programas que cumplen una tarea específica. Los datos se recogen varias veces y se encuentran repetidos. Esta redundancia, además de malgastar recursos, origina a menudo divergencias en los resultados.

Para solucionar estos inconvenientes la informática ha hecho uso de las bases de datos. En lo esencial, una base de datos no es otra cosa que una colección de información que existe durante un período largo, a menudo de muchos años. Con la expresión base de datos se designa una colección de datos que es administrada por un sistema de administración de bases de datos, que se abrevia DBMS (Data Base

Management System, sistema de administración de base de datos o simplemente sistema de bases de datos). Se espera que este sistema cumpla las siguientes actividades:

1. Permita a los usuarios crear otras bases de datos y especificar su **esquema** (estructura lógica de los datos) por medio de un lenguaje especializado denominado lenguaje de definición de datos.
2. Ofrezca a los usuarios la capacidad de **consultar** los datos (una consulta es un tecnicismo de base de datos que formula una pregunta sobre los datos) y modificarlos, usando para ello un lenguaje apropiado, llamado a menudo lenguaje de consulta o lenguaje de manipulación de datos.
3. Soporte el almacenamiento de cantidades muy voluminosas de datos durante un largo período, protegiéndolos contra accidentes o utilización no autorizada y permitiendo el acceso eficiente para hacer consultas y modificar la base de datos.
4. Controle el acceso simultáneo a los datos por parte de muchos usuarios, sin permitir que las acciones de uno de ellos afecte a los otros ni que los accesos simultáneos corrompan los datos accidentalmente.

Sin embargo, dichas bases de datos no trabajan necesariamente por separado, se ven asociadas muchas veces con las denominadas arquitecturas cliente/servidor, que se trata de un sistema de organización de la información en la cual la aplicación central o servidor almacena los ficheros y los pone a disposición de las aplicaciones clientes.

Por otro lado, las empresas modernas viven de datos. Las bases de datos de las empresas con frecuencia ocupan cientos de gigabytes (unidad de medida de la información, un gigabyte equivale a 109 bytes). Para manejar base de datos tan enormes, se requieren varios avances tecnológicos.

Las aplicaciones cliente/servidor centradas en bases de datos se dividen en dos categorías: sistemas de apoyo a la toma de decisiones (DSS: Decision Support Systems) y procesamiento de transacciones en línea (OLTP: On Line Transaction Processing). Estas dos categorías de cliente/servidor ofrecen soluciones de negocios absolutamente distintas.

Los sistemas de OLTP sirven para crear aplicaciones en todo género de actividad. Entre ellas pueden citarse los sistemas de reservaciones, punto de venta, sistemas de rastreo, control de inventarios y sistemas de control para fábricas manufactureras. Se trata por lo general de aplicaciones decisivas para el cumplimiento de objetivos que en el 100% de los casos requieren un tiempo de respuesta muy rápido. Las aplicaciones de OLTP también requieren rigurosos controles sobre seguridad e integridad de la base de datos. La confiabilidad y la disponibilidad del sistema general deben ser del más alto nivel. Los datos deben conservarse en forma consistente y correcta.

Los DSS sirven para analizar datos y generar reportes. Les ofrecen a los profesionales de negocios y buscadores de información los medios para obtener justamente la información que necesitan. Para cumplir exitosamente con sus funciones, un sistema de apoyo a la toma de decisiones debe brindarle al usuario un

acceso flexible a datos y las herramientas para manipularlos y presentarlos en todo tipo de formatos de reportes. El usuario debe estar en condiciones de elaborar consultas complejas, responder preguntas del género "¿Qué pasaría si?", buscar correlaciones entre los datos, convertirlos en gráficas y trasladarlos a otras aplicaciones, como hojas de cálculo y documentos de procesamiento de textos.

Habitualmente, los sistemas de apoyo a la toma de decisiones no son críticos en lo que respecta al tiempo, de modo que pueden tolerar tiempos de respuesta más lentos. Sus controles de integridad son deficientes, y sus capacidades de acceso a tablas múltiples son limitadas. Otro concepto que entra en juego son los sistemas de información ejecutiva (EIS: Executive Information System), son incluso más potentes, fáciles de usar y específicos que las herramientas DSS. También son más caros, lo que explica la presencia de la palabra "ejecutiva" en su nombre. En todo caso las distinciones entre EIS y DSS son cada vez menos claras.

Estas diferencias deben conocerse para poder apreciar lo que ofrece el Data Warehouse. Bill Inmon y proveedores como Teradata define un Data Warehouse como una base de datos independiente de apoyo para la toma de decisiones, que por lo general contiene grandes cantidades de información. Richard Hackathorn define un Data Warehouse como una colección de objetos que han sido inventariados para su distribución entre una comunidad de negocios. El Data Warehouse por tratarse de una estructura bastante amplia, se ayuda de otras herramientas como es el caso de OLAP. Las herramientas de procesamiento analítico en línea (OLAP: On Line Analytical Processing) permiten formular consultas más sofisticadas, y después visualizar los resultados correspondientes. Otro elemento relacionado con el Data

Warehouse lo constituye el Data Mining. Se emplea a menudo este término para designar el conjunto de herramientas que permiten al usuario acceder a los datos de la empresa y analizarlos.

## 1.6. HIPÓTESIS

El desarrollo de un sistema de apoyo a la de toma de decisiones, mejorará la gestión de la información de la Prueba de Aptitud Académica de la Universidad Técnica de Ambato.

## 1.7. OBJETIVOS

### 1.7.1. OBJETIVO GENERAL

- Desarrollar un sistema que permita apoyar los procesos de toma de decisiones asociadas con la gestión de la información de la Prueba de Aptitud Académica de la Universidad Técnica de Ambato.

### 1.7.2. OBJETIVOS ESPECÍFICOS

- Comprender los fundamentos del Data Warehouse como herramienta que apoye la toma de decisiones.
- Investigar varias técnicas de modelado del Data Warehousing.
- Generar informes que apoyen la toma de decisiones asociadas a la Prueba de Aptitud Académica de la Universidad Técnica de Ambato.

## 1.8. METODOLOGÍA DE TRABAJO

Se ha definido como tipo de investigación: la Investigación Bibliográfica o Documental, la cual se realizará tomando información de fuentes indirectas como: libros, revistas, folletos, manuales y documentos relacionados con el Data Warehouse; empleadas durante el proceso de recolección, procesamiento y análisis de la información.

Cabe destacar que se hará uso del Método Científico y para la recolección de datos, se apoyará en información secundaria disponible a partir de documentos y material impreso como libros, manuales, revistas, informes, documentos oficiales, etc., empleando para esto como técnica bibliográfica la Lectura Científica. También se utilizará información encontrada en Internet.

## 1.9. JUSTIFICACIÓN

En la actualidad la Universidad Técnica de Ambato posee grandes volúmenes de datos, que no han sido explotados adecuadamente. La dificultad de acceder, navegar y analizarlos constituye un problema.

El Data Warehouse como herramienta que apoya los procesos de toma de decisiones, provee una estructura que permite resolver el problema de manipular gran cantidad de datos. Obteniendo estos datos de diferentes fuentes, mejorando la recuperación y presentación de los datos, de tal forma que pueda responder a preguntas estratégicas fundamentales y esenciales en la gestión de la Prueba de Aptitud Académica.

Con esta nueva herramienta de apoyo a la toma de decisiones, los niveles directivos de la Universidad contarán con información oportuna y fácil de entender, referentes al área de aplicación tratada.

Por todo lo expuesto anteriormente, resulta indispensable la ejecución de este proyecto, mismo que servirá de base para otros relacionados con la toma de decisiones en otras áreas o entidades de la Universidad Técnica de Ambato.

# CAPÍTULO II

# 2. FUNDAMENTO TEÓRICO

## 2.1. Sistemas de Apoyo a la Toma de Decisiones (DSS)

El concepto y los límites de los DSS (Decision Support Systems) no han sido completamente determinados, pese a que su utilidad ha sido justificada en las organizaciones. En varias ocasiones llega a confundirse con términos como Data Warehouse y OLAP (On Line Analytical Processing, Procesamiento Analítico en Línea); lo cierto es que, independientemente del término que se utilice, todos estos conceptos tienen en la toma de decisiones el punto de encuentro.

DSS se refiere a cualquier sistema de software que permite el análisis de las diferentes variables del negocio para apoyar una decisión.

Son "Sistemas de información a nivel administrativo de la organización, que combinan datos y modelos analíticos sofisticados o herramientas de análisis de datos para apoyar la toma de decisiones semiestructurada y no estructurada" **[1]**.

Decisiones no estructuradas hacen referencia a todas aquellas fuera de la rutina en las que el tomador de decisiones debe dar criterio, evaluación y visión a la definición del problema; no hay un procedimiento de acuerdo para tomar tales decisiones. En cambio las decisiones estructuradas son repetitivas, rutinarias e implican un procedimiento definido para manejarlas **[1]**.

Igualmente a un DSS puede considerarse como un sistema que se basa en un Data Warehouse y crea una base de datos multidimensional, permitiéndole al usuario procesar analíticamente la información en línea (OLAP).

## 2.2. La necesidad de desarrollar un Data Warehouse (DW)

Para responder a la inquietud del por qué desarrollar un DW, se debe considerar como factores de influencia a la información y las organizaciones en general. De ahí que la importancia del DW se refleja en los siguientes aspectos:

- Mejorar la entrega de la información: información accesible, completa, consistente, correcta y oportuna. Información que los miembros de la organización necesitan, en el formato adecuado y en el tiempo requerido.
- Mejorar el proceso de toma de decisiones: con un mayor soporte de información se obtienen decisiones más rápidas, de igual manera, quienes son responsables de las decisiones de negocio adquieren mayor confianza en decisiones propias y de otros miembros.
- Se logra un impacto positivo sobre los procesos organizacionales: cuando al personal se le da acceso a una mejor calidad de la información, se dice que la organización puede lograr por si sola eliminar retardos en los procesos de la organización, mediante el uso compartido de las fuentes de información se puede integrar procesos y eliminar el procesamiento de datos innecesarios.

## 2.3. Objetivos del Data Warehouse

El objetivo principal del DW consiste en reunir y consolidar las bases de datos diferentes, que se mantienen en diferentes departamentos o áreas funcionales de la

organización como subsistemas de información independientes en una gran base de datos, recogiendo datos muy dispares y, muchas veces infrautilizados, procedentes de fuentes internas repartidas por toda la organización. También recogerá datos o información externa, que normalmente se recibe sobre las diferentes entidades u objetos de información, como pueden ser: clientes, proveedores, productos y servicios, canales, estructura organizativa, mercado, competencia, etc. En otras palabras los derivados de las relaciones de la organización con su entorno **[2]**.

"El principal rol del DW es servir como un repositorio de datos que almacena datos de diferentes orígenes, haciéndolos a estos accesibles a otros conjuntos de almacenes de datos – los Data Marts" **[3]**.

"La misión del DW es publicar la información de la organización, haciendo más efectivo el soporte a la toma de decisiones" **[4]**.

Por otra parte, según Dos Santos Romina en **[5]** se dice que el DW tiene por objetivos los siguientes aspectos:

- La integración de bases de datos heterogéneas (relacionales, documentales, geográficas, archivos, etc.).
- La ejecución de consultas complejas no predefinidas visualizando el resultado en forma gráfica y en distintos niveles de agrupamiento y totalización de datos.
- Agrupamiento y desagrupamiento de datos en forma interactiva.
- Análisis de problemas en términos de dimensiones. Por ejemplo, permite analizar datos históricos a través de una dimensión de tiempo.

- Control de calidad de datos para asegurar, no solo la consistencia de la base, sino también la relevancia de los datos en base a los cuales se toman las decisiones.

## 2.4. Beneficios y áreas de aplicación

Normalmente en el orden económico los beneficios que podemos obtener de un DW no tienen la inmediatez de los que pueden obtenerse mediante un eficiente sistema de información operacional, por lo general mediante el DW hemos de esperar el ahorro de gastos motivados por los cambios que puedan sugerirse en la gestión de nuestra organización a mediano y largo plazo.

Sin embargo, bajo otros puntos de vista podemos considerar como beneficios del DW a los siguientes:

- Menor costo en la toma de decisiones, ya que se suprime el despilfarro de tiempo que se podía producir al intentar ejecutar consultas de datos complejas y largas con bases de datos que estaban diseñadas específicamente para transacciones más cortas y sencillas.
- Mayor flexibilidad ante el entorno, debido a que el DW convierte los datos operacionales en información relacionada y estructurada, que genera el "conocimiento" necesario para la toma de decisiones.
- Mejor servicio al cliente, como efecto de una importante mejora en la calidad de gestión por parte de la organización. Esto constituye como pilar fundamental sobre el que descansa las organizaciones.

- Rediseño de procesos, tras ofrecer a los usuarios una capacidad de análisis de la información de su negocio que tiende a ser limitada y permite con frecuencia obtener una visión más profunda y clara de los procesos de negocio propiamente dichos, lo que a su vez permite obtener ideas renovadoras para el rediseño de los mismos.
- Se consigue una herramienta estratégica y táctica que permite obtener una ventaja competitiva.
- Se logra la habilidad para explorar y analizar datos con el fin de revelar la existencia de tendencias dentro de un negocio.
- Generar reportes globales o por áreas o secciones específicas.
- Crear bases de datos adicionales que se relacionen con clientes.
- Crear escenarios con relación a una decisión.
- Hacer pronósticos en temas puntuales como pueden ser las ventas, devoluciones, promociones, admisiones, deserciones, etc.
- Compartir informaciones entre departamentos.
- Realizar análisis multidimensionales.
- Generar y procesar datos adicionales para la toma de decisiones organizacionales.

En la actualidad el DW ha logrado posicionarse en varios sectores, siendo los más conocidos los siguientes:

- En lo referente a Ventas: se logra el análisis de ventas, detección de clientes importantes, análisis de productos, líneas, mercados, pronósticos, proyecciones, patrones de compra y hábitos del consumidor.

- En el Marketing: segmentación y análisis de clientes, seguimiento a nuevos productos y servicios.
- En el sector Bancario y Financiero: análisis de gastos, rotación de cartera, razones financieras, administración de relaciones, administración de riesgos de crédito.
- En la Manufactura: productividad según líneas, análisis de calidad, análisis de desperdicios, rotación de inventarios y partes críticas, materias primas, cumplimiento de pedidos y embarques, integración de proveedores y logística.
- En Recursos Humanos: permite comprender, ajustar y maximizar el rendimiento de los empleados ya sea basado en su ubicación geográfica o ubicación de trabajo, habilidades y otras consideraciones como de género y edad.
- En el área de la Educación (puede ser la educación Universitaria por ejemplo): se puede hacer un seguimiento de las carreras con mayor población estudiantil, las carreras en peligro de desaparecer, la distribución de los estudiantes según su lugar de procedencia, género, edad. Las calificaciones de los aspirantes a ingresar en las diferentes carreras universitarias, notas mínimas, máximas, promedios, la cantidad de matriculados, aprobados, reprobados, deserciones, etc.

Igualmente se puede indicar que otros sectores como: el E-Business, Comunicación, Cuidado de la Salud y Seguridad Social están siendo apoyados por los beneficios que otorga el DW.

## 2.5. Diferencias entre las bases de datos operacionales y las del Data Warehouse

Un DW es diferente de las bases de datos operacionales que soportan las aplicaciones de un Procesamiento de Transacciones en Línea[1] (OLTP, On-Line Transaction Processing) **[6]**.

- Un DW está orientado fundamentalmente al almacenamiento de información para su posterior utilización mediante el uso de herramientas adecuadas. Una base de datos tradicional debe estar diseñada para un manejo óptimo de la información, tomando en consideración aspectos como rapidez y seguridad.
- En un DW no se da mucha importancia a un diseño óptimo de la base de datos según normas clásicas (como es el caso de la normalización).
- Todo DW puede recopilar información a partir de un indeterminado número y tipo de fuentes.

La siguiente tabla reúne las diferencias fundamentales entre una base de datos operacional y un DW:

| BASES DE DATOS OPERACIONALES | BASE DE DATOS DEL DW |
|---|---|
| Sirven para las operaciones diarias. | Usada para análisis, toma de decisiones, búsqueda de patrones y tendencias. |
| Datos detallados y manejados en línea. | Datos categorizados, dimensionados y jerarquizados. |
| Utilizadas por usuarios administrativos y operadores de datos. | Utilizada por analistas y estadísticos. |

[1] OLTP se usa para definir cualquier sistema de software que reúne datos usando las transacciones (en el momento en que ocurren).

Continuación…

| | |
|---|---|
| Manejan alto grado de normalización. | Manejan un menor control de redundancias. |
| Contienen información detallada. | Contiene información resumida. |
| Los requerimientos son usualmente conocidos antes del diseño del sistema. | Los requerimientos no son totalmente comprendidos al inicio del diseño del sistema. |
| Admiten el acceso simultáneo de muchos usuarios que agregan y modifican datos. | Admiten el acceso simultáneo de muchos usuarios que consultan datos. |
| Contienen grandes cantidades de datos, incluidos datos extensivos utilizados para comprobar transacciones. | Contienen grandes cantidades de datos, sumarizados, consolidados y transformados. También de detalle pero solo los necesarios para el análisis. |
| Tienen estructuras de bases de datos complejas. | Tienen estructuras de bases de datos simples. |
| Se ajustan para dar respuesta a la actividad transaccional. | Se ajustan para dar respuesta a la actividad de consultas. |
| Las consultas analíticas que resumen grandes volúmenes de datos afectan negativamente a la capacidad del sistema para responder a las transacciones en línea. | Organizan los datos en estructuras simplificadas buscando la eficiencia de las consultas analíticas más que del proceso de transacciones. |
| Los datos que se modifican con frecuencia interfieren en la coherencia de la información analítica. | Proporcionan datos estables que representan el historial de la empresa u organización. Se actualizan periódicamente con datos adicionales, no con transacciones frecuentes. |
| La seguridad se complica cuando se combina el análisis en línea con el proceso de transacciones en línea. | Simplifica los requisitos de seguridad. |

**Tabla 2.1:** Diferencias entre bases de datos operacionales y base de datos del DW

## 2.6. ¿Qué es el Data Warehouse?

Hasta el momento ya ha sido utilizado el término DW, es por ello que conviene indicar su definición bajo varios puntos de vista.

Según el considerado "padre" del DW William Inmon: "Data Warehouse es una colección de datos orientada a temas específicos, integrado, variante en el tiempo, no volátil y que soporta la administración del proceso de toma de decisiones" **[7]**.

Ralph Kimball y Joe Caserta, consideran que el "Data Warehouse es un sistema que extrae, limpia, ajusta y entrega fuentes de datos dentro de un depósito dimensional y luego apoya e implementa consultas y análisis con el propósito de la toma de decisiones" **[4]**.

"Un Data Warehouse es una base de datos que contiene datos de múltiples sistemas operacionales que han sido consolidados, integrados, agregados y estructurados para que puedan ser usados para apoyar el análisis y proceso de toma de decisiones de un negocio" **[8]**.

"Un Data Warehouse es una base de datos especializada que reúne los datos de una variedad de bases de datos existentes para apoyar la gestión de las necesidades de información" **[9]**.

"Un Data Warehouse es una base de datos específicamente estructurada para la consulta y el análisis. Típicamente contiene datos del negocio que representan la historia de la organización. Los datos son por lo general menos detallados y de mayor vida de los datos de un procesamiento de transacciones en línea (OLTP)" **[10]**.

También se puede definir al DW:

- Como una arquitectura: Data Warehousing, que es una arquitectura basada en el entorno cliente/servidor.
- Como un depósito de datos: representa el proceso de reunir información histórica de una organización en un depósito central.
- Como una herramienta: herramientas orientadas a ofrecer nuevas visiones de la información almacenada en una base de datos.
- Como un continuo proceso de mezcla: almacenamiento y presentación de información consolidada a partir de fuentes no homogéneas.

## 2.7. Características

Basado en la definición de William Inmon **[7]**, se consideran como características del DW a las siguientes:

### 2.7.1. Orientado al tema

Esta característica indica que el DW se organiza alrededor de los temas principales de la empresa. De esta manera los datos se estructuran o clasifican por temas o en base a aspectos que son de mayor interés, contrariamente a los datos que son tomados como producto de procesos orientados a aplicaciones operacionales.

Un DW se diseña para consultar eficientemente información relativa a las actividades básicas de la organización como producción, ventas y compras, y no para soportar los procesos que se realizan en ella, como gestión de pedidos y facturación.

### 2.7.2. Integrado

De todos los aspectos, este es el más importante. La información que se encuentra al interior de un DW está siempre integrada.

En su sentido más amplio el DW es un proyecto de empresa. Esto significa que de alguna manera todos los datos provenientes de las aplicaciones operativas departamentales, generan información relacionada. Así por ejemplo la consolidación de todas las informaciones respecto de un cliente dado es necesaria para dar una vista homogénea de dicho cliente a los analistas. Antes de integrarse en el DW, los datos deben formatearse y unificarse para llegara a un estado coherente, es decir, un dato debe poseer una codificación y descripción única.

Las diferencias únicamente dependerían del punto de vista de los usuarios, de los programadores o en sí del uso que se le esté dando a la información. La integración de datos se aprecia de varias maneras: en convenciones de nombres, en la medida uniforme de variables, en la codificación de estructuras consistentes, en atributos físicos de los datos, etc.

Como ejemplo real se puede indicar que existen desarrolladores de aplicaciones que codifican el campo GENERO en varias formas. En una Aplicación A: GENERO puede ser representado con "F" para femenino y con "M" para masculino. En una Aplicación B: "1" para femenino y "0" para masculino. En una Aplicación C: GENERO se encuentra representado por las palabras "Femenino" y "Masculino".

No importa como GENERO llegue al DW. Pueda que "F" y "M" sea la mejor opción. Lo importante es que sea de cualquier fuente o aplicación de donde venga, el GENERO debe llegar al DW en un estado integrado uniforme.

### 2.7.3. Variante en el tiempo

Los datos en el DW son precisos para un cierto momento, no necesariamente ahora; por eso se dice que los datos en el DW son variantes en el tiempo. La varianza en el tiempo de los datos se manifiesta de muchas maneras, debido a que el DW contiene datos en un largo plazo.

El DW contiene un lugar para almacenar datos con una antigüedad que puede ser desde 5 a 10 años, o incluso más antiguos, para poder ser usados en comparaciones, tendencias y previsiones. Estos datos no se modificarán.

Toda estructura clave en un DW contiene implícita o explícitamente un elemento relacionado con el tiempo. Esto necesariamente no sucede en el ambiente operacional.

### 2.7.4. No volátil

La información es útil sólo cuando es estable. Los datos operacionales cambian sobre una base momento a momento. La perspectiva más grande, esencial para el análisis y la toma de decisiones, requiere de una base de datos estable.

Los procesos de inserción, actualización y borrado se realizan regularmente en un ambiente operacional sobre una base de registro a registro. En este punto resulta preciso indicar que la manipulación básica de los datos que ocurre en el DW en mucho más simple.

Solo ocurren dos únicas operaciones: la carga inicial y el acceso a los datos. No hay necesidad de actualizaciones (en su sentido más general). Los datos no serán cambiados o modificados de ninguna manera una vez que ellos han sido introducidos en el DW, solamente podrán ser cargados, leídos y/o accedidos.

## 2.8. Estructura

El DW según un eje histórico y un eje sintético, se encuentra estructurado por varias clases de datos. La figura que se detalla a continuación, muestra esta estructura y posiciona a los datos unos respecto de otros en un marco de arquitectura de datos [7].

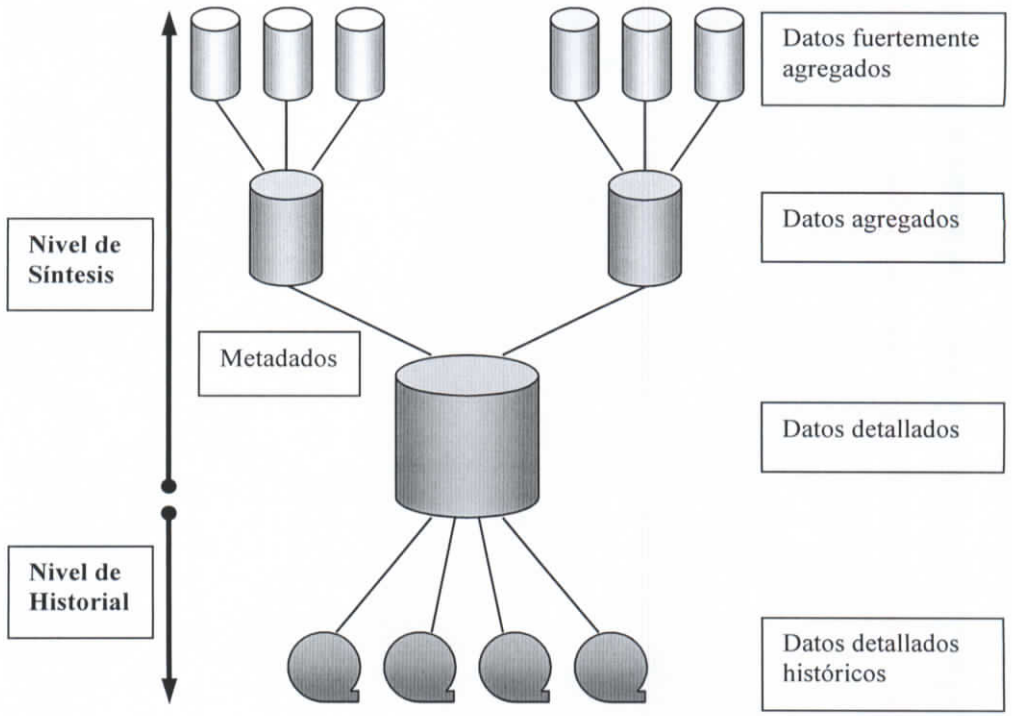


**Figura 2.1:** Estructura del DW

### 2.8.1. Datos detallados

Reflejan o son consecuencia de los eventos más recientes. Las inserciones regulares de datos surgidos de los sistemas de producción habitualmente se realizarán a este nivel. Los volúmenes de datos pueden ser significativamente más grandes que los gestionados tradicionalmente, debido principalmente a las características antes mencionadas del DW.

Por otra parte el nivel de detalle almacenado en el DW no es necesariamente idéntico al nivel gestionado en los sistemas operacionales. El dato insertado puede

ser una agregación[2] o una simplificación de informaciones sacadas de los sistemas de producción.

### 2.8.2. Datos agregados

Se utilizan a menudo porque corresponden a elementos de análisis representativos de las necesidades de los usuarios. Constituyen ya un resultado de análisis y una síntesis de la información contenida en el sistema de decisión. Por lo tanto deben ser fácilmente accesibles; para ello se hace uso de estructuras multidimensionales que permiten a los usuarios navegar por los datos según una lógica intuitiva. Una estructura multidimensional, caracteriza una base de datos en forma de una tabla multidimensional.

En el caso de un agregado, esta información se compone del contenido presentado (media de las ventas, malos resultados, etc.) y de la unidad sobre la que se realiza la agregación (por semestre, por meses, por sucursales, por productos, etc.).

### 2.8.3. Metadatos

Reúnen todas las informaciones respecto al DW y los procesos asociados. Se integran en un denominado referencial. Las principales informaciones van orientadas al usuario, a los equipos responsables de los procesos de transformación de datos, a los equipos responsables de los procesos de creación de agregados, a los

[2] Agregación: partición horizontal de una relación según valores de atributos seguida de una agrupación por una función de cálculo, por ejemplo una cuenta, máximo, mínimo, etc.

equipos de administración de base de datos, etc. Muchos dicen que metadatos son "datos sobre los datos".

### 2.8.4. Datos históricos

También llamados datos historiados. Uno de los objetivos del DW es conservar en línea los datos históricos. Cada nueva inserción de datos proveniente del sistema de producción no destruye los valores anteriores, sino que crea una nueva ocurrencia del dato. El soporte de almacenamiento de los datos históricos depende de la frecuencia de acceso, el tipo de acceso y los costos de soportes.

## 2.9. Framework general

El desarrollo, uso y mantenimiento de un DW son tareas bastante complejas y en ocasiones llevan períodos de tiempo muy largos.

Una razón de su complejidad es el rango de técnicas que se requieren para formular, desarrollar, implementar, desplegar y explotar u proyecto de esta magnitud. Por este motivo conviene simplificar la idea general que se tiene y ser más preciso en la apreciación. De esta manera resulta necesario conocer el framework bajo el cual se desenvuelve el DW.

A continuación se muestra sintetizado el framework general del DW que está formado por tres ámbitos: las aplicaciones, los componentes funcionales y las infraestructuras **[11]**.

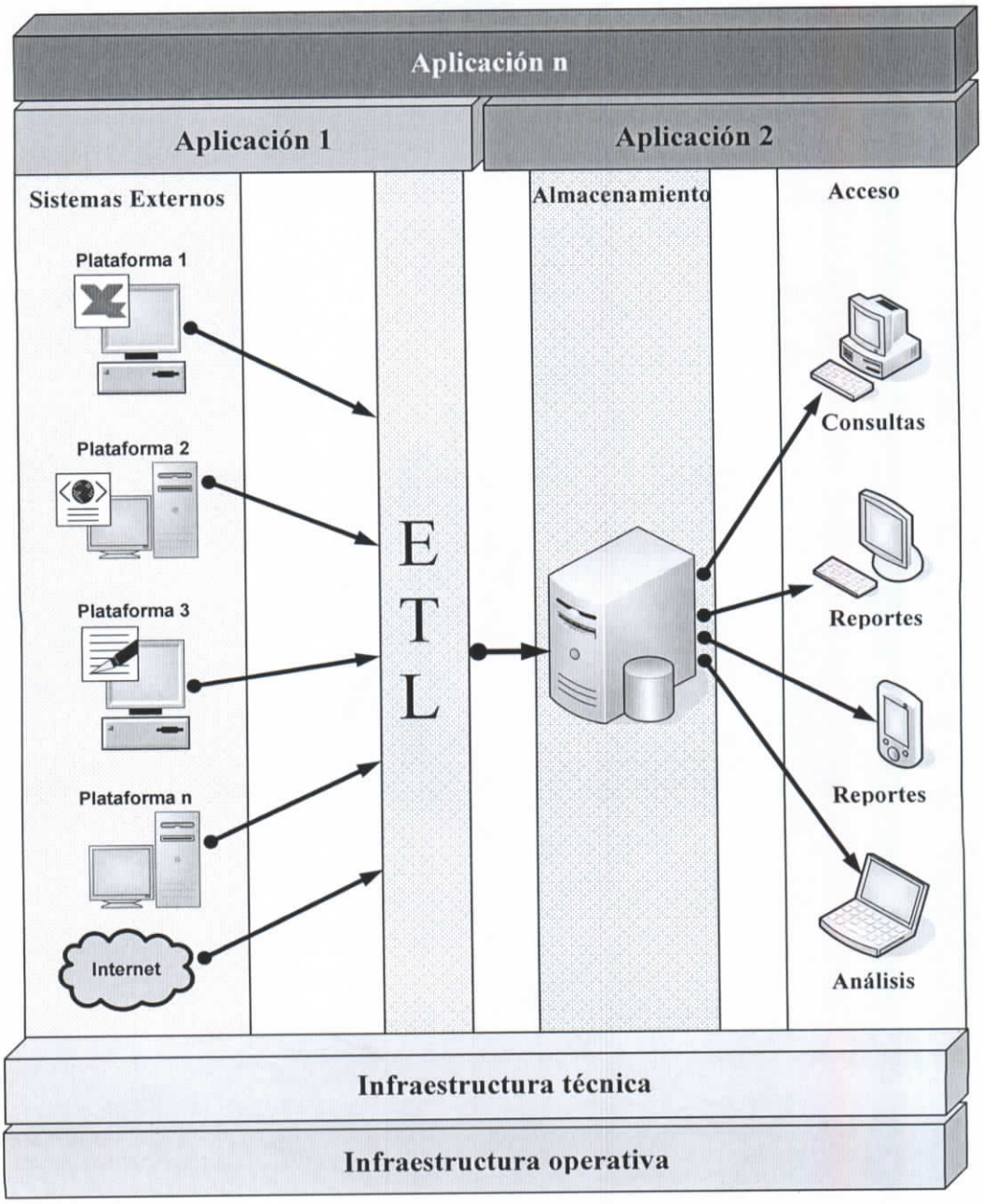


**Figura 2.2:** Framework general del DW

### 2.9.1. Aplicaciones

Se debe tener claro que un DW no se puede construir en una sola iteración. El conjunto de temas que lo conforman se descompone en un grupo de aplicaciones,

conocidas también como "iniciativas". En sí, una iniciativa es un proyecto de decisión que entra en juego durante el desarrollo de un DW en un método iterativo.

El ámbito o perímetro de cada aplicación debe estar claramente definido: actividades y objetivos de la empresa, actores o personajes involucrados, frecuencia y periodicidad de los análisis, entre otros. Estas aplicaciones lógicamente deben ser controlables y proporcionar resultados tangibles en un plazo no mayor a los seis meses, que se puede considerar como plazo medio en la realización de una aplicación. La descomposición en aplicaciones lamentablemente no solo proporcionan grandes ventajas, también producen ciertos problemas sobre varios aspectos o temas relacionados con la infraestructura técnica y operativa. Una aplicación también puede ser un programa, el mismo que debe estar integrado en la infraestructura general y que aporte de una u otra forma a las actividades de toma de decisiones.

### 2.9.2. Componentes funcionales

Son tres las actividades consideradas como componentes funcionales: la adquisición de los datos, su almacenamiento y el acceso por parte de los usuarios finales.

#### 2.9.2.1. Adquisición de los datos

Para simplificar la visión de esta actividad, se hace referencia a ETL. ETL son las siglas en inglés de Extraer, Transformar y Cargar (Extract, Transform and Load). Es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos, limpiarlos y cargarlos en otra base de datos, Data Mart o Data

Warehouse para analizar, o en otro sistema operacional para apoyar un proceso de negocio **[12]**.

ETL es el conjunto de procesos mediante los cuales los datos provenientes de fuentes operacionales son preparados para el DW **[13]**.

También se puede indicar que ETL son los pasos por los que atraviesan los datos para ir desde sistemas OLTP (o la fuente de datos utilizada) a una base de datos dimensional.

#### 2.9.2.1.1. Extracción

La primera parte de un proceso ETL consiste en extraer o recolectar los datos útiles desde las fuentes.

Primeramente se debe identificar los datos considerados como más importantes a fin de extraer la menor cantidad posible, luego planificar estas extracciones para evitar las saturaciones en cuanto a la red, entradas y salidas y unidad central de los sistemas de origen.

La mayoría de los proyectos de almacenamiento de datos, consolidan datos de diferentes sistemas. Cada sistema separado puede usar una organización diferente de datos o formatos distintos. Los formatos de las fuentes normalmente se encuentran en bases de datos relacionales o ficheros planos, pero pueden incluir

bases de datos no relacionales u otras estructuras diferentes. La extracción convierte los datos en un formato preparado para iniciar el proceso de transformación.

Una de las razones que dificultan la tarea de extracción es la redundancia de datos en los sistemas operacionales, redundancia que los programas de extracción deben detectar. Por ejemplo, el elemento de datos que almacena el nombre del cliente puede existir en varios archivos y bases de datos de origen. Estas ocurrencias redundantes deben consolidarse, con el procesamiento y búsqueda en las tablas que eso implica.

Además, es necesario examinar las interdependencias operacionales entre los distintos archivos y bases de datos de donde se va a extraer la información para determinar la secuencia de ejecución de los programas de extracción **[4][12]**.

Para poder extraer datos, se puede hacer uso de tecnologías como **[11]**:

- Gateways o pasarelas, proporcionados generalmente por los Sistemas de Gestión de Bases de Datos (SGBD); estos gateways en ocasiones resultan insuficientes porque están orientados esencialmente a datos y traen consigo dificultades cuando se trata de solucionar problemas de transformación complejos.
- Utilidades de replicación, utilizados si los sistemas de producción y el sistema de decisión son homogéneos y si la transformación a aplicar es ligera o en ocasiones innecesaria. Resulta interesante conocer que la replicación es un mecanismo de copia de datos de una base de datos a otra u otras generalmente

situadas en uno o más servidores, en ocasiones en un medio heterogéneo. Los SGBD proponen mecanismos de replicación automáticos y trasparentes.

- Herramientas específicas de extracción, son sin duda la mejor solución operativa al problema de la extracción, pero debido a su elevado precio, no queda (en ocasiones) otra alternativa que crear nuestra propia aplicación o mecanismo de extracción, haciendo uso de nuevas tecnologías como puede ser el caso de XML (eXtensible Markup Language).

#### 2.9.2.1.2. Transformación

La transformación (también conocida como limpieza) es la etapa por la que puede atravesar una base de datos para estandarizar los datos de las distintas fuentes, normalizando y fijando una estructura. La fase de transformación aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados. Algunas fuentes de datos requerirán alguna pequeña manipulación de los datos, por otra parte en otros casos será necesario aplicar varias transformaciones **[12]**.

Algunos problemas típicos cuando se extraen los datos del entorno operacional se mencionan a continuación **[4]**:

- Claves primarias inconsistentes: las claves primarias en las bases de datos que actúan como fuentes no siempre coinciden con las que se definen en el DW. Por ejemplo, si existen diez archivos de clientes, cada uno con una clave diferente, estas claves se tienen que consolidar o transformar en una única clave en el DW.

- Valores inconsistentes: son datos duplicados que existen en la organización, es decir, elementos que tienen una o más copias exactas, pero debido a actualizaciones inconsistentes y a otras anomalías, dejaron de ser copia exacta del original. Estos valores también deben conciliarse durante el proceso de ETL.
- Datos con diferentes formatos: los elementos tales como fechas o valores monetarios cuentan con una variedad de formatos. Esto da lugar a la necesidad de transformar estos datos a un formato único con el que será almacenado en el DW.
- Valores erróneos: la corrección de valores incorrectos o que violen las reglas de negocios pueden ser muy extensa y complicada. Los programas de transformación realizan cálculos y búsquedas en las tablas para determinar los valores correctos.
- Sinónimos y homónimos: los datos redundantes no siempre son fáciles de reconocer porque el mismo elemento de datos puede tener diferentes nombres en las distintas fuentes. Otra situación que puede darse es que se use el mismo nombre en varias fuentes para referirse a elementos diferentes. Los sinónimos y homónimos no deben existir en el DW, por lo tanto todos estos elementos deben renombrarse con nombres estándares definidos.

Luego de trasformar los datos debido a la existencia de tipos de datos y longitudes incompatibles, o datos inconsistentes o incorrectos, una gran porción del proceso de transformación está destinado a la integración, derivación, agregación y totalización de los datos:

- Integración: es resultado de la integración es que cada elemento de dato único sea conocido por un nombre estándar. Muchos de los datos se renombran en forma acorde a ciertos estándares de nombramiento en el DW, por ejemplo, renombrar Cliente_Cod como Cliente_Codigo.
- Derivación: se crean nuevos datos a partir de datos atómicos en las fuentes, mediante cálculos, búsquedas y lógica procedural. Por ejemplo, generar un nuevo código para clasificar clientes basándose en cierta combinación de datos existentes, calcular el precio total como resultado de multiplicar el precio unitario por la cantidad vendida, o unir la columna Nombre con la columna Apellido para obtener una única columna llamada NombreCliente en el DW. Otro ejemplo consiste en dividir elementos de datos y que cada una de las porciones resultantes conforme una columna en el DW. Por ejemplo, dividir una columna de tipo fecha o timestamp en sus componentes como día, mes, semestre y año.

Las reglas técnicas y de negocios que determinan las transformaciones que se tienen que aplicar se extraen de manuales, memos, emails, programas, y también son propuestas por gente de la organización que recuerda cuándo y por qué se crearon las distintas reglas.

La parte más importante de la transformación de datos consiste en el precálculo de los mismos para que las consultas al DW respondan más eficientemente.

- Totalización (Sumarización): consiste en procesar valores numéricos para obtener valores generales como cantidades, promedios, máximos, mínimos, totales. Estos valores son los que componen las tablas de hechos, y se pueden

calcular y almacenar en distintos niveles, por ejemplo, calcular los totales de venta por departamentos y los totales por regiones.

- Agregación: se refiere a crear datos derivados a partir de la unión de varios datos atómicos, en forma horizontal. Por ejemplo, agregar los elementos de datos de un cliente a partir de la tabla Clientes y de la tabla Ventas para armar un historial de los movimientos del cliente de la empresa.

La mayoría de los datos se totalizan y se agregan basándose en patrones de los reportes requeridos y en la estructura de la base de datos multidimensional elegida.

Es importante examinar en que momento realizar cada transformación. Hay un solo proceso ETL, así que las transformaciones aplicables a todos los datos de origen, como las conversiones de tipos, se deberían realizar en primer lugar. Las transformaciones específicas del DW, como agregaciones y totalizaciones, deberían realizarse hacia el final del proceso.

Por otra parte el propósito de la consolidación de datos es asegurar que toda la información que ingresa al proceso de ETL coincide con la que sale del mismo.

Finalmente, la gente de negocios espera calidad y consistencia en los datos, y esto se logra aplicando todas las transformaciones necesarias y realizando los chequeos correspondientes.

#### 2.9.2.1.3. Carga

La fase de carga es el momento en el cual los datos de a fase anterior son cargados en el destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de procesos diferentes. Algunos DW sobrescriben información antigua con nuevos datos. Los sistemas más complejos pueden mantener un historial de los registros de manera que se pueda hacer una auditoria de los mismos y disponer de un rastro de toda la historia de un dato.

El proceso de carga se puede aplicar de dos maneras: se puede insertar filas nuevas en las tablas mediante código escrito a medida, o hacer una carga masiva usando alguna herramienta de importación del SGBD. Este último enfoque es el más eficiente y el más usado por las organizaciones. Luego de completados los procesos de extracción y transformación, no debería ser muy complicado completar el proceso de carga. Sin embargo, es necesario considerar aspectos como la integridad referencial y los índices **[4]**.

Integridad referencial: debido al gran volumen de los datos, en ocasiones se desactiva esta opción durante la carga para acelerar el proceso, pero la aplicación de ETL tiene que estar capacitada para hacer los chequeos necesarios. Si esto no es así, el DW se volverá corrupto en pocas semanas o meses. La razón por la que se tiende a desactivar la integridad referencial surge del hecho de que se cargan solo datos con relaciones ya existentes en el ambiente operacional, y no se crean nuevas relaciones. Sin embargo, quizá las relaciones existentes no estén bien definidas o sean nulas, o tal vez las bases de datos de origen no sean relacionales. De todas

maneras se de activar una vez completado el proceso de carga, para que el SGBD determine cualquier violación de integridad entre datos relacionales.

Índices: las bases de datos con rendimiento pobre frecuentemente son la consecuencia de un esquema deficiente de índices. Es necesario definir índices en forma eficiente, y en gran cantidad, debido al gran volumen de datos en el DW. Construir los índices en el momento de la carga retarda mucho el proceso, por lo tanto se recomienda eliminar todos los índices antes del proceso de carga, cargar los datos y luego crear los índices nuevamente.

#### 2.9.2.2. Almacenamiento

El componente básico de soporte del almacenamiento es el SGBD. Además del almacenamiento, el SGBD debe proponer extensiones para responder a las características del acceso a la decisión. Estas tecnologías se relacionan principalmente con el paralelismo de las consultas y con diversas optimizaciones propuestas para acelerar las selecciones y las agrupaciones de conjuntos.

Debido a la importancia del historial en un DW, la estructuración física de los datos es también importante. Una partición física de las tablas en unidades menores, según el criterio tiempo, aporta rendimientos estables, facilidades para la recuperación, indexaciones y las reestructuraciones.

Otro aspecto que hay que considerar es el tipo de datos que se van a almacenar y manipular, lo que implica observar las capacidades que tienen los SGBD para

soportar estructuras multimedia compuestas de documentos, imágenes, sonidos, videos que en ocasiones son necesarios almacenarlos. Finalmente es beneficioso utilizar SGBD con últimas tecnologías, que aporten en la evolución del hardware (scalability), la independencia ya sea en el número como en el tipo de procesadores, los discos, la memoria y la evolución de los sistemas operativos.

#### 2.9.2.3. Acceso

Definir una arquitectura global que de soporte a los accesos de decisión impone opciones tecnológicas no estructuradas. Un grupo de usuarios pretenderá efectuar consultas simples o complejas sobre los datos que les interesen; otro grupo querrá efectuar análisis sobre informaciones muy estructuradas y agregadas; otro servicio necesitará hacer simulaciones a partir de la información existente; varias personas se conectarán vía Internet para adquirir cotizaciones, etc. Es entonces razonable pensar que al DW van a acceder personas con distintos puntos de vista de la información.

### 2.9.3. Infraestructuras

Para responder a las necesidades actuales de la información y el conocimiento, la informática se encarga de definir e integrar una arquitectura global sobre la que se basarán las aplicaciones de decisión. Por tal motivo es necesario considerar dos niveles de infraestructura en un DW: técnica y operativa.

#### 2.9.3.1. Infraestructura técnica

Es el conjunto de componentes materiales y programas (software). Está formado de productos que implementan las tecnologías elegidas, integrados en un conjunto coherente y homogéneo. Estas opciones técnicas afectan a los componentes materiales y programas junto con los componentes funcionales que son la alimentación, el almacenamiento y el acceso al DW.

#### 2.9.3.2. La infraestructura operativa

Se compone de todos los procesos que permiten, a partir de los datos de producción, crear y gestionar el sistema de decisión. Las grandes funciones de esta infraestructura conciernen a la administración y el uso del sistema de decisión. Esta última función es muy importante porque afecta a la ordenación y a la gestión de los flujos de datos de los sistemas originales al sistema destinatario **[11]**.

## 2.10. Data Mart como estrategia de diseño del Data Warehouse

El uso efectivo del Data Mart (DM) en un ámbito de Data Warehousing es importante para ganar efectividad y en ocasiones también determinante para el éxito del proyecto. Las organizaciones conocen que un DW corporativo es complejo de construir y de usar. Normalmente requiere un importante equipo de desarrolladores y programadores, inversión en hardware y software, tiempo y recursos. Las necesidades de diferentes áreas de la empresa deben ser analizadas en conjunto por lo que resultan de gestión complicada, es por ello que se tiende al desarrollo del DM.

### 2.10.1. Definición

Varias son las definiciones que se han emitido sobre este término, incluso varios autores afirman que el término DM es sinónimo de base de datos multidimensionales e incluso de OLAP, a continuación se recopilan varias definiciones:

"Data Mart es una base de datos orientada al tema puesta a disposición de los usuarios en un contexto de decisión descentralizado" **[11]**.

"Un Data Mart es una base de datos separada del Data Warehouse; a veces consiste en un subconjunto del Data Warehouse en la misma base de datos. La información soporta la toma de decisiones en un área específica del negocio" **[14]**.

"Data Mart es un conjunto de datos flexible, idealmente basado en los datos más atómicos posibles extraídos de fuentes operacionales, y presentados en modelos simétricos (dimensionales) de cara a inesperadas consultas por parte de los usuarios" **[13]**.

"Data Mart es un depósito de datos que almacena un subconjunto o una agregación de datos del Data Warehouse. Un Data Mart puede ser visto como un pequeño y local Data Warehouse" **[15]**.

"Data Mart es una estructura de datos que es dedicada a proporcionar las necesidades analíticas de un grupo de personas, que puede ser el departamento de contabilidad o el departamento financiero. Existen dos tipos de Data Mart: los

independientes y los dependientes; los primeros son aquellos construidos directamente a partir de las aplicaciones legadas. Por el contrario los Data Mart dependientes son construidos a partir de datos provenientes de un Data Warehouse" [7].

Un DM es una aplicación o extensión de un DW construido para soportar una línea de negocio simple pero manteniendo siempre los valores de integridad, no volatilidad y orientación temática de un DW.

Al hablar de los DM, se puede indicar que existen varios enfoques o estrategias que involucran a su desarrollo: los DM como DW locales (pertenecen a un departamento o área específica del negocio), DM como componentes de un DW (su unión forman un DW) y DM como subconjuntos de un DW (forman parte de un DW).

La siguiente figura explica las tres situaciones indicadas del DM como estrategia de diseño del DW:

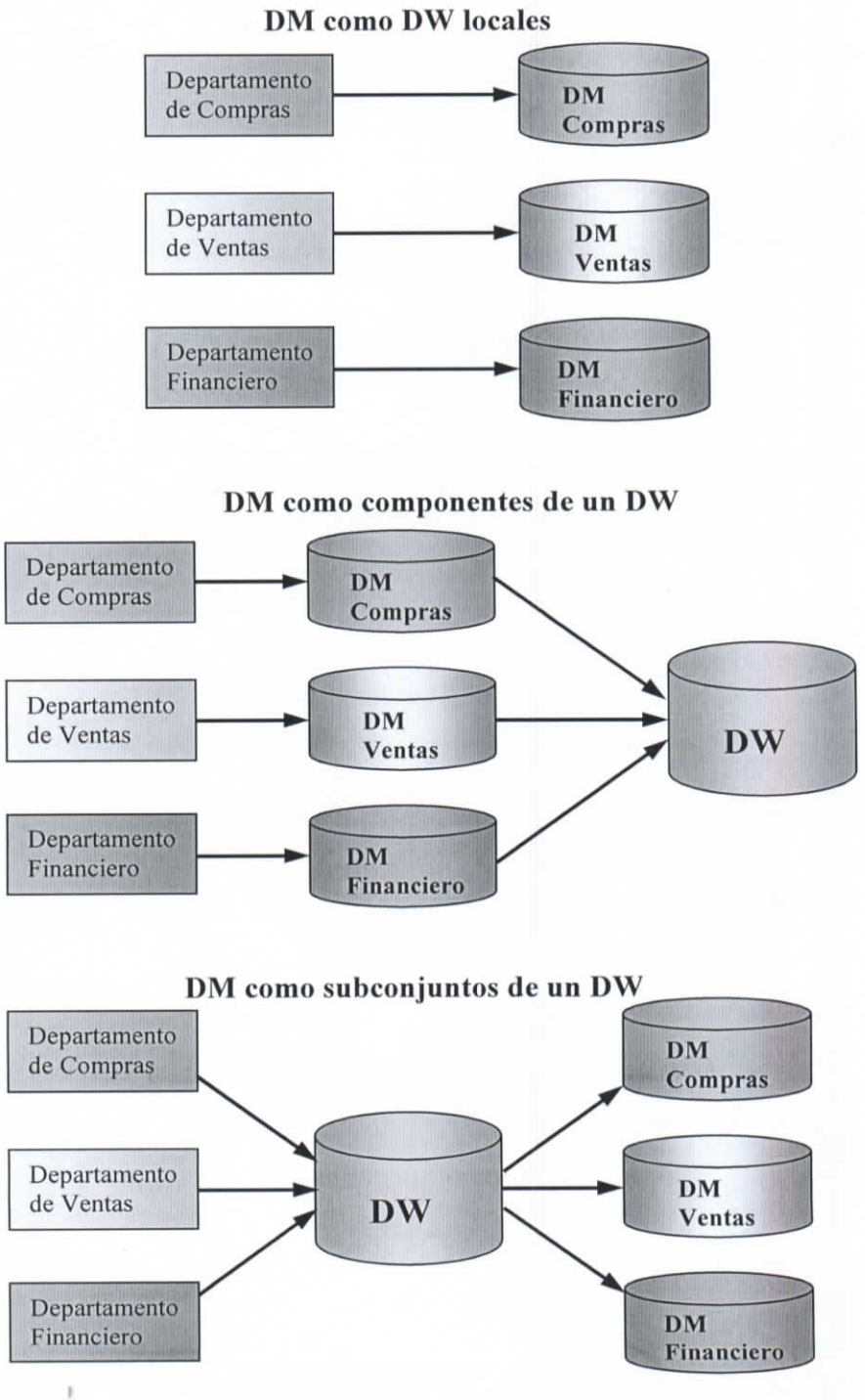


**Figura 2.3:** DM como estrategia de diseño del DW

### 2.10.2. Diferencias con el Data Warehouse

Las principales diferencias a considerar por quienes están involucrados en el desarrollo de un DM o DW son [7]:

- Un DM es creado para satisfacer las necesidades específicas de un departamento o área de acuerdo a los objetivos de los mismos, en cambio, un DW se crea para satisfacer las necesidades globales de la organización.
- Los datos contenidos por el DM, son más recientes históricamente que los contenidos por el DW.
- Debido a que los DM contienen un menor volumen de datos comparado con los DW, los DM son más fácilmente entendibles, administrables y navegables.
- Los DM normalmente no contienen datos operacionales detallados a diferencia del DW.
- El DW es diseñado principalmente a partir de un modelo de datos. El modelo de datos corporativo refleja la abstracción de las necesidades de información de la organización. En cambio el DM refleja los requerimientos específicos necesitados por un departamento determinado.
- El DW lleva consigo un mayor horizonte de tiempo durante su desarrollo comparado con un DM.

### 2.10.3. Razones para crear un Data Mart

Entre las más notables tenemos:

- Proporcionar a los usuarios acceso a los datos que ellos necesitan para analizarlos más a menudo.

- Entregar los datos en una forma que concuerda la vista colectiva de los mismos por un grupo de usuarios en un departamento o función de negocio específico.
- Mejorar el tiempo de respuesta en cuanto a consultas del usuario final debido a la reducción en el volumen de datos a ser accesados.
- Proveer datos apropiadamente estructurados para satisfacer los requerimientos de las herramientas de acceso del usuario final.
- Los DM usan menos datos de tal manera que la tarea de carga es más sencilla.
- El costo de implementación de un DM normalmente es menor que el requerido para establecer un DW.
- Los potenciales usuarios de un DM son más claramente definidos y pueden ser más fácilmente objetivisados para obtener soporte en un proyecto de este tipo.

## 2.11. Modelamiento multidimensional

Una de las cuestiones que enfrentan los profesionales que intervienen en proyectos de Data Warehousing es la del modelo básico de almacenamiento o diseño de la base de datos del DW. Hay dos modelos que por lo general se los puede considerar: el relacional y el multidimensional. El modelo relacional es ampliamente difundido por el enfoque "Inmon", mientras que el modelo multidimensional es más conocido y altamente utilizado bajo la línea de diseño del DW de "Kimball".

Los dos modelos gozan de ventajas y desventajas, lo que hay que tener claro es que el enfoque relacional tiene mayor incidencia en un entorno organizacional a largo

plazo, mientras que el enfoque multidimensional resulta más apropiado a corto plazo y a un tema o área de acción más reducida o específica de toda la organización. Se menciona también que el modelamiento multidimensional apoya al funcionamiento y análisis de información proveniente de un DM **[7]**.

Por otra parte, el presente trabajo de investigación está orientado a un tema o proceso específico de la Universidad Técnica de Ambato que realiza semestralmente con sus aspirantes (la Prueba de Aptitud Académica, PAA). Con esta referencia y otras características antes comentadas, resulta necesario indicar que este proyecto entra en el campo de desarrollo de un DM. Al ser este el caso, y con la premisa de que se requiere un modelado del tipo multidimensional, a continuación se amplia más sobre el tema.

"Los sistemas de gestión de bases de datos multidimensionales, DM o en ocasiones también llamados como procesamiento OLAP, proporcionan un sistema de información con la estructura que permite a la organización tener un acceso muy flexible a los datos, para rebanarlos y rotarlos (Slice & Dice) de cualquier número de formas, y para dinámicamente explorar la relación entre resúmenes y datos detallados. Los sistemas de gestión de bases de datos multidimensionales ofrecen flexibilidad y control al usuario final, y como tal encajan bien en un ambiente de DSS" **[7]**.

"El modelamiento multidimensional, también conocido como modelado dimensional es una técnica de diseño lógico que busca presentar los datos en un

área de trabajo o framework estándar que es intuitivo y que permite un alto rendimiento de acceso" **[16]**.

"El modelamiento dimensional es una técnica para conceptualizar y visualizar modelos de datos que son descritos mediante aspectos comunes del negocio" **[17]**.

El modelado dimensional parte del hecho de que el objetivo principal de un sistema de decisión es el análisis del rendimiento. Dicho rendimiento puede materializarse a través de un conjunto de indicadores. Se enfoca en datos numéricos como: balances, cantidades, ocurrencias, promedios, sumas, tasas, totales, etc. Así por ejemplo en un hospital la tasa de ocupación de camas será un indicador importante, mientras que el número de artículos en stock interesará a la logística de una empresa que vende productos. En el DM el modelo dimensional es más expresivo y simple que el modelo entidad/relación. Un modelo dimensional tiene asociado otros conceptos como: medidas, hechos y dimensiones.

### 2.11.1. Medidas, hechos y dimensiones

**Medidas o métricas**, son características cualitativas o cuantitativas de los aspectos u objetos de negocio que se desean analizar en las organizaciones. Las medidas cuantitativas están dadas por valores o por cifras porcentuales. Por ejemplo: las ventas en dólares, cantidad de unidades en stock, cantidad de unidades de productos vendidos, horas trabajadas, el promedio de piezas producidas, el porcentaje de aceptación de un producto, el consumo de combustible de un vehículo, la cantidad de preguntas respondidas, etc.

**Hechos**, un hecho es una colección de datos relacionados con los temas, que consta de las medidas y datos de contexto. Cada hecho normalmente representa un tema del negocio, una transacción comercial, o un acontecimiento que puede ser utilizado en el análisis de la organización o proceso de negocio.

**Dimensiones**, una dimensión es una colección de propiedades a lo largo de las cuales conducimos nuestro análisis de los hechos. Las dimensiones nos posibilitan una vista de los hechos bajo diferentes contextos. Son objetos del negocio con los cuales se puede analizar la tendencia y el comportamiento del mismo. Las definiciones de las dimensiones se basan en políticas de la compañía o del mercado, e indican la manera en que la organización clasifica o interpreta su información para segmentar el análisis en sectores, facilitando la observación de los datos. Para determinar las dimensiones requeridas para analizar los datos, podemos basarnos en varias preguntas:

- ¿Cuándo?, la respuesta a esta pregunta permite establecer la dimensión del tiempo y visualizar de manera comparativa el desempeño del negocio.
- ¿Dónde?, esta pregunta nos ubica en un área física o imaginaria donde se están llevando a cabo los movimientos que se desean analizar. Estos lugares pueden ser zonas geográficas, divisiones internas de la organización, sucursales, etc.
- ¿Qué?, es el objeto del negocio, o el objeto de interés para determinada área de la compañía. Para estos casos se tienen los productos y/o servicios, la materia prima como elemento de interés para la división de abastecimientos,

los vehículos para la sección de transporte, las maquinarias para el área de producción, etc.

- ¿Quién?, plantea una estructura de los elementos que inciden directamente sobre el objeto de interés. En estos casos se hace referencia al área comercial o de ventas, o a los empleados de la organización si se está llevando a cabo un análisis a nivel de talento humano por ejemplo.
- ¿Cuál?, habla de hacia donde se enfoca el esfuerzo de la organización o de una determinada área del negocio, para hacer llegar los productos o servicios. Una dimensión que surge de esta pregunta puede ser clientes.

Además, una dimensión puede estar compuesta de varios niveles de información, y estos niveles están formados por miembros. Por ejemplo, la dimensión tiempo está compuesta de niveles como: día, semana, mes, trimestre, año; cada uno de estos niveles a su vez están formados por miembros. Los miembros del nivel día pueden ser: lunes, martes, miércoles, etc. A los miembros de una dimensión se los puede agregar en uno o más niveles, dichos niveles constituyen las llamadas jerarquías de la dimensión. En la dimensión tiempo se pueden distinguir las jerarquías como días, meses, etc.

### 2.11.2. Representación física

En una base de datos multidimensional, el modelo de datos está constituido por lo que se denomina un Cubo Multidimensional, Cubo de Información, Cubo OLAP o simplemente Cubo.

"Nos referimos a Cubos de información cuando hablamos sobre bases de datos multidimensionales, en las cuales su almacenamiento físico es a partir de vectores multidimensionales. Para acceder a los datos sólo es necesario indexarlos a partir de los valores de las dimensiones o ejes" **[18]**.

En los cubos OLAP la información se representa por medio de matrices multidimensionales o cuadros de múltiples entradas que nos permite realizar distintas combinaciones de sus elementos para visualizar los resultados desde distintas perspectivas o puntos de vista y variando los niveles de detalle. Esta estructura es independiente del sistema transaccional de la organización, facilita y agilita la consulta de información histórica ofreciendo la posibilidad de navegar y analizar los datos. Los ejes del cubo son las dimensiones y los valores que se presentan en la matriz son las medidas. Una instancia del modelo está determinada por un conjunto de datos para cada eje del cubo y un conjunto de datos para la matriz.

En el siguiente gráfico se puede apreciar la representación de un cubo denominado CuboAutosVendidos, que está formado de las dimensiones: auto, tiempo y región. Hace relación a la cantidad de unidades (autos) que se han comercializado durante un período de tiempo.

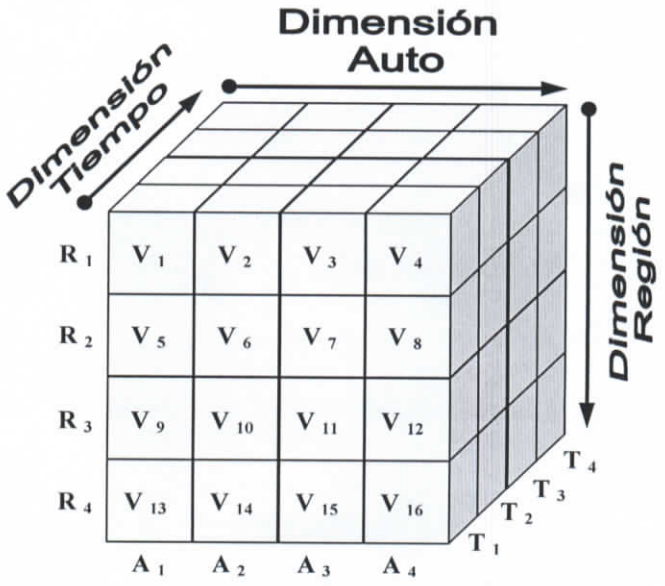


**Figura 2.4:** Representación de un cubo

En la dimensión Auto: $A_1$, $A_2$, $A_3$, $A_4$ pueden representar los modelos o marcas de dichos Autos (por ejemplo Optra, Aveo, Evolution, Spark); en la dimensión Tiempo: $T_1$, $T_2$, $T_3$, $T_4$ representan los trimestres asociados a las ventas (primer trimestre, segundo trimestre, tercer y cuarto trimestre); en la dimensión Región: $R_1$, $R_2$, $R_3$, $R_4$ hace referencia a la región donde se ha comercializado el auto (Costa, Sierra, Oriente o Galápagos). Finalmente $V_1$, $V_2$, $V_3$, $V_4$,….., $V_{16}$ corresponden a las unidades de los modelos de autos vendidos en una región dada y en un trimestre determinado.

## 2.12. Esquemas de modelamiento multidimensional

Conociendo que a los cubos en el ámbito dimensional también se les conoce como tablas de hechos, donde cada eje del cubo corresponde a una dimensión de la misma

y por otra parte en vista de la complejidad de representar más de tres dimensiones, hace necesario la utilización de otras técnicas de modelado dimensional, a las que también se les conoce como esquemas o modelos en estrella (star) y copo de nieve (snowflake).

### 2.12.1. Star

Esta técnica de modelado dimensional, consiste en distinguir físicamente las tablas de hechos de las tablas de dimensiones. La tabla de hechos se coloca en el centro del modelo, y las tablas de dimensiones gravitan alrededor. Es por esta razón que este modelo representa visualmente una estrella **[16]**.

El identificador de la tabla de hechos es una clave múltiple compuesta de la concatenación de claves de cada una de las dimensiones de análisis. Así por ejemplo, una cifra de negocio almacenada en una tabla de hechos, podrá ser identificada por la dimensión tiempo, cliente, producto, región, etc.

Como ya ha sido dicho con anterioridad, alrededor de la tabla de hechos figuran los elementos que caracterizan las dimensiones de análisis. Estas características se agrupan en las tablas de dimensiones. Así por ejemplo, la tabla correspondiente a la dimensión Productos incluirá todas las informaciones interesantes a analizar sobre el producto como: color, categoría, precio, etc.

Este modelo resulta ser asimétrico, pues hay una tabla dominante en ele centro con varias conexiones a las otras tablas. Además se caracteriza por la legibilidad y el

rendimiento. En cuanto a la legibilidad, se puede indicar que este modelo es muy elocuente para el usuario y presenta de manera clara su finalidad. Está claramente orientado al tema y define con notoriedad los indicadores de análisis. En lo referente al rendimiento, los caminos de acceso a la base de datos son previsibles. En este modelo, la tabla de hechos puede soportar varios millones de filas pero las tablas de dimensiones serán mucho más reducidas.

Resulta más fácil de controlar y optimizar los accesos a las tablas de hechos porque se accede a ellas tras haber efectuado selecciones sobre las tablas de dimensiones. Estas selecciones darán como resultado identificadores, que son los únicos puntos de entrada para acceder a las tablas de hechos. De esta manera puede evitarse un recorrido total por estas tablas con una buena indexación. Será por tanto más fácil tener tiempos de respuesta proporcionales al resultado esperado. Además las tablas de hecho sólo contienen valores de informaciones numéricos e identificadores. Son importantes en número de filas, pero el tamaño de cada fila es reducido.

El siguiente gráfico, muestra un modelo en estrella simple:

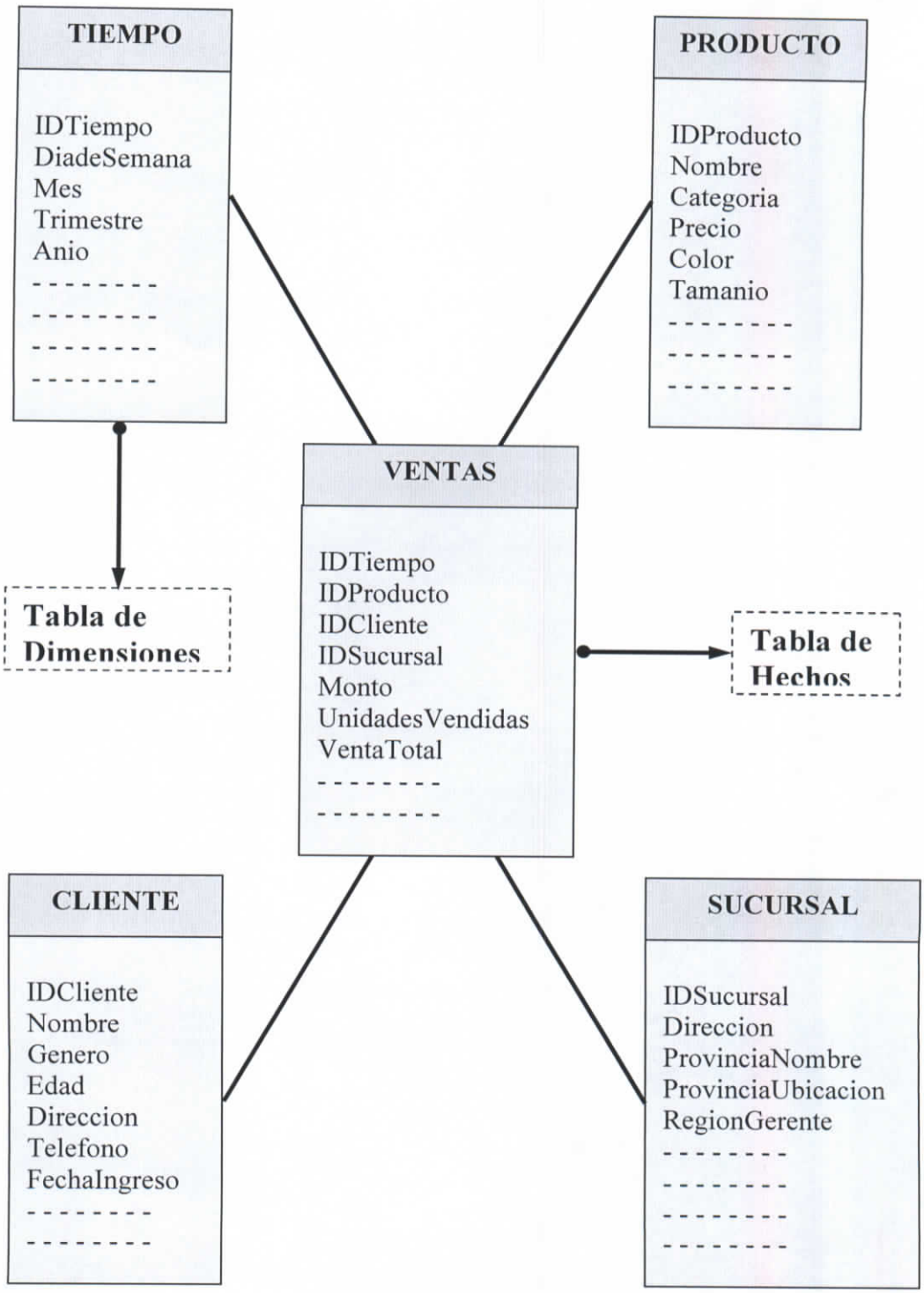


**Figura 2.5:** Ejemplo de un modelo en estrella

### 2.12.2. Snowflake

El copo de nieve es simplemente una estrella cuyas puntas se descomponen a su vez en jerarquías. Modelar en copo de nieve significa conservar los principios del modelo en estrella, es decir, la o las tablas de hechos y afinar el modelado de las tablas de dimensiones que son divididas en subtablas **[16]**.

Resulta interesante este modelado ya que por una parte, normaliza las dimensiones, reduciendo el tamaño de cada una de las tablas; y por otra parte este modelo permite formalizar la noción de jerarquía en el interior de una dimensión. Por ejemplo para la dimensión Tiempo, se podrá definir subentidades como año, mes, semana, etc.

Por otra parte, se nota que los modelos en copo de nieve son un tanto más complejos de gestionar que los modelos en estrella.

El siguiente es un ejemplo de modelado en copo de nieve, donde se visualiza que varias tablas de dimensiones se han dividido en al menos una nueva subtabla.

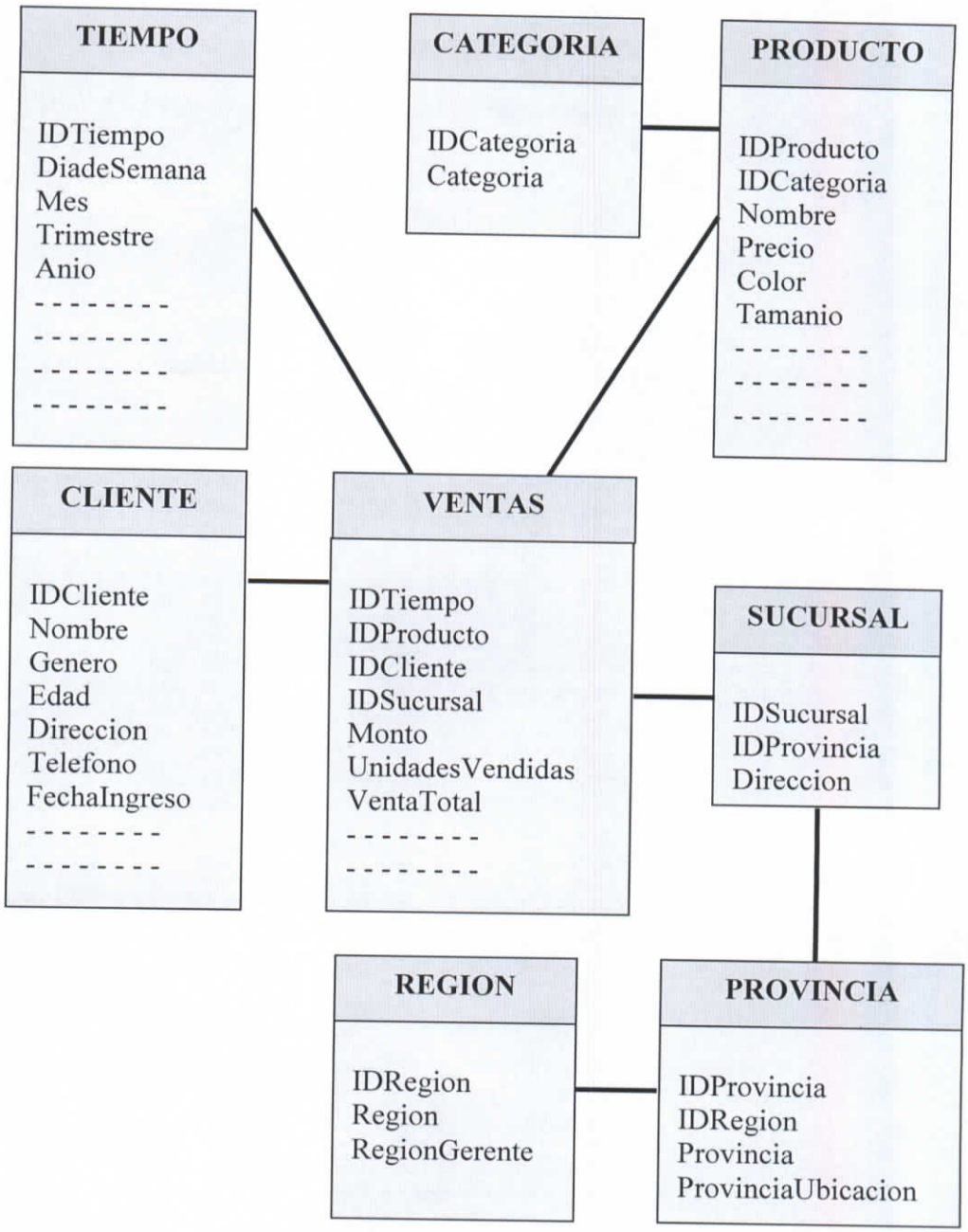


**Figura 2.6:** Ejemplo de un modelo en copo de nieve

## 2.13. Arquitecturas OLAP

OLAP se define como el análisis multidimensional e interactivo de la información del negocio a escala empresarial. El análisis multidimensional consiste en combinar distintas áreas de la organización, y así ubicar ciertos tipos de información que revelen el comportamiento del negocio **[17]**.

Se dice que el análisis es interactivo porque los usuarios de las herramientas OLAP se mueven con facilidad desde una perspectiva del negocio a otra. Así por ejemplo pueden estar observando las ventas anuales por sucursal y pasar a ver la sucursal con más ganancias en los últimos seis meses, además con la posibilidad de elegir entre diferentes niveles de detalle como son las ventas diarias o mensuales. Este tipo de exploración es lo que distingue a OLAP de las herramientas simples de consulta que incluso generan reportes.

El aspecto de la multidimensionalidad es útil debido a que permite a los analistas de negocios examinar sus indicadores clave o medidas como: ventas, costos y ganancias, desde distintos puntos de vista, como períodos de tiempo, productos y regiones. Estas perspectivas constituyen las dimensiones mediante las cuales se explora la información.

OLAP se maneja a escala empresarial ya que es robusto y escalable al punto de permitir satisfacer las necesidades de análisis de información de toda la organización. Es decir, que se comparte y cruza la información en todas las áreas de la misma.

Se dice que OLAP tiene como sus componentes a: un modelo multidimensional de la información para el análisis interactivo; un motor OLAP que procesa las consultas multidimensionales sobre los datos; y un mecanismo de almacenamiento para guardar los datos que se van a analizar. Este componente puede ser externo como un DW o como un sistema de gestión de bases de datos relacional (SGBDR).

Quienes usan OLAP se enfocan en los conceptos del negocio, trabajando intuitivamente con ellos, sin necesidad de conocer aspectos técnicos tales como el formato físico de los datos, instrucciones de lenguajes como SQL, nombres de tablas o columnas de las bases de datos u otras arquitecturas relacionadas con OLAP.

Por otra parte, la estructura que almacena los datos de una aplicación tipo OLAP tiene que proveer consultas eficientes, escalabilidad y acceso multiusuario. Las bases de datos relacionales están optimizadas para obtener un rendimiento óptimo en consultas simples y frecuentes, pero no funcionan de manera ideal para las consultas complejas y multidimesionales de estas aplicaciones. Así se tiene que las herramientas OLAP que usan almacenamiento multidimensional son llamadas MOLAP, las que almacenan los datos en bases relacionales se les llama herramientas ROLAP, y las que combinan los dos enfoques anteriores son OLAP Híbrido o herramientas HOLAP.

### 2.13.1. ROLAP

A pesar de que las bases de datos relacionales no están optimizadas para el análisis multidimensional, tienen ciertas ventajas. En particular, son escalables a conjuntos más grandes de datos e incluyen soporte para replicación, rollback y recuperación. Además, en la mayoría de las organizaciones es probable que la gente de sistemas esté más familiarizada con la gestión de bases de datos relacionales que multidimensionales **[16]**.

Las herramientas del tipo ROLAP brindan análisis multidimensional sobre datos almacenados en una base de datos relacional. Lo hacen a través de un mapeo entre los datos en el DW a un modelo multidimensional, usando consultas SQL.

Debido a su naturaleza, ROLAP es ideal para grandes bases de datos **[17]**.

### 2.13.2. MOLAP

Las bases de datos OLAP multidimensionales usan estructuras de tipo arreglo para almacenar los datos. Estas estructuras están indexadas con el fin de proveer un tiempo de acceso óptimo a cualquier elemento **[16]**.

Se pueden distinguir dos enfoques en la forma de organizar estas estructuras, las bases de datos multidimensionales que usan una arquitectura de hipercubo y las que usan multicubos.

La arquitectura hipercubo almacena un único gran cubo en el cual cada medida está referenciada y totalizada en todas las dimensiones del mismo.

En la arquitectura de multicubos los datos se guardan en más de un cubo. Por ejemplo, una base de datos puede contener un cubo que almacena las ventas mensuales, por región y por producto, y otro que guarde la información correspondiente a costos departamentales y mensuales.

Por su naturaleza, MOLAP es apropiado para pequeños y medianos conjuntos de datos **[17]**.

### 2.13.3. HOLAP

Existen dos enfoques para definir a las herramientas HOLAP.

Por un lado, se habla de una base de datos multidimensional que puede recuperar y analizar información detallada. Esta es la definición más aceptada de HOLAP. Se habla de una herramienta que almacena los datos totalizados en la base de datos multidimensional y los datos detallados en el SGBDR. Los usuarios trabajan con un único modelo de los datos, y acceden de forma transparente a los dos tipos de almacenamiento. Lo interesante de este enfoque es su facilidad de uso **[16]**.

Por otro lado, existen herramientas HOLAP que consisten de un almacenamiento multidimensional, con preconsolidación opcional. Se extrae un conjunto de datos de un SGBDR y luego se construye un cubo multidimensional en el cliente. La

diferencia con el enfoque anterior es que en este caso no se cuenta con una capa de manejo de base de datos que abstraiga al usuario de la implementación técnica. Algunos vendedores incluyen la opción de preconsolidar valores en el almacenamiento, y otros almacenan los datos y hacen las consolidaciones en el momento en que se requieren.

# CAPÍTULO III

## 3. METODOLOGÍAS PARA EL DESARROLLO DEL DATA WAREHOUSE

En un proyecto de desarrollo de software los informáticos basan sus actividades en el empleo de varias metodologías, conocidas también como ciclo de vida o ciclo de desarrollo y destinadas a la solución de tópicos específicos.

El ciclo de vida de los sistemas de procesamiento transaccional usualmente pasa por varias fases, entre las que se pueden considerar: planificación, análisis, diseño, desarrollo, testeo o pruebas e implementación. El desarrollo de sistemas cliente / servidor y distribuidos han modificado el ciclo de vida del desarrollo de sistemas, pero también se mueve dentro de un conjunto predecible de tareas, que cubren desde el comienzo del ciclo hasta la implementación de un sistema en producción funcionando completamente.

Igualmente, el desarrollo de un DW para el soporte de la toma de decisiones, también pasa por varias fases predecibles, a menudo simplemente llamadas Ciclo de vida para el soporte de las decisiones, el cual, a pesar que puede contener muchas de las fases antes mencionadas, es diferente del ciclo de vida del desarrollo de sistemas tradicionales. Las diferencias entre estos dos enfoques de desarrollo aparecen porque los objetivos y las estructuras de datos de los sistemas basados en transacciones en línea y sistemas de soporte para la toma de decisiones son diferentes.

El foco principal de un DW es el dato, no el procesamiento del negocio y su funcionalidad asociada. La funcionalidad del proceso de negocio operacional no es el mayor componente del ciclo de vida para el soporte de las decisiones. Esto permite un ciclo de vida de desarrollo mucho más rápido ya que el modelamiento de procesos y otras tareas asociadas con el desarrollo de la funcionalidad del negocio generalmente no son necesarias.

Existen tres puntos de vista bajo los que se han originado y funcionan las metodologías relacionadas con el DW: de arriba hacia abajo (de lo general a lo particular), de abajo hacia arriba (de lo particular a lo general) y una combinación de los anteriores **[14] [17]**.

**De arriba hacia abajo**, se identifican primero los requerimientos empresariales que debe cubrir el DW propuesto, que son los principales conductores de su implementación. El establecimiento del ámbito (y por ende, de los requerimientos) se especifica como las fronteras de datos que definirán el territorio del DW. Este punto de vista puede producir las siguientes ventajas: los requerimientos empresariales delinean claramente las fronteras de la implementación del DW; otra ventaja es que la tecnología es conducida por el negocio y no a la inversa; también se puede indicar que resulta fácil comunicar los beneficios del DW a quienes toman las decisiones y a los inversionistas **[14] [17]**.

Igualmente surgen ciertas desventajas de este punto de vista: en ocasiones, pueden quedar oportunidades fuera del horizonte empresarial. Estas oportunidades perdidas son el resultado de demasiada concentración; otro inconveniente es que la

tecnología puede impulsar el negocio y ofrecer una ventaja competitiva que, en principio, no es muy obvia para las actividades; finalmente las expectativas iniciales pueden restringir la persecución de objetivos con recompensas potencialmente mayores.

**De abajo hacia arriba**, generalmente comienza con experimentos y prototipos basados en tecnología. Se selecciona un subconjunto específico, bien entendido, de la problemática empresarial y se formula una solución para este subconjunto. Por lo regular, la implementación bajo este punto de vista es más rápida, ya que comprende menos gente tomando menos decisiones para resolver un problema empresarial más reducido. Este enfoque permite que una organización avance con un gasto considerablemente menor y que evalúe los beneficios de la tecnología antes de establecer compromisos significativos **[14] [17]**.

Como ventajas de este punto de vista se tiene: en ocasiones, las necesidades de implementación y de comenzar de una vez, supera por mucho el análisis de arriba hacia abajo y las consideraciones a largo plazo; en las primeras etapas de madurez de la tecnología, este punto de vista permite a una organización evaluar los beneficios de la tecnología sin grandes compromisos; también la participación de poca gente trabajando en un ámbito reducido puede acelerar la implementación y la toma de decisiones.

Las desventajas asociadas son: después de la implementación inicial, es bueno retroceder y observar cómo puede ampliarse la solución para dar servicio a toda la empresa; la falla de un solo proyecto de abajo hacia arriba puede demorar la

implementación de una tecnología benéfica en potencia; por último, el equipo inicial debe retroceder e integrarse a un mayor equipo para ampliar el ámbito de la solución inicial.

**Combinado**, bajo este punto de vista una organización puede explotar la naturaleza planeada y estratégica del punto de vista de arriba hacia abajo, al tiempo que conserva la rápida implementación y aplicación oportunista del punto de vista de abajo hacia arriba. Este punto de vista depende de dos componentes: primeramente una arquitectura de arriba hacia abajo, estándares y un equipo de diseño que aplique la experiencia de proyecto en proyecto y que pueda retroceder y convertir las decisiones tácticas en decisiones estratégicas; en segundo lugar, un equipo de proyecto de abajo hacia arriba que se concentre en implementar una solución empresarial muy enfocada, estrecha, pero de largo alcance, en un período de tiempo reducido **[14] [17]**.

En la actualidad, se puede hacer uso de un considerado número de metodologías asociadas al DW, mundialmente reconocidas y que han aportado significativamente al desarrollo exitoso de proyectos de esta magnitud. A continuación se detallan brevemente varias metodologías, consideradas como las más importantes y más usadas.

## 3.1. Rapid Data Warehousing de SAS Institute

El Instituto SAS considera que suceden dos causas por las que el desarrollo de un DW fracase: la experiencia de los desarrolladores y el tiempo excesivo que puede

durar el proyecto. Los consultores de SAS indican que su metodología intenta resolver estos inconvenientes, dividiendo al proyecto en pequeñas fases a las que en ocasiones llaman "figuras" con lo que se logra además un retorno rápido de la inversión **[19]**. El siguiente gráfico sintetiza la metodología:

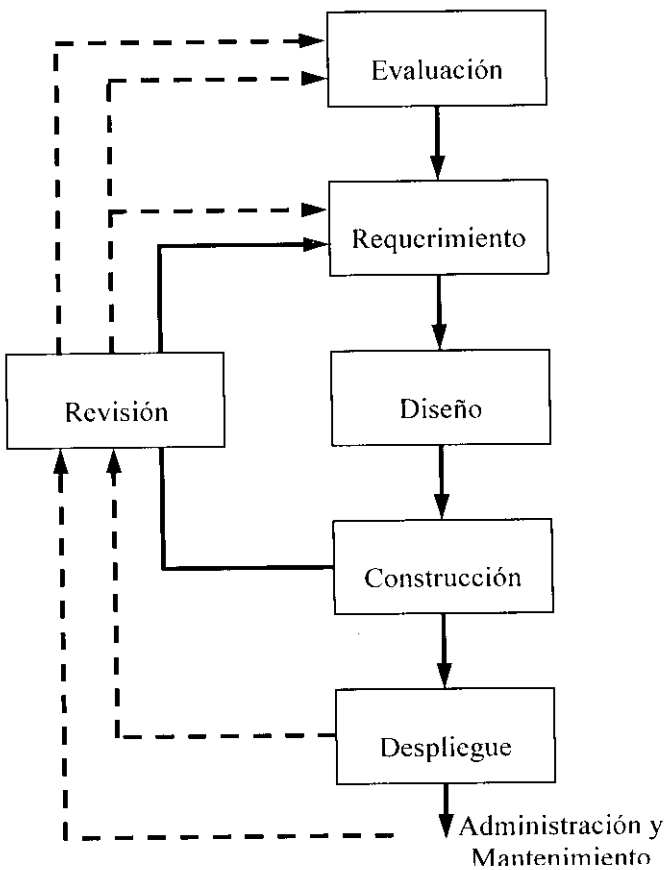


**Figura 3.1:** Fases de la metodología: Rapid Data Warehousing Methodology

A continuación se describen brevemente las fases de esta metodología:

**Evaluación**

La fase de evaluación es crucial para determinar si la organización está dispuesta a emprender en un proyecto de Data Warehousing y de qué alcance. Si una organización no está preparada para llevar a cabo el proyecto descrito en la

evaluación, se puede tomar la decisión de aplazar el proyecto. El equipo de evaluación puede recomendar una serie de medidas o pasos para lograr dicha disposición.

Mediante la realización de una fase de evaluación, los riesgos que podrían dar lugar a gastos adicionales y finalmente al fracaso de los proyectos son identificados, minimizados o eliminados.

Para algunas organizaciones, la construcción de un DW resulta una iniciativa preliminar a otras de negocio más centrada, como la minería de datos o la gestión de relaciones con clientes (CRM). La metodología se integra a la perfección con otras metodologías incluso aunque no sean relacionadas con el Instituto SAS.

**Requerimientos**

La fase de requisitos se orienta a un alto nivel de necesidades de todo el DW. Al mantener una perspectiva de alto nivel, el proyecto puede evitar la denominada "parálisis de análisis" que, a menudo, daña otros enfoques como el de cascada, manteniendo el tiempo total dedicado a esta fase al mínimo. Tanto los requerimientos del negocio como los requerimientos técnicos (incluyendo la infraestructura) necesarios son reunidos durante esta fase.

Entrevistas, talleres, y el análisis de los documentos y sistemas existentes se pueden utilizar para recopilar y confirmar los hechos necesarios. El documento resultante de los requisitos es revisado por todas las partes afectadas. En ese documento se incluye la identificación de los objetivos de negocio, así como un análisis de

viabilidad técnica. Las recomendaciones pueden incluir el número de proyectos a construir y la prioridad de cada construcción.

**Diseño**

La fase de diseño se centra en la construcción de un proyecto a la vez. Al restringir el ámbito de aplicación antes de avanzar a otro nivel de detalle, el almacén puede seguir evolucionando rápidamente de manera sistemática. Las actividades para esta fase incluyen:

- Análisis detallado de los requisitos para la construcción seleccionada.
- Detallados diseños físicos y lógicos para el modelo de datos.
- Especificación detallada de un modelo para el proceso de extracción, transformación y carga.
- Creación de un modelo para la aplicación o selección de herramientas de explotación.
- Diseño de aspectos adicionales como de seguridad y modelos de metadatos.

El documento de requisitos se utiliza como entrada a esta fase, que produce un documento detallado de diseño para la construcción seleccionada. Para construcciones posteriores, el documento de diseño se basa en los anteriores que pueden ser utilizados como entrada para asegurar que el trabajo incorpora decisiones antes tomadas.

Tanto el director del proyecto como el arquitecto del DW pueden esperar estar involucrados o participar a tiempo completo en esta fase. Representantes técnicos y de negocios pueden proporcionar comentarios a las líneas de diseño.

**Construcción**

Durante la fase de construcción, los equipos de implementación codifican y se encargan de poblar el DW, desarrollan las aplicaciones para el usuario final y realizan el análisis y la presentación de informes. Los usuarios del negocio y de TI (Tecnologías de la Información) rigurosamente prueban o examinan el DW y las aplicaciones para verificar que todos los criterios de aceptación se cumplen. Un aspecto importante de esta etapa es la preparación para la implantación del sistema de producción. Los consultores y gestores del proyecto revisan los procesos para la creación, actualización y mantenimiento del DW con el administrador del mismo.

Un documento de mantenimiento es preparado, que guarda esta información como referencia futura. Los individuos de las unidades de negocio y las TI de la organización están implicados en la explotación de las aplicaciones de pruebas como una manera eficaz de la transferencia de conocimientos.

**Despliegue**

La fase de despliegue consiste en la implantación del DW y de las aplicaciones para el usuario final y al personal de TI. Hay que asegurarse de que los usuarios estén bien capacitados y que las aplicaciones son fácilmente accesibles a los datos, lo que ayuda a promover la aceptación generalizada de todo el proyecto.

Los usuarios del negocio que más rápido obtengan algún beneficio del DW, son probablemente quienes sigan apoyando los esfuerzos de desarrollo o mejora.

**Revisión**

Tres ejecuciones de revisión se llevan a cabo en esta fase con cada construcción:

Una después de la fase de construcción para evaluar el proceso de implementación y aprender de los éxitos y fracasos.

En el lapso de 3 a 6 meses después de revisar la fase de despliegue y asegurarse de que la transición hacia el apoyo ha ido sin problemas y que los usuarios tengan acceso al DW.

Luego de 18 a 24 meses después de la primera construcción en la que se examina cualquier tipo de beneficio tangible, se puede calcular el retorno de la inversión (ROI), y se puede asegurar que el medio ambiente del DW continúa satisfaciendo las necesidades de la comunidad empresarial.

Una vez que el DW ha sido implementado, se requiere una administración y mantenimiento continuo del mismo. Estas tareas incluyen actividades como:

- Adición de nuevos usuarios.
- Capacitación a los usuarios.
- Adición de consultas, tablas y vistas previamente unidas, datos agregados.
- Gestión contínua de la calidad de los datos.
- Gestión de los sistemas DW, incluyendo almacenamiento, monitoreo y administración de datos.
- Suministro de mejoras como apoyo a cambio ligero de necesidades.

## 3.2. Ralph Kimball

Ralph Kimball es considerado como uno de los mayores defensores del modelado multidimensional, su metodología es iterativa, con cada iteración se enfoca un aspecto simple del negocio y está asociado a la creación de un esquema en estrella sencillo basado en un DM.

El siguiente gráfico muestra las fases o etapas que forman parte de esta metodología:

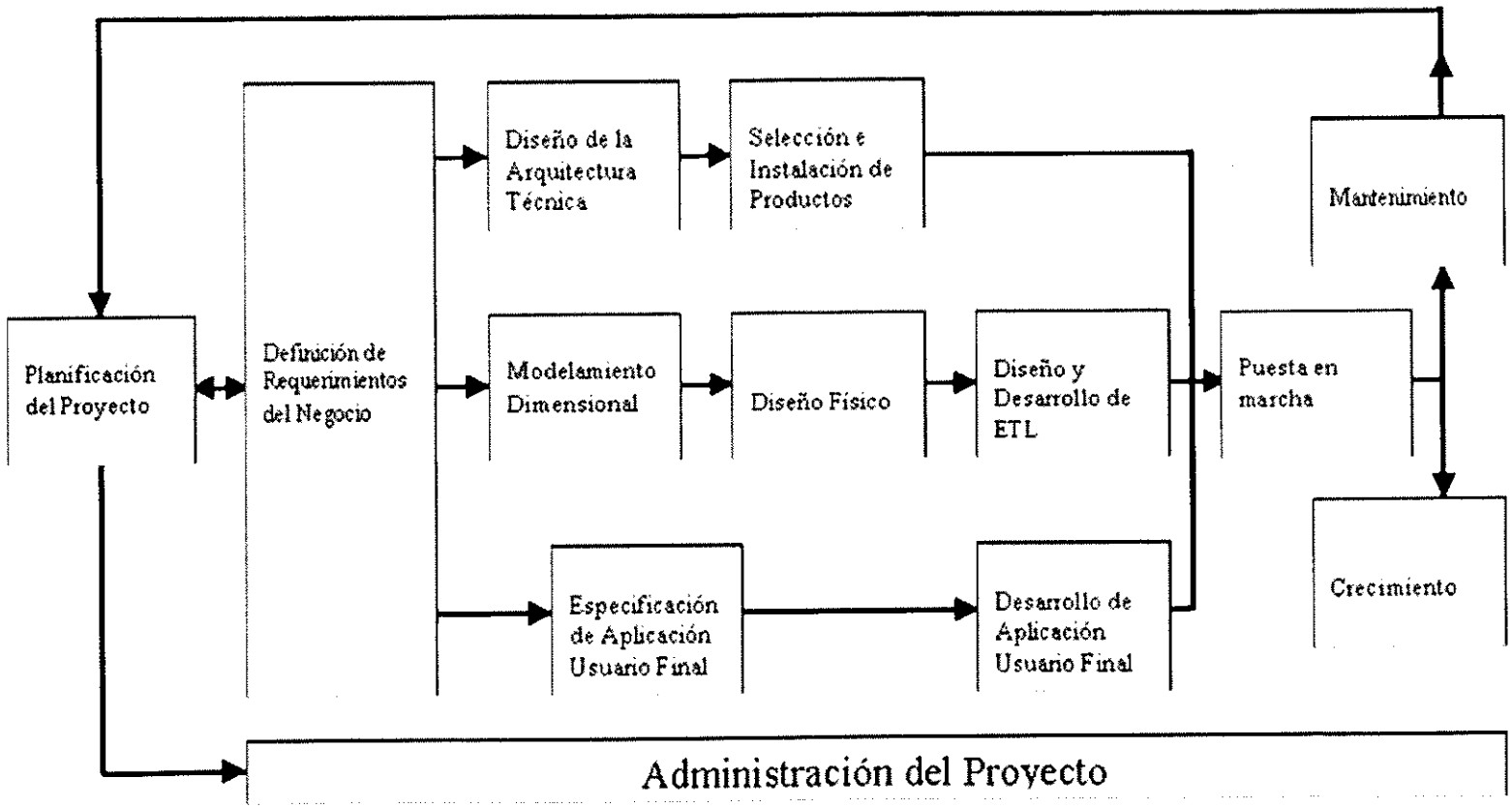


**Figura 3.2**: Fases de la metodología propuesta por Ralph Kimball

Las etapas que cubre esta metodología son las siguientes **[16]**:

**Planificación del Proyecto**

En esta etapa se direcciona la definición y alcance del proyecto incluyendo una evaluación de la preparación de la empresa para utilizar la tecnología de DM así como la justificación de negocios. Luego se enfoca en los recursos y niveles de conocimiento del personal requerido, junto con las asignaciones de tareas, duración y secuenciamiento. El plan resultante integra todas las tareas asociadas con el proyecto y documenta las personas o departamentos involucrados. La planificación depende de la definición de requerimientos como se ve por la flecha bidireccional entre los bloques.

**Definición de Requerimientos del Negocio**

El objetivo de esta etapa es entender las necesidades del negocio y determinar los requerimientos de los usuarios finales. La forma de obtener requerimientos de usuarios analíticos difiere de aquella usada para determinar los de los usuarios operacionales. Los diseñadores del DM deben entender los factores claves que dirigen el negocio para determinar efectivamente y traducirlos en consideraciones de diseño. Los requerimientos de negocio establecen la base para las tres rutas paralelas: datos, tecnología y aplicaciones de usuario final.

**Ruta de Datos: Modelamiento Dimensional**

La definición de los requerimientos de negocio determina los datos necesarios para atender a los usuarios analíticos. Se comienza construyendo una matriz que represente los procesos claves y su dimensionalidad. La matriz sirve como el

"blueprint" o plan para asegurar que el DM sea extensible a través de la organización en el tiempo. Luego se conduce un análisis más detallado de los sistemas operacionales que serán fuentes relevantes. Juntando este análisis con nuestro entendimiento de los requerimientos del negocio desarrollamos entonces el modelo dimensional, donde identificamos las tablas de datos granulares, los atributos dimensionales y la jerarquía de exploración ("drill"). Se completa el diseño lógico de la base de datos con las estructuras apropiadas de datos y las relaciones primarias y externas de claves. Igualmente se desarrolla el plan de agregaciones primario. Todo este conjunto de actividades se termina con el desarrollo del mapa de datos y su flujo desde la fuente hacia el objetivo.

**Ruta de Datos: Diseño Físico**

El diseño físico de la base de datos consiste en definir las estructuras físicas necesarias para soportar el diseño lógico; incluye definir los nombres estándares y establecer el ambiente de bases de datos así como la indexación preliminar y las estrategias de particionamiento.

**Ruta de Datos: Diseño y Desarrollo de la Preparación de los Datos ("Staging")**

El proceso de preparación de datos tiene tres pasos principales: extracción, transformación y carga. El proceso de extracción casi siempre muestra problemas de calidad en los datos, que han estado ocultos en los sistemas operacionales. Ya que la calidad de los datos impacta significativamente a la credibilidad del DM, hay que resolver aquí este problema, tanto para la carga inicial como para la carga regular incremental posterior.

**Ruta Tecnológica: Diseño de la Arquitectura Técnica**

Los ambientes de un DM requieren la integración de muchas tecnologías. El diseño de la arquitectura tecnológica establece el marco arquitectónico y la visión. Hay que considerar tres factores: los requerimientos de negocio, el ambiente técnico actual y las direcciones técnicas estratégicas planificadas.

**Ruta Tecnológica: Selección e Instalación de Productos**

Utilizando el diseño de la arquitectura técnica, como marco, se especifican los componentes estructurales tales como: plataforma de hardware, administrado de bases de datos, herramientas de ETL.

**Ruta de Aplicación: Especificación de la Aplicación del Usuario Final**

Se define un conjunto de aplicaciones estándares ya que no todos los usuarios requieren acceso ad hoc[1] al DM. Se describen los universos ("templates"), los parámetros a ser utilizados por los usuarios, los cálculos requeridos. Estas especificaciones aseguran que los usuarios y el grupo de desarrollo tengan un entendimiento común de las aplicaciones que se entregarán.

**Ruta de Aplicación: Desarrollo de la Aplicación del Usuario Final**

De acuerdo a las especificaciones de la aplicación se desarrollan las aplicaciones de usuario final, lo que involucra modificar la metadata y construir los informes especificados. Las aplicaciones y las plantillas de reportes pueden ser entregadas

[1] Ad hoc: implica que el sistema permite al usuario personalizar una consulta en tiempo real, en vez de estar atado a las consultas prediseñadas para informes.

bajo un ambiente Web, bajo un ambiente de una herramienta o como una aplicación personalizada.

**Puesta en Marcha**

La puesta en marcha representa la convergencia de tecnología, datos y aplicaciones de usuario final accesibles desde el desktop u otro mecanismo del usuario. Se requiere de una planificación detallada para asegurar que todas las piezas del rompecabezas se ajusten apropiadamente. Se capacita a los usuarios en todos los aspectos de la convergencia. Adicionalmente se define el soporte a los usuarios así como las comunicaciones y estrategias de retroalimentación antes de dar acceso a los usuarios al DM.

**Mantenimiento**

A partir de esta etapa se brinda soporte y capacitación continua a los usuarios, se pone foco en el "back room", asegurándose que los procesos y procedimientos garanticen una operación efectiva del DM.

**Crecimiento**

Se establecen los procesos de priorización para tratar con las demandas de los usuarios para evolución y crecimiento. Una vez que se establecen las prioridades se vuelve al comienzo del ciclo.

**Administración del Proyecto**

La administración del proyecto asegura que las actividades del ciclo de vida dimensional del negocio sigan por un buen camino y en sincronización. Estas

actividades se enfocan en monitorear el estado del proyecto, rastreo de problemas y cambia el control para preservar el límite del alcance. Finalmente, la administración del proyecto incluye el desarrollo de un plan de comunicación comprensible que se direcciona o enfoca hacia el negocio y hacia la información de los sistemas organizacionales.

### 3.3. Sakhr Youness

Sakhr Youness está convencido de que la fase de diseño de cualquier proyecto (aplicable a cualquier disciplina) es la más importante. Cuando se diseña un proyecto de software es esencial empezar de manera correcta, aún si esto significa usar más tiempo y recursos de los que se creían eran necesarios. A un mejor diseño se puede conseguir un mejor producto. Esto se debe a que un buen diseño minimiza el número de iteraciones que el proyecto va a sufrir durante su ciclo de vida o de desarrollo, logrando de esta manera un significativo ahorro o reducción de tiempo y recursos.

El siguiente gráfico representa la metodología de Youness:

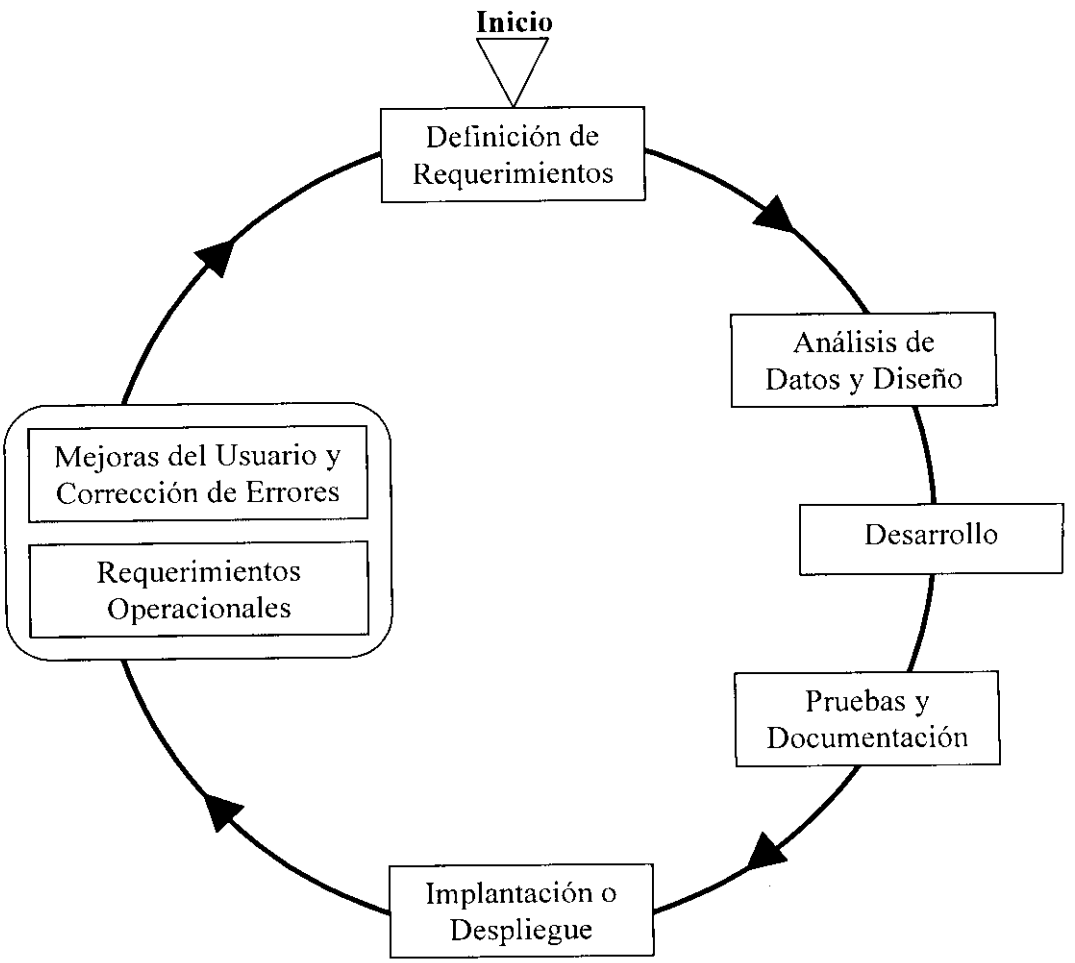


**Figura 3.3:** Fases de la metodología propuesta por Sakhr Youness

La metodología sugerida por Youness considera los siguientes pasos [17]:

**Definición de Requerimientos**

Constituye el primer paso de este ciclo de desarrollo, es el resultado de entrevistas, observaciones y estudios previos que define y recoge las necesidades o requerimientos del negocio.

**Análisis de Datos y Diseño**

Con los requerimientos del proyecto ya en mano, los prototipos pueden ser desarrollados para seleccionar el mejor diseño; esto permite conocer los datos con que contamos, los formatos que deseamos y los que vamos a usar. Se diseñan las diferentes estructuras de datos al igual que las herramientas que nos facilitarán el acceso de los mismos.

**Desarrollo**

Durante esta etapa se procede con la construcción de la o las bases de datos que se va a usar y de las herramientas de software que permitirán el acceso a las mismas.

**Pruebas y Documentación**

Conjuntamente con la etapa de desarrollo, se pueden realizar los pasos de pruebas y de documentación que son realizados en gran medida por los usuarios o clientes que igualmente forman parte de la construcción.

**Implantación o Despliegue**

Una vez que los usuarios han probado las aplicaciones y se han corregido ciertos problemas, se implanta o libera el proyecto de DW para completo uso de los clientes. Pero esto no significa que ya ha sido terminado completamente el proyecto.

Posterior a la liberación del proyecto es posible que se presenten o planteen varios problemas relacionados con el mantenimiento o con el aparecimiento de nuevos

**Requerimientos Operacionales.**

Aparecen además nuevas **Mejoras o necesidades** sugeridas **por los Usuarios** y también la necesidad nuevamente de realizar una **Corrección de Errores**. Estas etapas complementarias pueden exigir la modificación del proyecto y dar lugar a nuevos análisis y desarrollo de las nuevas funcionalidades deseadas, o de otra forma se puede realizar la fijación de errores durante los primeros días de funcionamiento en tiempo real.

## 3.4. W. H. Inmon

El autor de esta metodología manifiesta que una de las más grandes diferencias entre los sistemas operacionales y el DW se encuentra en la manera de construirlos. Considera que el ciclo de vida para el desarrollo del DW es exactamente contrario al ciclo de vida clásico para el desarrollo de sistemas.

Cuando se utiliza el ciclo de vida clásico para el desarrollo de sistemas, se tienen las siguientes etapas: recolección de requerimientos, análisis, diseño, programación, pruebas, integración e implementación. Este ciclo se maneja en base a los requerimientos, para construir un sistema es necesario primero comprender los requerimientos de los usuarios y luego continuar con las etapas de diseño y desarrollo.

El ciclo de vida para el desarrollo del DW (conocido también como ciclo de vida del desarrollo de sistemas hacia atrás, o ciclo reverso), funciona de manera diferente. Comienza con los datos, con éstos en mano son integrados, luego

examinados para determinar que pueden aportar y que defectos pueden tener. Si se decide continuar con el desarrollo, se codifican los programas para la manipulación de datos; los resultados obtenidos por estos programas son analizados, y finalmente, los requerimientos del sistema son comprendidos. Como se puede apreciar, este ciclo de desarrollo se basa en el manejo de los datos. Entonces se tendría las siguientes etapas: recolección de datos, integración, pruebas, codificación de programas, análisis de resultados y comprensión de requerimientos.

Inmon indica que en esta metodología todas las actividades que se realicen deben ser documentadas, el gráfico que a continuación se muestra detalla las actividades de esta metodología. Es necesario indicar que varias de ellas son iterativas [7].

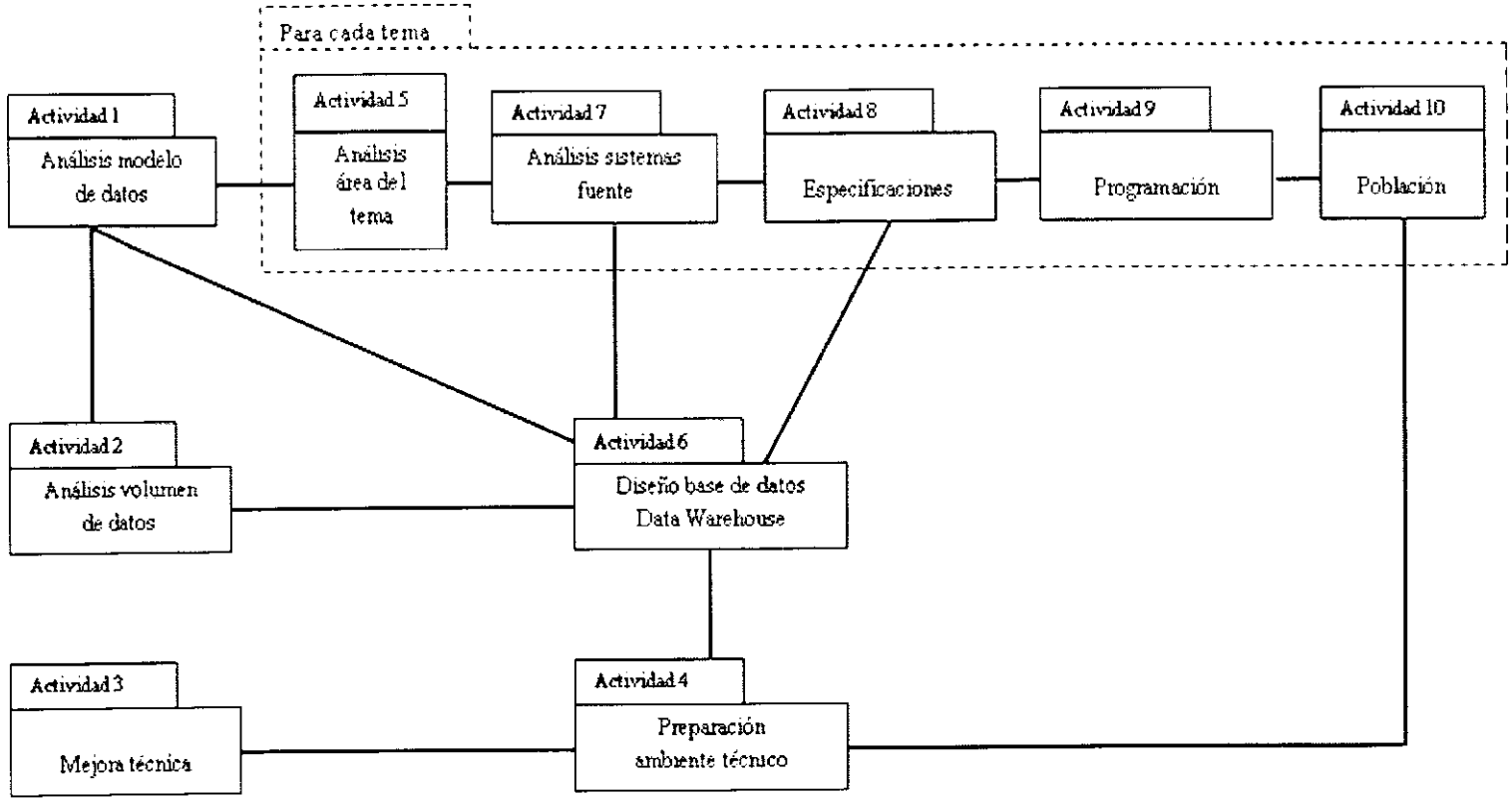


**Figura 3.4:** Fases de la metodología propuesta por W. Inmon

**Actividad 1: Análisis del modelo de datos**

Como actividad precedente: el compromiso de construir un DW; como siguiente actividad: 5, 2 y 6.

En un principio, el modelo de datos necesita estar bien definido y debe tener:

- Identificados todos los temas principales y las áreas bajo los cuales se desenvuelven.
- Claramente definidos los límites del modelo.
- Separados los datos primitivos de los derivados.
- Lo siguiente para cada área: claves, atributos, grupos de atributos, relaciones entre grupos de atributos, ocurrencia múltiple de datos, tipos de datos.

A la salida de esta etapa se debe confirmar que la organización disponga de un modelo de datos sólido, caso contrario no se puede continuar con el desarrollo del proyecto, mientras el problema no sea solucionado.

**Actividad 2: Análisis del volumen de datos**

Como actividad precedente: 1; como siguiente actividad: 6.

Una vez que el modelo de datos ha sido analizado y elevado a un nivel suficiente de calidad se continúa con el análisis de datos, que en breves términos es conocer que cantidad de información puede sostener el DW.

Como resultado de esto: si el DW va a contener gran cantidad de información se hace necesario considerar múltiples niveles de granularidad[2]; caso contrario no se necesita un plan de diseño para múltiples niveles de granularidad.

**Actividad 3: Mejora técnica**

Como actividad precedente: el compromiso de construir un DW; como siguiente actividad: 4.

Satisface los siguientes criterios:

- Habilidad para organizar los datos acorde al modelo de datos.
- Habilidad para administrar gran cantidad de información.
- Habilidad para acceder a la información de manera más flexible.
- Habilidad para poder cargar periódicamente información en masa.
- Habilidad para acceder o guardar datos de manera inmediata.
- Habilidad para enviar y recibir información a una amplia variedad de tecnologías.

**Actividad 4: Preparación del ambiente técnico**

Como actividad precedente: 3; como siguiente actividad: 6 y 10.

Una vez que la configuración de la arquitectura para el DW ha sido establecida, el siguiente paso es identificar técnicamente cómo dicha configuración puede ser acomodada. Varios de los típicos o cuestiones que aparecen en este punto son:

[2] 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. http://es.wikipedia.org/wiki/Granularidad

- La cantidad de dispositivos de almacenamiento directo son necesarios.
- El tipo de enlace que se va a tener a través de la red y dentro de la misma.
- Minimizar y/o aliviar los conflictos de procesamiento entre los programas que compiten el acceso a la información.
- El volumen y la naturaleza de tráfico que puede ser generado por las tecnologías que controlan al DW.

**Actividad 5: Análisis del área del tema**

Como actividad precedente: 1; como siguiente actividad: 7.

En este punto el área del tema (o departamento de la empresa por ejemplo) que va a ser poblado es seleccionado. El área puede ser significativamente extensa o pequeña para ser implementada; si debido a algún cambio dicha área es verdaderamente grande y compleja, se puede seleccionar un subconjunto para la implementación. Como resultado de este paso se tiene el alcance de esfuerzo en términos del tema.

**Actividad 6: Diseño de la base de datos del DW**

Como actividad precedente: 1, 7 y 2; como siguiente actividad: 8.

El DW (la base de datos) es diseñado en base del modelo de datos. Algunas características del último diseño incluyen:

- Un acomodamiento de los diferentes niveles de granularidad (si los hay).
- Una orientación de los datos a los temas principales de la corporación.
- La presencia únicamente de datos primitivos.
- La ausencia de la información que no vaya a servir para la toma de decisiones.

- El tiempo variante de cada registro de datos.
- La desnormalización física de los datos donde sea aplicable (donde el rendimiento lo garantice).

Como resultado de este paso se diseña la base de datos del DW, aunque aún no son cubiertas en detalle todas sus necesidades. Es aceptable diseñar las principales estructuras del DW inicialmente y luego de poco tiempo completar los detalles.

**Actividad 7: Análisis de sistemas fuente**

Como actividad precedente: 5; como siguiente actividad: 8 y 6.

Una vez que el tema o área a ser poblado es identificado, la siguiente actividad es identificar la (o las) fuente (s) de datos para cada tema en los ambientes de los sistemas existentes, resulta normal disponer una variedad de fuentes u orígenes de datos para los sistemas de soporte a la decisión; en este punto se presentan inconvenientes en la integración de los datos. Las siguientes son las consideraciones hacia donde va dirigida esta etapa de la metodología:

- Las estructuras de las claves y su resolución cuando éstas son llevadas de los sistemas operacionales a los ambientes de ayuda a la decisión.
- Atribuciones
  - ¿Qué hacer cuando existen múltiples fuentes de datos a escoger?
  - ¿Qué hacer cuando no existen múltiples fuentes de datos a escoger?
  - ¿Qué transformaciones, codificaciones, decodificaciones, conversiones, etc. tienen que hacerse como datos seleccionados para llevarlos a los ambientes soporte de decisiones?

- ¿Cómo se puede estructurar un sistema de soporte a la toma de decisiones a partir de estructuras operacionales?
- ¿De qué manera las relaciones operacionales aparecerán en el sistema de toma de decisiones?

Como resultado de esta etapa se tiene un mapeado de datos que va desde los ambientes operacionales al sistema de toma de decisiones.

**Actividad 8: Especificaciones**

Como actividad precedente: 7 y 6; como siguiente actividad: 9.

Una vez que las interfaces entre los sistemas operacionales y el sistema de soporte a las decisiones han sido delineadas, el siguiente paso es formalizarlos en términos de programas. Algunos de los problemas que se pueden dar son los siguientes:

- ¿Cómo puedo conocer qué información operacional examinar?
  - ¿Existe un archivo delta?
  - ¿Existen sistemas de auditoría que se puedan usar?
- ¿Cómo se puede guardar los resultados una vez examinados?
  - ¿Está el sistema de soporte de decisiones prellamado y preformateado?
  - ¿Es la información añadida?
  - ¿Es la información reemplazada?
  - ¿Son hechas actualizaciones en el ambiente de decisión?

El resultado de este paso es el programa que lleva los datos del ambiente operacional al DW.

**Actividad 9: Programación**

Como actividad precedente: 8; como siguiente actividad: 10.

Este paso incluye todas las actividades estándar de programación como son:

- Desarrollo de pseudocódigo.
- Codificación.
- Compilación.
- Pruebas.

**Actividad 10: Población o alimentación**

Como actividad precedente: 9 y 4; como siguiente actividad: Uso del DW.

Este paso trae consigo nada más que la ejecución del sistema para el soporte de toma de decisiones desarrollado. Ciertos inconvenientes se encaminan a:

- La frecuencia de la alimentación o poblado de los datos.
- La depuración de los datos ingresados.
- El envejecimiento de los datos alimentados (corriendo programas marcadores).
- Administrar múltiples niveles de granularidad, etc.

Como resultado de este paso se tiene un Data Warehouse funcional y con datos.

# CAPÍTULO IV

## 4. CASO DE ESTUDIO: DESARROLLO DE UN SISTEMA DE APOYO A LA TOMA DE DECISIONES QUE GESTIONE LA INFORMACIÓN DE LA PRUEBA DE APTITUD ACADÉMICA DE LA UNIVERSIDAD TÉCNICA DE AMBATO

En la actualidad resulta vital para las organizaciones, tener acceso oportuno a la información que facilite la toma de decisiones, es por esto, que los sistemas de información[1] se convierten en una pieza clave en la competitividad de las organizaciones.

La Universidad Técnica de Ambato no se encuentra ajena a este panorama, en el que la información juega un rol importante para la adecuada gestión de la Institución. Estos hechos han significado una reacción por parte de las autoridades, respecto a la creación de nuevos sistemas de información y de gestión que permitan mejorar la administración académica.

Resulta interesante indicar en este apartado que todo aspirante que desee ingresar a la Universidad Técnica de Ambato en calidad de estudiantes, tiene que de manera obligatoria rendir y aprobar la denominada Prueba de Aptitud Académica (PAA). La mencionada prueba tendrá efecto en las fechas y el horario establecido por el

[1] Componentes interrelacionados para reunir, procesar, almacenar y distribuir información para apoyar la toma de decisiones, la coordinación, el control, el análisis y la visualización de una organización [1].

cronograma académico. Para ello se hace uso de un sistema que funciona en la intranet de la Universidad bajo un entorno de trabajo tipo Web. La PAA evalúa a los aspirantes acerca de tópicos como Matemática, Aptitud Verbal, Técnicas de Estudio y de Especialidad. Tiene un total de 75 preguntas de selección que se generan de manera aleatoria e irrepetible de un macro banco de preguntas, lo que garantiza que no hay pruebas totalmente iguales para más de un aspirante. Dicha prueba debe ser resuelta en un tiempo no mayor a los 90 minutos.

El puntaje de aprobación depende de la carrera en la que se ha aplicado y de la cantidad de aspirantes que la han rendido. Para ello es el propio sistema que califica y selecciona a los aspirantes que pueden matricularse en primer semestre de la Universidad basado en la nota de referencia que se tiene para cada un de las diferentes carreras.

Es por esto que el presente trabajo propone la creación de un DM que gestione la información obtenida de la Prueba de Aptitud Académica de la Universidad Técnica de Ambato. Dicha herramienta cuenta con el apoyo de las autoridades así como del área de Desarrollo de la Dirección de Sistemas Informáticos y Redes de Comunicación (DISIR) de la Universidad Técnica de Ambato que es una unidad de apoyo administrativo, encargada de administrar los sistemas informáticos y redes de comunicación de la Universidad Técnica de Ambato, así como también, la de capacitar a la comunidad universitaria y a la colectividad de acuerdo a las necesidades planteadas por ellos.

El desarrollo del prototipo se va a basar en la Metodología sugerida por Sakhr Youness, ya que esta puede ser aplicada para pequeñas y grandes tareas de Data Warehousing. Su autor sugiere el uso de herramientas proporcionadas por Microsoft en lo referente a las bases de datos, herramientas de ETL y aplicaciones de explotación de la información.

Haciendo una analogía de las etapas de la Metodología de Sakhr Youness, tengo lo siguiente: la etapa denominada Definición de Requerimientos yo le llamo Requerimientos y va a ser el resultado de varias entrevistas.

La etapa de Análisis de Datos y Diseño la nombro como Fuentes de Información, en la que escojo la información con la que se va a trabajar. En lo referente al Diseño se va a realizar el Modelo Multidimensional del Data Mart de la Prueba de Aptitud Académica de la Universidad Técnica de Ambato.

En lo referente a la etapa de Desarrollo, la he dividido en dos partes: la carga de las tablas de Dimensiones y la carga de la tabla de Hechos. Para ello se va hacer uso de las herramientas de ETL de SQL Server 2005.

Las Pruebas y Despliegue del prototipo se las puede hacer mediante herramientas de explotación de información de SQL Server 2005 o de otras herramientas de mayor difusión como es el caso de Microsoft Excel en nuestro caso constituye la generación de informes.

La etapa final propuesta por Youness de Requerimientos Operacionales, puede constituirse como una mejora continua del prototipo propuesto.

## 4.1. Generalidades

Según la base legal: "la Universidad Técnica de Ambato es una Institución de Educación Superior, de derecho público, con domicilio principal en la ciudad de Ambato, provincia del Tungurahua, creada mediante Ley N°. 69-05 del 18 de abril de 1969. Se rige por la Constitución y Leyes de la República del Ecuador, la Ley de Educación Superior, el Reglamento General a la Ley de Educación Superior, el Reglamento General del Sistema Nacional de Evaluación y Acreditación, los Reglamentos del CONESUP y del CONEA, este Estatuto y sus Reglamentos, Guía de Auditoria para Universidades y Escuelas Politécnicas y las disposiciones que adopten sus organismos y las autoridades universitarias, en el ámbito de su competencia" **[20]**.

El Gobierno de la Universidad Técnica de Ambato, será ejercido en orden jerárquico, por los siguientes Organismos y Autoridades:

- Honorable Consejo Universitario;
- Rector/a;
- Vicerrectores/as Académico y Administrativo;
- Consejos Directivos de Facultades;
- Decanos/as de Facultades;
- Subdecanos/as de Facultades;
- Coordinadores/as de Carreras; y,

- ❖ Coordinadores/as de Departamentos.

Para cumplir con los objetivos propuestos, la Universidad Técnica de Ambato se conforma de Unidades Académicas, que son las Facultades u otro tipo de unidades que se crearen, con Carreras de Pregrado y con Programas de Posgrado que cumplen las funciones básicas de la Universidad: Docencia o Formación de Profesionales, Investigación, Extensión; apoyadas por Departamentos Especializados ya existentes y con los que se establecieren en lo posterior.

La Universidad Técnica de Ambato es una comunidad académica, constituida por Docentes, Estudiantes, Empleados/as y Trabajadores/as. Estructural y funcionalmente se organiza en los siguientes niveles: Directivo, Ejecutivo, Asesor, Apoyo Administrativo, Apoyo Académico y Operativo **[20]**.

Actualmente la Universidad Técnica de Ambato tiene la cantidad estimada de 14000 estudiantes, repartidos en: 9 facultades, aproximadamente 42 carreras de pregrado (entre la modalidad presencial y semipresencial) y alrededor de 20 programas de posgrado.

El siguiente gráfico muestra el organigrama estructural de la Universidad Técnica de Ambato:

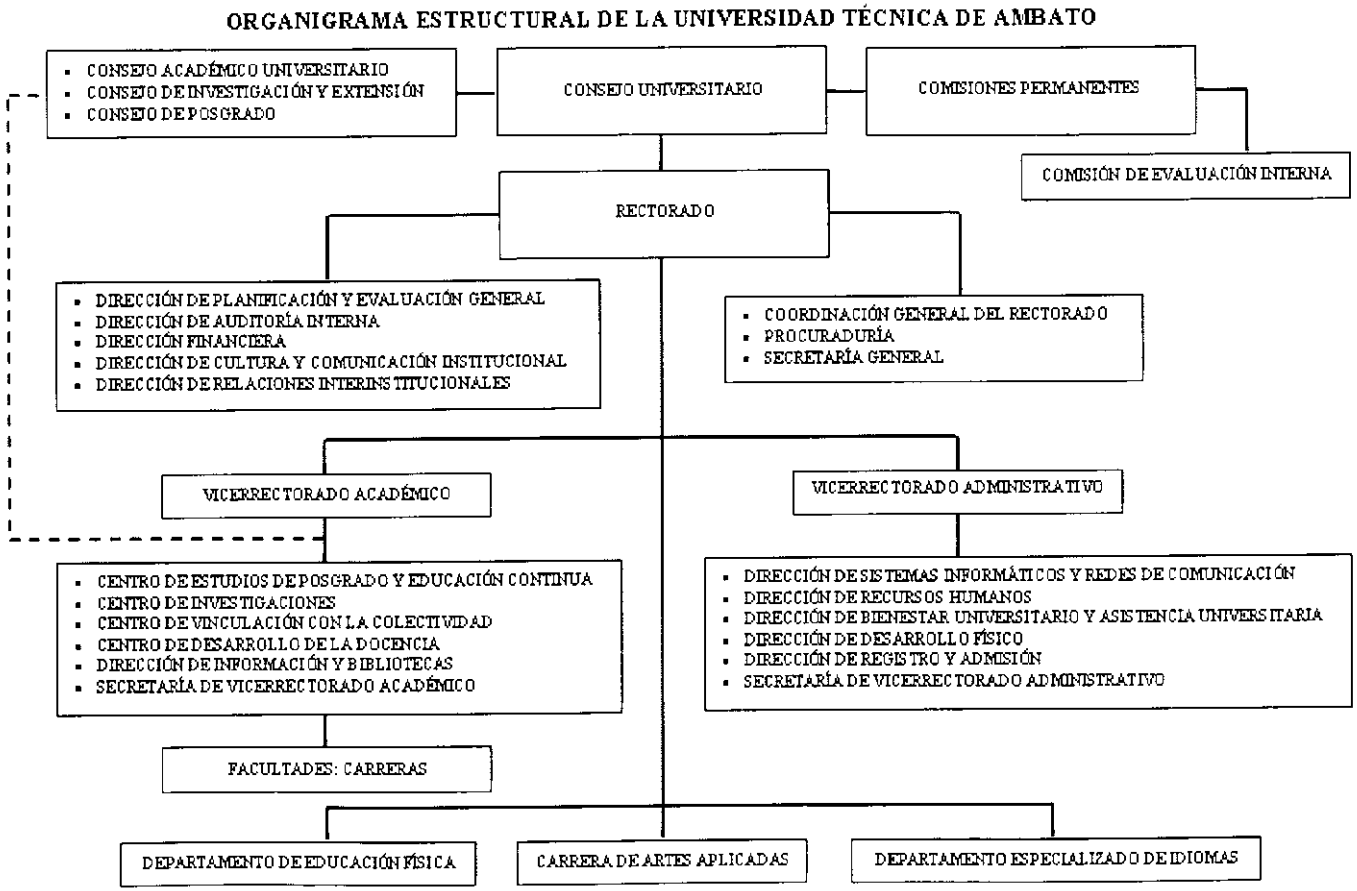


**Figura 4.1:** Organigrama Estructural de la Universidad Técnica de Ambato

## 4.2. Usuarios considerados

A nivel directivo se han considerado como usuarios y beneficiarios del presente proyecto a las principales autoridades de la Universidad Técnica de Ambato, encabezadas por el Ing. Luis Amoroso en calidad de Rector Rector; el Dr. Galo Naranjo como Vicerrector Académico y el Dr. Remigio Medina como Vicerrector Administrativo.

El aspecto operativo y de desarrollo se ve apoyado por el área de Desarrollo de la Dirección de Sistemas Informáticos y Redes de Comunicación (DISIR) cuyo director es el Ing. Edison Alvarez.

Otros organismos propios de la Universidad Técnica de Ambato que también serán beneficiados con este proyecto son: la Dirección de Planificación y Evaluación General, la Dirección de Registro y Admisión, el Centro de Investigaciones y el Centro de Desarrollo de la Docencia.

Finalmente a nivel externo como usuarios tenemos a los directivos y autoridades de los diferentes colegios del país, los aspirantes que rinden la Prueba de Aptitud Académica y la colectividad en general que verán satisfecha su demanda de información.

## 4.3. Requerimientos

Esta etapa está orientada a conocer las actividades y objetivos de las diferentes áreas en estudio, de manera que se pueda captar las necesidades de información para el DM.

Para determinar estos requerimientos se establecen entrevistas, se estudian documentos e informes existentes y se revisan los modelos de datos de los sistemas de información existentes.

Se trata de presentar reportes que permitan realizar estadísticas de los aspirantes que pretenden ingresar a la Institución. Dichos reportes requieren de un gran esfuerzo para su desarrollo ya que sus cálculos se los hacen en forma manual, lo que ocasiona el retrazo de otras tareas y no existe total garantía en la calidad de la información proporcionada.

Es así que los informes requeridos de acuerdo a las necesidades de los diferentes usuarios, ayudarán a conocer información como:

- Cantidad de aspirantes inscritos a la Prueba de Aptitud Académica.
- Cantidad de aspirantes que rindieron la Prueba de Aptitud Académica.
- Cantidad de aspirantes que no rindieron la Prueba de Aptitud Académica.
- Cantidad de aspirantes que rindieron la Prueba de Aptitud Académica según la Carrera.
- Cantidad de aspirantes que rindieron la Prueba de Aptitud Académica según la Facultad.

- Cantidad de aspirantes que rindieron la Prueba de Aptitud Académica según la modalidad (Presencial o Semipresencial) de estudio.
- Cantidad de aspirantes que rindieron la Prueba de Aptitud Académica según el tipo de colegio del que provienen (Fiscal o Particular).
- Cantidad de aspirantes que rindieron la Prueba de Aptitud Académica según el colegio del que provienen.
- Cantidad de aspirantes que rindieron la Prueba de Aptitud Académica y que aprobaron o reprobaron.
- Cantidad de aspirantes que rindieron la Prueba de Aptitud Académica según la ubicación del colegio del que provienen (país, provincia o cantón).
- Cantidad de aspirantes que rindieron la Prueba de Aptitud Académica en una fecha determinada.
- Los puntajes obtenidas por los aspirantes (totales y por área).
- La cantidad de preguntas correctas, incorrectas y no contestadas por los aspirantes.
- Los puntajes máximos y mínimos obtenidos.
- Las Facultades y Carreras con mayor y menor cantidad de aspirantes.

En lo referente a las herramientas de software que se van a utilizar para el desarrollo del prototipo y aprovechando el convenio que tiene la Universidad Técnica de Ambato con Microsoft, se utilizará SQL Server 2005 con sus aplicaciones SQL Server Management Studio y SQL Server Business Intelligence Development Studio.

Como parte de SQL Server Business Intelligence Development Studio también se necesita usar Microsoft SQL Server 2005 Integration Services (SSIS) que es una plataforma que permite generar soluciones de integración de datos de alto rendimiento, entre las que se incluyen paquetes de extracción, transformación y carga (ETL) para el almacenamiento de datos. Integration Services incluye herramientas gráficas y asistentes para generar y depurar paquetes, tareas para realizar funciones de flujo de trabajo, como las operaciones de FTP, tareas para ejecutar instrucciones SQL o para enviar mensajes de correo electrónico, orígenes y destinos de datos para extraer y cargar datos, transformaciones para limpiar, agregar, mezclar y copiar datos, un servicio de administración, entre otras actividades. Integration Services reemplaza a los Servicios de Transformación de Datos (DTS) que se incluyó por primera vez como componente de SQL Server 7.0 **[21]**.

También se usará Microsoft SQL Server 2005 Analysis Services (SSAS) que proporciona funciones de procesamiento analítico en línea (OLAP) y minería de datos para soluciones Business Intelligence. Analysis Services combina los mejores aspectos del análisis tradicional basado en OLAP y la elaboración de informes basada en relaciones al permitir a los programadores definir un único modelo de datos, denominado Unified Dimensional Model (UDM), a partir de uno o más orígenes de datos físicos. Todas las consultas de usuario final desde aplicaciones OLAP, de elaboración de informes y de Business Intelligence personalizadas obtienen acceso a los orígenes de datos subyacentes a través del modelo UDM, que proporciona una única vista empresarial de estos datos relacionales. Analysis Services proporciona un amplio conjunto de algoritmos de minería de datos para

permitir a los usuarios empresariales recopilar los datos mediante la búsqueda de patrones y tendencias específicos. Estos algoritmos de minería de datos se pueden utilizar para analizar los datos a través de un modelo UDM o directamente a partir de un almacén de datos físico **[22]**.

Por razones de sencillez y familiaridad, para la visualización de informes y reportes se puede hacer uso de Microsoft Excel.

## 4.4. Fuentes de información

Las fuentes de información corresponden a los sistemas transaccionales OLTP que actualmente se encuentran en uso en la Universidad.

Para este caso de estudio existen dos aplicaciones: el Sistema de Gestión Estudiantil UTAm@tico y el módulo de Inscripciones de la Prueba de Aptitud Académica del UTAm@tico. Además forma parte de la primera aplicación la base de datos BD_UTAMATICO y como parte de la segunda aplicación se tiene la base de datos BD_EVALUACION. Ambas bases de datos son del tipo SQL Server 2005 y por razones obvias de seguridad y de confidencialidad no se pueden mostrar los diagramas detallados de estas dos bases de datos.

Pero se debe indicar que la base de datos BD_UTAMATICO tiene entre otras las siguientes tablas: UTA_CANTONES, UTA_COLEGIOS, UTA_CURSOS, UTA_CARRERAS,UTA_DOCENTES,UTA_ESTUDIANTES,UTA_FACULTADES, UTA_MATERIAS, UTA_PAISES, UTA_PROVINCIAS etc.

Por otra parte la base de datos BD_EVALUACION posee tablas consideradas como principales a las siguientes: EVA_ARANCELES, EVA_ASPIRANTES, EVA_ORDENES, EVA_NOTAS, EVA_PRUEBAS, etc.

Con la finalidad de obtener una mejor comprensión de las fuentes de datos que se van a utilizar, así como evitar ciertas complicaciones de rendimiento al momento de la extracción de la información, de su transformación y llamado, se desarrollaron varias vistas de las bases de datos que se van a utilizar. Dichas vistas contienen la información necesaria para nuestro prototipo y que son obtenidas de las bases de datos antes mencionadas.

Así por ejemplo se tiene la vista vwPAAFacultadesCarreras que me va a proporcionar información como la que se puede visualizar en la tabla a continuación detallada:

| NOMBRE | TIPO | DESCRIPCIÒN |
|---|---|---|
| CodigoFacultad | Varchar | Código de las diferentes facultades |
| NombreFacultad | Varchar | Nombre de las facultades existentes |
| CodigoCarrera | Varchar | Código de las diferentes carreras |
| NombreCarrera | Varchar | Nombre de las carreras existentes |
| NotaReferencia | Tinyint | Nota mínima de aprobación de la PAA para cada carrera |

**Tabla 4.1:** Estructura de la vista vwPAAFacultadesCarreras

La siguiente vista usada se denomina vwPAAColegiosCantones, y facilita información como:

| NOMBRE | TIPO | DESCRIPCIÒN |
|---|---|---|
| CodigoColegio | Varchar | Código de los diferentes colegios registrados |
| NombreColegio | Varchar | Nombre de los colegios |
| NombreCanton | Varchar | Nombre del cantón donde se encuentra un colegio |
| NombreProvincia | Varchar | Nombre de la provincia donde se encuentra un colegio |
| NombrePais | Varchar | Nombre del país donde se encuentra un colegio |
| TipoColegio | Varchar | El tipo de colegio muestra si es Fiscal o Particular |

**Tabla 4.2:** Estructura de la vista vwPAAColegiosCantones

Otra vista usada es vwAspirantes que proporciona como información lo que muestra la siguiente tabla:

| NOMBRE | TIPO | DESCRIPCIÒN |
|---|---|---|
| CEDULA | Varchar | Cédula o pasaporte de los aspirantes |
| NOMBRE | Varchar | Nombre producto de concatenar los apellidos y nombres de los diferentes aspirantes |

**Tabla 4.3:** Estructura de la vista vwAspirantes

También se usa la vista denominada vwHechos2 que incluye información como:

| NOMBRE | TIPO | DESCRIPCIÒN |
|---|---|---|
| CEDULA | Varchar | Cédula o pasaporte de los aspirantes |
| NOMBRE | Varchar | Nombre de los aspirantes, concatenado apellidos y nombres |
| FECHA_INICIO | Datetime | Fecha en que rindió la PAA un aspirante |
| NOTA | Tinyint | El total de las respuestas correctas contestadas |
| INCORRECTAS | Tinyint | El total de las respuestas incorrectas contestadas |
| MATEC | Tinyint | Respuestas correctas contestadas de matemáticas |

Continuación...

| | | |
|---|---|---|
| MATEI | Tinyint | Respuestas incorrectas contestadas de matemáticas |
| APVEC | Tinyint | Respuestas correctas contestadas de aptitud verbal |
| APVEI | Tinyint | Respuestas incorrectas contestadas de aptitud verbal |
| TECC | Tinyint | Respuestas correctas contestadas de técnicas de estúdio |
| TECI | Tinyint | Respuestas incorrectas contestadas de técnicas de estúdio |
| ESPC | Tinyint | Respuestas correctas contestadas de especialidad |
| ESPI | Tinyint | Respuestas incorrectas contestadas de especialidad |
| CODIGO_CARRERA | Varchar | Código de las diferentes carreras |
| NOMBRE_CARRERA | Varchar | Nombre de las carreras |
| CODIGO_FACULTAD | Varchar | Código de las facultades existentes |
| NOMBRE_FACULTAD | Varchar | Nombre de las facultades |
| CODIGO_COLEGIO | Varchar | Código de los colegios registrados |
| NOTA_REQUERIDA | Tinyint | Nota mínima de aprobación de la PAA para cada carrera |
| ESTADO_PRUEBA | Varchar | Indica si ha aprobado o reprobado la PAA |
| NOMBRE_COLEGIO | Varchar | Nombre del colegio |
| NOMBRE_CANTON | Varchar | Nombre del cantón donde se encuentra un colegio |
| NOMBRE_PROVINCIA | Varchar | Nombre de la provincia donde se encuentra el colegio |
| NOMBRE_PAIS | Varchar | Nombre del país donde se encuentra el colegio |
| TIPO_COLEGIO | Varchar | Tipo de colegio que indica si es fiscal o particular |

**Tabla 4.4:** Estructura de la vista vwHechos2

## 4.5. Modelo Multidimensional del Data Mart de la Prueba de Aptitud Académica de la Universidad Técnica de Ambato

Basado en los requerimientos establecidos y en los sistemas operacionales disponibles, se define el modelo en copo de nieve o multidimensional del DM de la PAA de la Universidad técnica de Ambato en el que se consideran como dimensiones: DIM_ASPIRANTE (información básica de los aspirantes), DIM_FACULTAD (información básica de las diferentes facultades), DIM_CARRERA (información básica de las carreras de la Universidad), DIM_TIEMPO (información sobre las fechas en que rindieron la PAA), DIM_PAA (información que registra si un aspirante ha aprobado o no la PAA), DIM_COLEGIO (información básica sobre los colegios registrados), DIM_TIPOCOLEGIO (indica si un colegio es fiscal o particular) y como tabla de hechos: HECHOS_PAA (almacena las medidas o valores a analizar, en nuestro caso las notas de los aspirantes).

Los campos de las tablas de dimensiones y de la tabla de hechos van a tener las siguientes características:

| DIM_ASPIRANTE | | |
|---|---|---|
| **NOMBRE** | **TIPO** | **DESCRIPCIÒN** |
| ASPIRANTEID | int | Clave sustituta (autonumérico) |
| ASPIRANTENK | varchar(10) | Clave natural u original |
| ASPIRANTENOMBRE | varchar(102) | Nombre del aspirante |

**Tabla 4.5:** Estructura de la dimensión DIM_ASPIRANTE

| DIM_FACULTAD | | |
|---|---|---|
| **NOMBRE** | **TIPO** | **DESCRIPCIÒN** |
| FACULTADID | int | Clave sustituta (autonumérico) |
| FACULTADNK | varchar(2) | Clave natural u original |
| FACULTADNOMBRE | varchar(40) | Nombre de la facultad |

**Tabla 4.6:** Estructura de la dimensión DIM_FACULTAD

| DIM_CARRERA | | |
|---|---|---|
| **NOMBRE** | **TIPO** | **DESCRIPCIÒN** |
| CARRERAID | int | Clave sustituta (autonumérico) |
| ESPECIALIDADNK | varchar(4) | Clave natural u original |
| FACULTADNK | varchar(2) | Clave original de facultad |
| ESPECIALIDADNOMBRE | varchar(50) | Nombre de la carrera |
| ESPECIALIDADMODALIDAD | varchar(15) | Presencial o Semipresencial |
| ESPECIALIDADREFERENTE | tinyint | Nota referencial a partir de la cual aprueban la PAA. Varía según la carrera |

**Tabla 4.7:** Estructura de la dimensión DIM_CARRERA

| DIM_TIEMPO | | |
|---|---|---|
| **NOMBRE** | **TIPO** | **DESCRIPCIÒN** |
| TIEMPOID | Int | Clave sustituta (autonumérico) |
| TIEMPOFECHA | datetime | Clave natural u original |
| TIEMPODIA | tinyint | Número de día en que rindió la PAA |
| TIEMPOMES | tinyint | Número de mes en que rindió la PAA |
| TIEMPOANIO | smallint | El año en que rindió la PAA |

**Tabla 4.8:** Estructura de la dimensión DIM_TIEMPO

| DIM_PAA | | |
|---|---|---|
| **NOMBRE** | **TIPO** | **DESCRIPCIÒN** |
| PAAID | varchar(1) | Clave sustituta |
| PAAESTADO | varchar(10) | Nombre del estado de la PAA (Aprobado o Reprobado) |

**Tabla 4.9:** Estructura de la dimensión DIM_PAA

| DIM_COLEGIO | | |
|---|---|---|
| **NOMBRE** | **TIPO** | **DESCRIPCIÒN** |
| COLEGIOID | int | Clave sustituta (autonumérico) |
| TIPOID | varchar(1) | Clave foránea |
| COLEGIONK | varchar(5) | Clave natural u original |
| COLEGIONOMBRE | varchar(32) | Nombre del colegio |
| COLEGIOPAIS | varchar(20) | Nombre del país donde se encuentra el colegio |
| COLEGIOPROVINCIA | varchar(20) | Nombre de la provincia donde se encuentra el colegio |
| COLEGIOCANTON | varchar(20) | Nombre del cantón donde se encuentra el colegio |

**Tabla 4.10:** Estructura de la dimensión DIM_COLEGIO

| DIM_TIPOCOLEGIO | | |
|---|---|---|
| **NOMBRE** | **TIPO** | **DESCRIPCIÒN** |
| TIPOID | varchar(1) | Clave sustituta |
| TIPONOMBRE | varchar(25) | Tipo de un colegio (puede ser Fiscal o A/D o Particular) |

**Tabla 4.11:** Estructura de la dimensión DIM_TIPOCOLEGIO

| HECHOS_PAA | | |
|---|---|---|
| **NOMBRE** | **TIPO** | **DESCRIPCIÒN** |
| ASPIRANTEID | Int | Clave sustituta |
| TIEMPOID | Int | Clave sustituta |
| FACULTADID | Int | Clave sustituta |

Continuación...

| | | |
|---|---|---|
| COLEGIOID | Int | Clave sustituta |
| PAAID | varchar(1) | Clave sustituta |
| CARRERAID | Int | Clave sustituta |
| ASPIRANTENK | varchar(10) | Clave foránea |
| ESPECIALIDADNK | varchar(4) | Clave foránea |
| CMATE | tinyint | Respuestas correctas de matemáticas |
| CAPVE | tinyint | Respuestas correctas de aptitud verbal |
| CTECE | tinyint | Respuestas correctas de técnicas de estudio |
| CESPE | tinyint | Respuestas correctas de especialidad |
| NOTA | tinyint | Total de respuestas correctas |
| IMATE | tinyint | Respuestas incorrectas de matemáticas |
| IAPVE | tinyint | Respuestas incorrectas de aptitud verbal |
| ITECE | tinyint | Respuestas incorrectas de técnicas de estudio |
| IESPE | tinyint | Respuestas incorrectas de especialidad |
| INCORRECTAS | tinyint | Total de respuestas incorrectas |
| VACIAS | tinyint | Número de preguntas sin contestar |

**Tabla 4.12:** Estructura de la tabla de hechos HECHOS_PAA

Basado en estos elementos, el modelo multidimensional de la base de datos PAA_MART, de nuestro estudio es como se muestra en la siguiente figura:

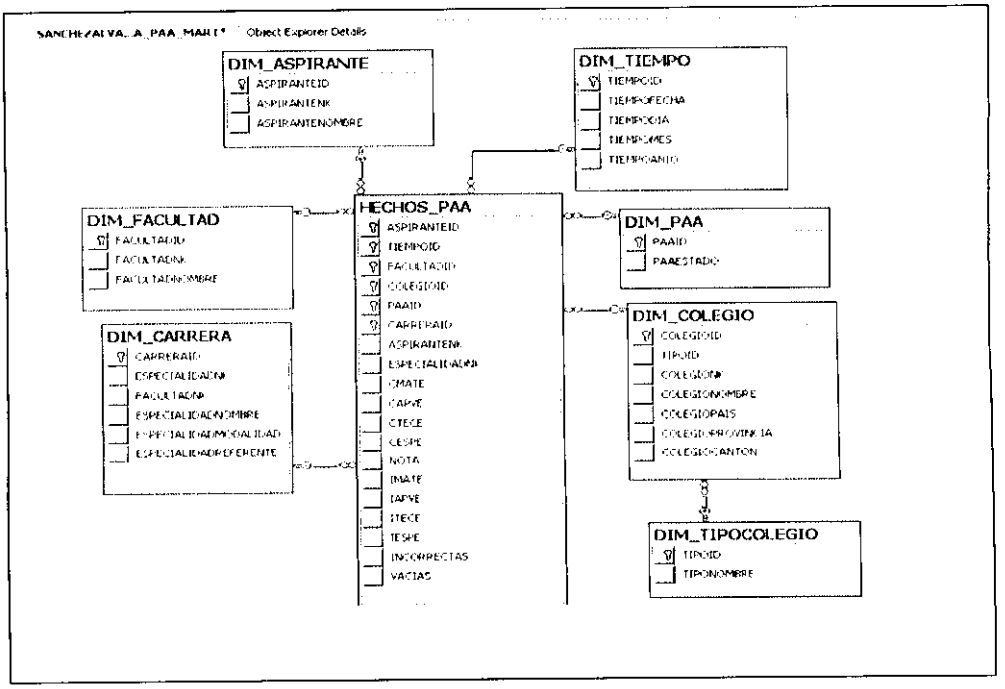


**Figura 4.2:** Modelo multidimensional de PAA_MART

Adicionalmente y por recomendación de varios autores de importantes libros como: W. H. Inmon y Ralph Kimball **[7][4]**, sugieren que los datos provenientes de los distintos orígenes o aplicaciones, no deben ser transferidos directamente al modelo multidimensional del DM. Indican que se debe tener un área de "preparación" o STAGE a la que se debe cargar la información deseada y sobre la cual es factible realizar aún varias purificaciones, modificaciones o transformaciones (en general las tareas de extracción y transformación). Una vez realizadas estas acciones se trasladará la información de las tablas de stage al Modelo Dimensional del DM (la tarea de carga propiamente dicha).

Las tablas de stage no deben estar relacionadas entre sí, no tienen claves y están formadas por los mismos campos y tipos de datos de las dimensiones y hechos a utilizar. Con el fin de tener cierto tipo de control las tablas de stage pueden incluir campos adicionales que no tienen las tablas de dimensiones y de hechos. Siguiendo el consejo de los expertos antes mencionados, nuestra área de stage incluirá las tablas que se muestran en la siguiente figura:

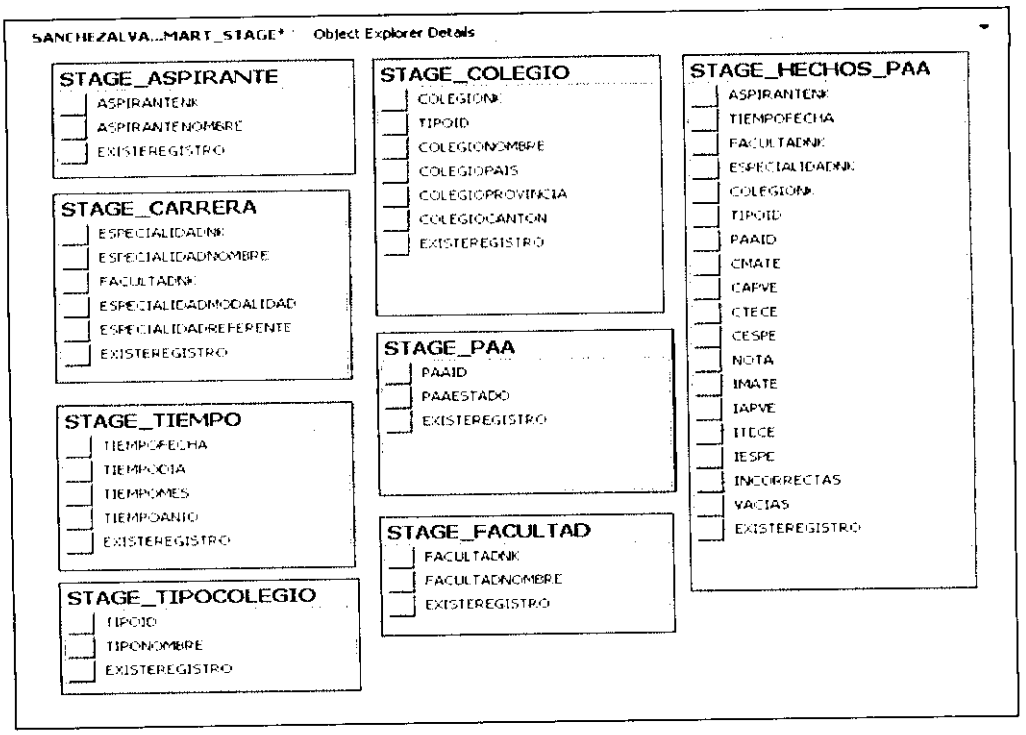


**Figura 4.3:** Tablas de Stage

Continuando con la etapa de desarrollo, en nuestro caso de estudio hay que proceder con el "poblamiento" del DM propuesto, esta tarea se la realiza en dos actividades que tienen un orden específico: la carga de las tablas de dimensiones y la carga de la tabla de hechos. Hay que recordar que las bases de datos de origen así como las de destino se encuentran en el SGBD SQL Server 2005 y que para la extracción, transformación y carga de la información se utiliza SQL Server 2005 Integration Services (SSIS).

Para realizar las diferentes actividades de ETL, se crea los denominados "paquetes". Un paquete es una colección organizada de conexiones, elementos de flujo de control, elementos de flujo de datos, controladores de eventos, variables y configuraciones que se pueden ensamblar con la ayuda de las herramientas gráficas

de diseño proporcionadas por SISS o mediante programación. El paquete es la unidad de trabajo que se recupera, ejecuta y guarda y está formado por un flujo de control y un flujo de datos **[23]**.

Un flujo de control consta de una o más tareas y contenedores que se ejecutan cuando se ejecuta el paquete.

Por otra parte un flujo de datos consta de los orígenes y destinos que extraen y cargan datos, las transformaciones que modifican y extienden datos, y las rutas que vinculan orígenes, transformaciones y destinos. Para poder agregar un flujo de datos a un paquete, el flujo de control de paquetes debe incluir una tarea Flujo de datos.

Esta tarea es el ejecutable del paquete SISS que crea, organiza y ejecuta el flujo de datos. Se abre una instancia independiente del motor de flujo de datos para cada tarea Flujo de datos de un paquete **[23]**.

El siguiente gráfico muestra un paquete individual que contiene un flujo de control con una tarea Flujo de datos que, a su vez, contiene un flujo de datos:

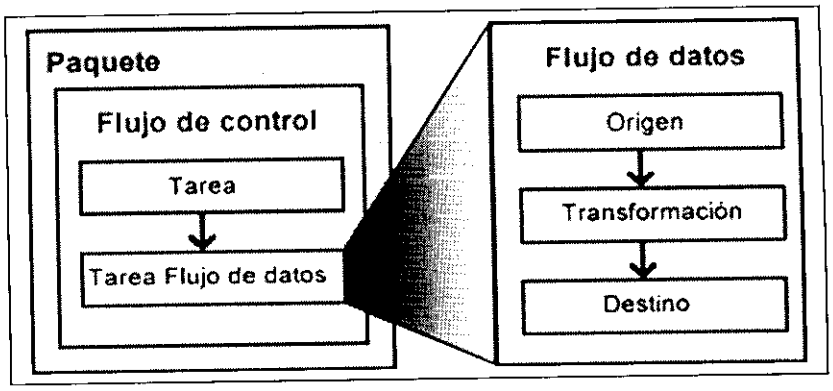


**Figura 4.4:** Paquete individual

Para el prototipo se crearon cuatro paquetes que cumplen tareas específicas. El primero (denominado ETL0), sirve como práctica y permite blanquear la tabla de hechos y las dimensiones del DM PAA_MART. El segundo (llamado ETL1), limpia las tabla de hechos y de las dimensiones del área de stage o preparación, adicionalmente realiza la carga de los valores iniciales de las tablas STAGE_TIPOCOLEGIO y STAGE_PAA. El tercer paquete (con nombre ETL2), facilita la carga de las tablas de stage con los datos provenientes de los diferentes orígenes. Finalmente el cuarto paquete (cuyo nombre es ETL3), una vez que realiza la limpieza y carga de todas las dimensiones, procede con la limpieza y carga de la tabla de hechos del DM.

Previo a la carga definitiva de las dimensiones y los hechos es necesario realizar varias actividades. Así por ejemplo, el paquete creado inicialmente y llamado ETL0, se encarga de blanquear la tabla de hechos y todas las tablas de dimensiones (en ese orden) del DM.

Para desarrollar esta tarea se incluyó en el flujo de control la herramienta denominada: "Execute SQL Task" que permite ejecutar una consulta SQL y se la configuró con las sentencias SQL que permiten la eliminación de la tabla de hechos HECHOS_PAA, así como de todas las tablas de dimensiones del DM.

La siguiente figura muestra como luce el paquete inicial ETL0:

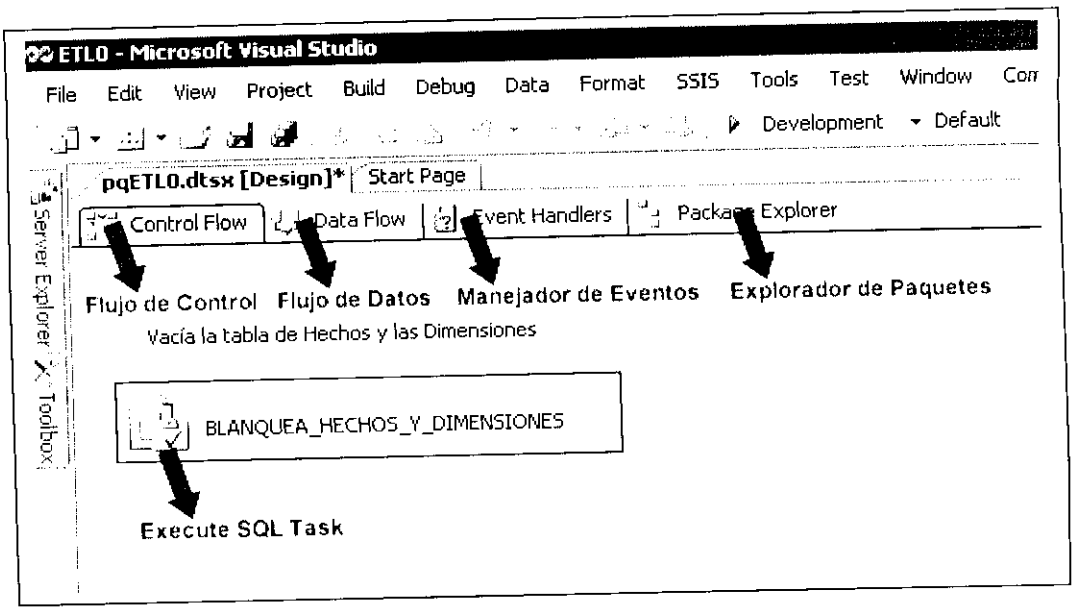


**Figura 4.5:** Paquete ETL0

Igualmente se tiene el paquete ETL1 que es utilizado para limpiar todas las tablas de stage. Una vez realizada esta acción se cargan con valores iniciales a las tablas STAGE_TIPOCOLEGIOS y STAGE_PAA.

Para esto en el flujo de control se incluyen tres "Execute SQL Task", cada una de ellas tienen como contenido sentencias SQL de borrado e inserción de datos. La siguiente figura muestra las actividades que se realizan en el flujo de control del paquete ETL1:

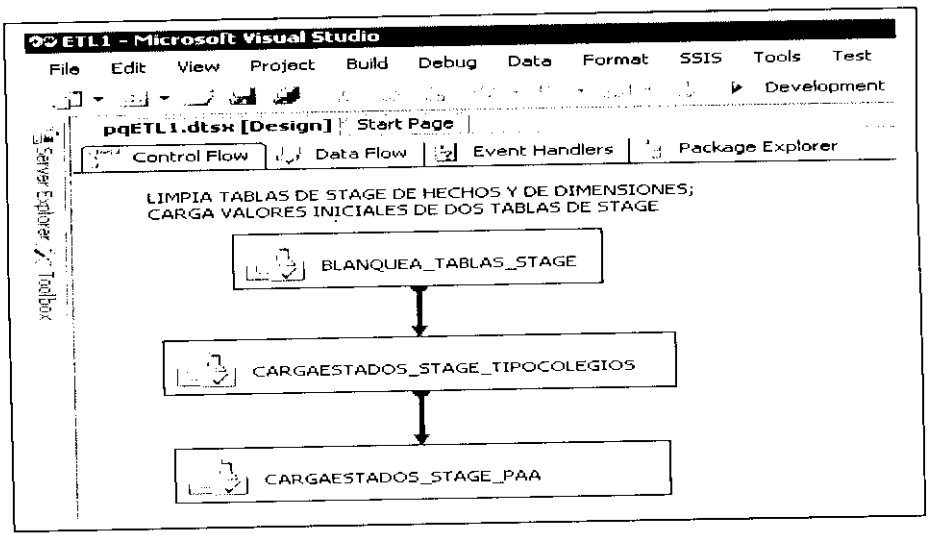


**Figura 4.6:** Paquete ETL1

A continuación se usa el paquete ETL2 que tiene en el flujo de control a dos "Data Flow Task" enlazados.

El primero contiene el flujo de datos que extrae, transforma y carga los datos hacia las tablas de stage correspondientes a las dimensiones.

El segundo hace igual trabajo, pero para la tabla de stage correspondiente a los hechos. Igualmente hay que recordar que el orden en que se ejecuta el flujo de control es importante. En este paquete se crearon dos variables locales denominadas fechaInicio y fechaFin, que permiten controlar como rangos las fechas en que se efectuaron la recepción de la PAA.

Esta es su apariencia gráfica:

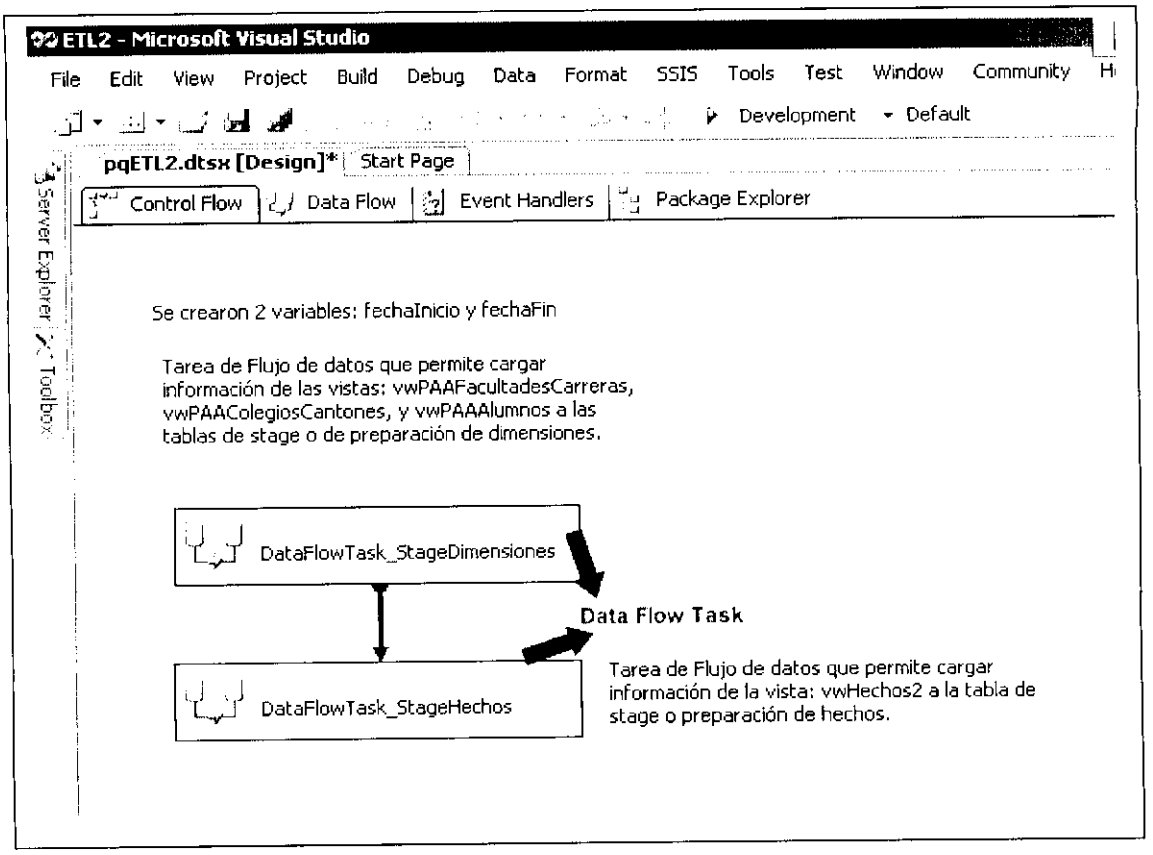


**Figura 4.7:** Paquete ETL2

Lo siguiente que viene es el detalle de DataFlowTask_StageDimensiones. Para accionarla en la tarea de flujo de datos, se ejecutan varios flujos de datos propiamente dichos, y que utilizan un conjunto de herramientas que realizan su cometido.

Se utilizan cinco "OLE DB Source" como flujo de datos de origen, y son utilizados para extraer datos de las carreras, colegios, aspirantes, del tiempo y de las facultades.

Se usan cinco "OLE DB Destination" como flujo de datos de destino, y son utilizados para cargar los datos extraídos y transformados (según sea necesario) a las tablas de stage.

Como parte de las transformaciones de datos, se usa cinco "Derived Column" que usan expresiones para actualizar columnas. Uno se usa para obtener la modalidad de las carreras, otros para reemplazar la denominación de algunas carreras y una última para quitar espacios innecesarios de los nombres y apellidos de los aspirantes.

Como flujo de datos de transformación, también se usan dos "Condicional Split" que permiten evaluar y direccionar datos hacia un subconjunto específico de estos. Uno es usado para seleccionar las unas carreras según un cierto criterio y el otro para seleccionar un rango de fechas.

Finalmente se ocupa un "Union All" que permite unir varios conjuntos de datos.

Como detalle de DataFlowTask_StageHechos, puedo indicar que esta tarea está constituida por un "OLE DB Source" que extrae los datos provenientes de la vista de hechos, un "Derived Column" que facilita el cálculo de las preguntas vacías y un "OLE DB Destination" que carga los datos en la tabla de stage de hechos.

A continuación se presenta los diagramas de flujo de datos de las actividades antes descritas:

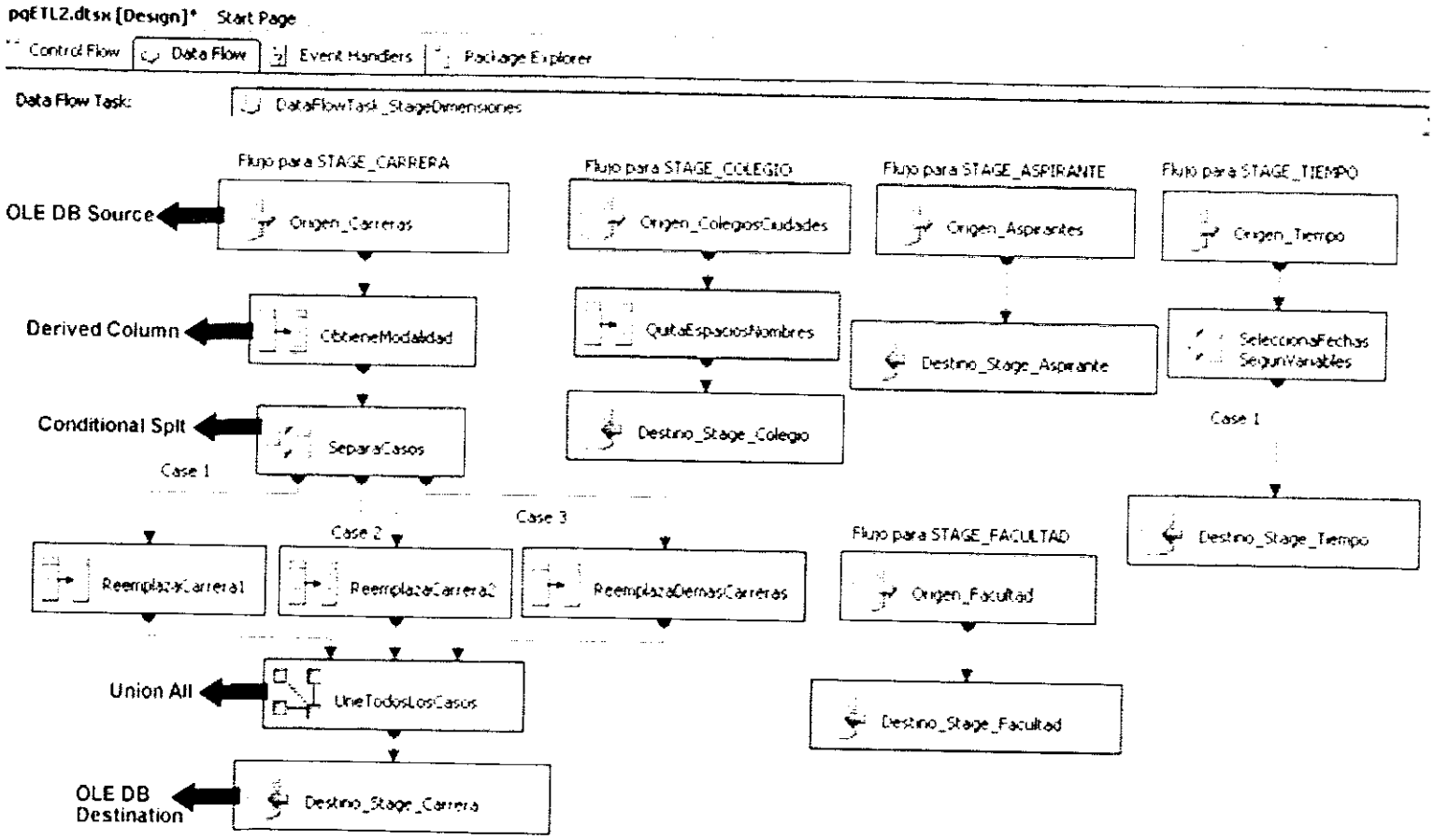


**Figura 4.8:** Flujo de datos del Paquete ETL2

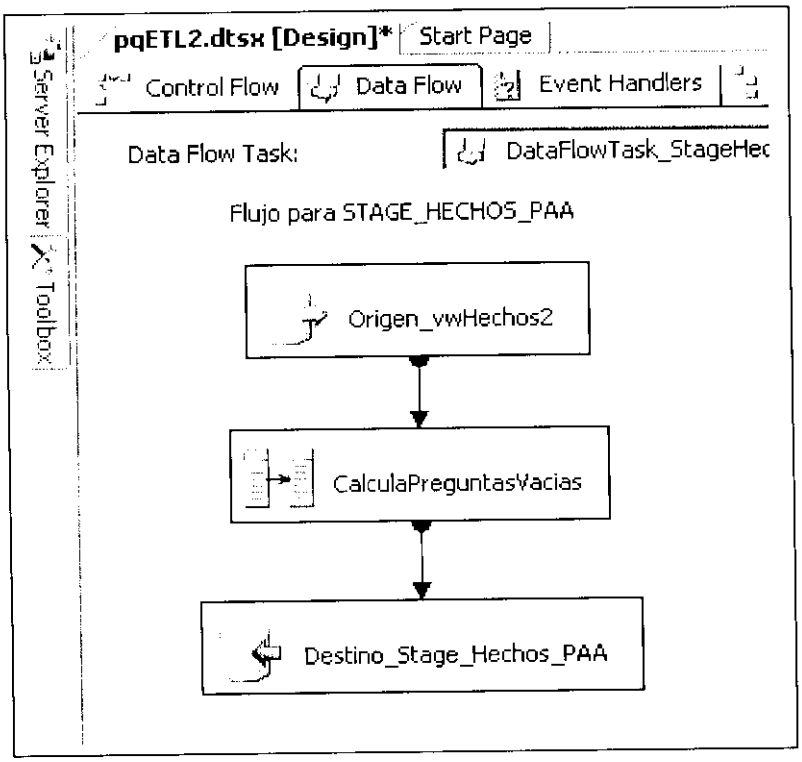


**Figura 4.9:** Flujo de datos del Paquete ETL2 STAGE_HECHOS

## 4.6. Carga de las tablas de Dimensiones

El objetivo de esta tarea consiste en pasar los datos de las tablas de stage hacia las dimensiones del DM. Para ello se utiliza el paquete ETL3, cuyo flujo de control consta de 14 "Execute SQL Task" y una "Data Flor Task".

Los 14 elementos ejecutan sentencias SQL para limpiar y cargar datos en todas las tablas de dimensiones y el otro componente es destinado a la carga de la tabla de hechos.

El diagrama de este flujo de datos se muestra a continuación:

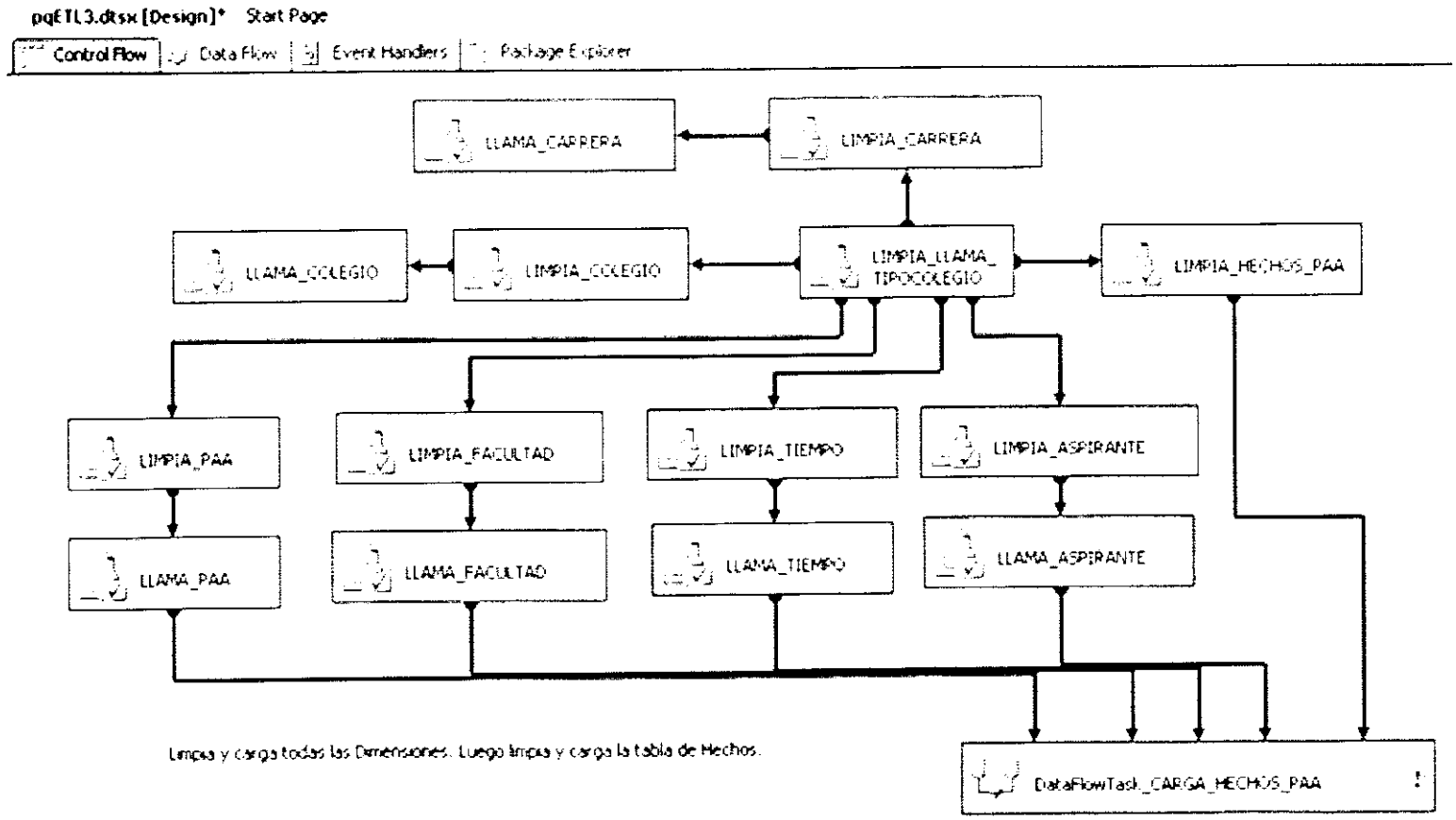


**Figura 4.10:** Flujo de control del Paquete ETL3

## 4.7. Carga de la tabla de Hechos

Luego de que las dimensiones han sido cargadas exitosamente, se ejecuta la tarea de flujo de datos que se encarga de llevar los datos desde la tabla de hechos de stage hacia la tabla de hechos del DM.

Para efectuar esta tarea, el flujo de datos del paquete ETL3 está constituido por un origen y destino de datos, así como siete "Lookup". Esta transformación Lookup usa las claves almacenadas en la tabla STAGE_HECHOS para recuperar las claves foráneas de las dimensiones y así poder cargar la información a la tabla de hechos HECHOS_PAA del DM.

El siguiente es el gráfico que indica como funciona flujo de datos del paquete ETL3 para la carga de la tabla de hechos:

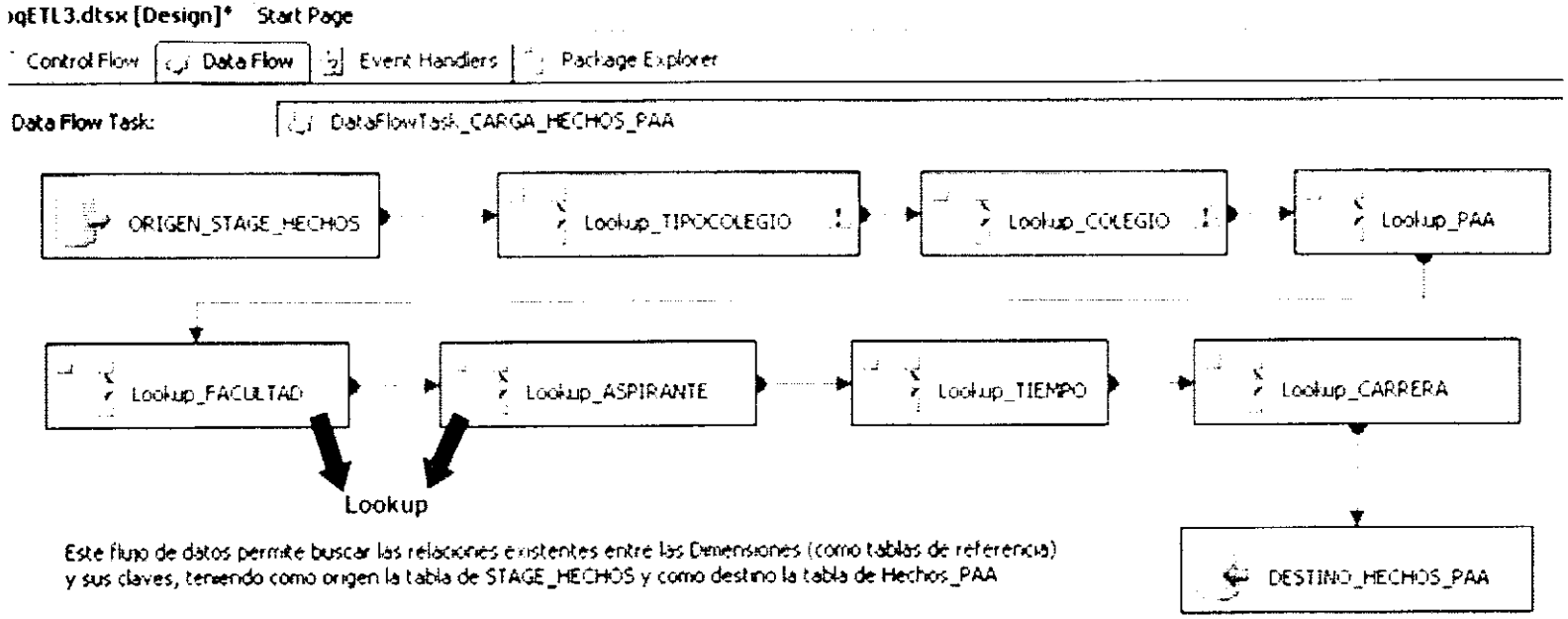


**Figura 4.11:** Flujo de datos del Paquete ETL3

## 4.8. Generación de informes

Para poder generar los informes necesarios con la información proveniente de nuestro DM denominado PAA_MART, se necesita de la creación del cubo de información.

Esta actividad está desarrollada con el empleo de Microsoft SSAS, y para lograrlo hay que seguir varios pasos:

- Lo primero que se hace es crear un origen de datos, que contiene una cadena de conexión y las credenciales necesarias para poder recuperar información proveniente de nuestro DM.
- A continuación se crea una vista del origen de datos (Data Source Views - DSV), que contiene la representación lógica de las tablas que son provenientes de uno o más orígenes de datos. En sí constituye la estructura del cubo.
- El paso a seguir lo constituye el desarrollo del cubo propiamente dicho, para ello se puede usar el asistente de creación que permite seleccionar de manera sencilla todas las medidas que necesitamos, así como las diferentes dimensiones que vamos a usar para poder visualizar varios reportes.
- Luego se realiza el procesamiento del cubo que realmente se hace en dos tareas internas: el procesamiento de las dimensiones y el procesamiento del cubo propiamente dicho. La primera tarea lee la información de las tablas de dimensiones y estructura un mapa de la dimensión para cada atributo y la

segunda combina dichos mapas de dimensiones con un mapa multidimensional del cupo, tareas que se realizan de manera eficiente.

- Por último se puede usar el Browser de Microsoft SSAS para mirar los resultados de las consultas que se deseen realizar. El uso de dicha herramienta facilita el despliegue y análisis de varios tópicos de la información.

Otra herramienta que se puede usar para visualizar la información proporcionada por el cubo constituye Microsoft Excel.

Las siguientes son figuras que muestran: la estructura generada en SSAS de nuestro cubo de información, dos ejemplos de informes generados por el Browser de SSAS (el primero muestra los aspirantes que han aprobado la PAA de la facultad de Contabilidad y Auditoría, de la carrera de Economía, provenientes de un colegio fiscal o a distancia y que han sido ordenados según su nota de manera descendente; el segundo muestra un informe que se relaciona con los aspirantes de la facultad de Contabilidad y Auditoría que han aprobado la PAA, que pertenecen a la carrera de Contabilidad y Auditoría y que provienen del Ins. Tec. Hispano América, igualmente ordenado por su nota con un desglose de la misma. Adicionalmente se puede observar el nombre de varios aspirantes), así como varios ejemplos de informes que se pueden obtener mediante el uso de Excel.

El primer ejemplo de Excel, muestra un gráfico que detalla el porcentaje de aspirantes que han aprobado la PAA, clasificados por colegio, tipo de colegio y ubicación geográfica.

El segundo ejemplo muestra un informe de la cantidad de aspirantes inscritos para la PAA según las facultades disponibles en la Universidad Técnica de Ambato.

El tercer ejemplo de Excel, indica la cantidad de aspirantes inscritos según la provincia de procedencia.

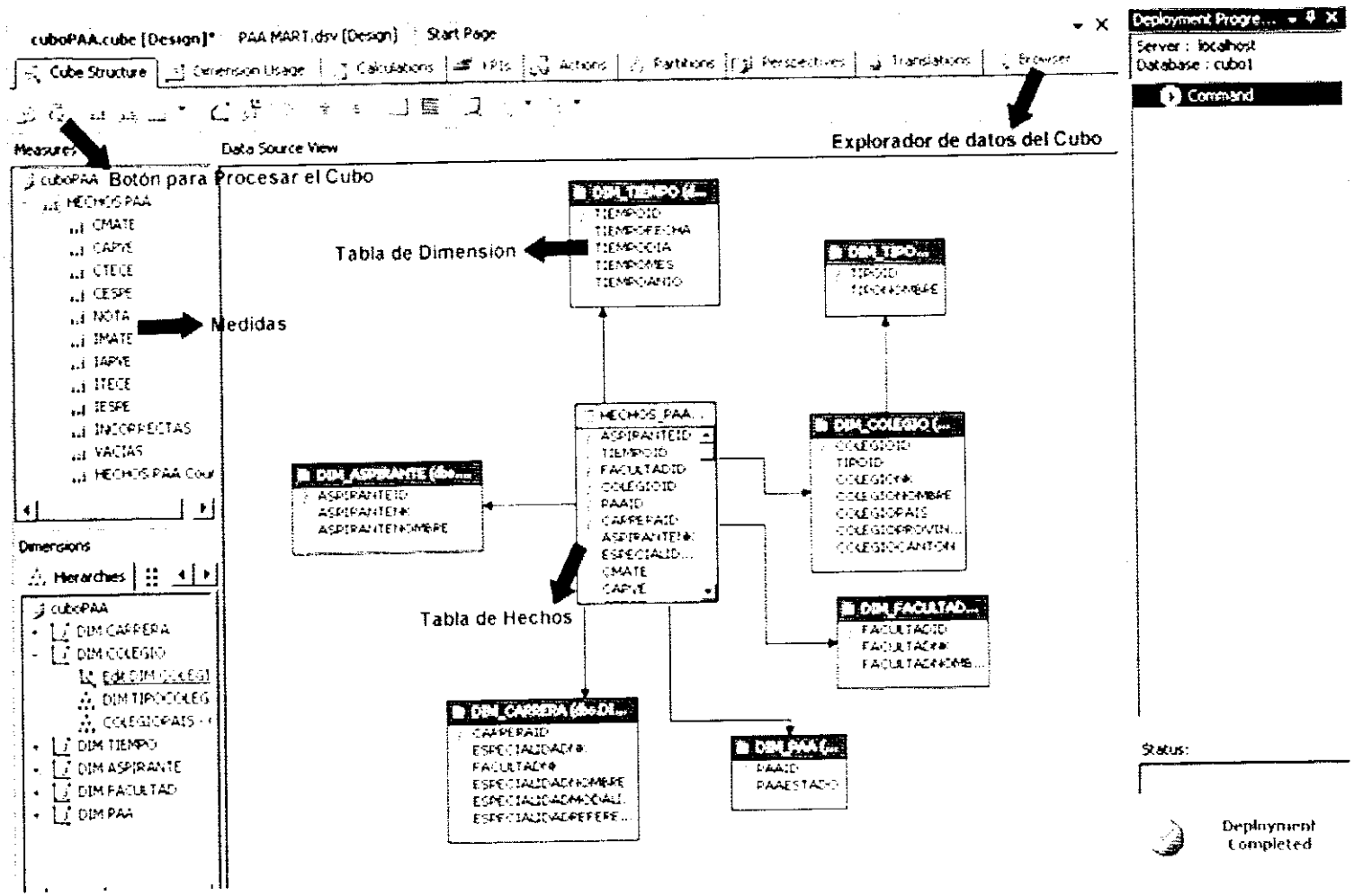


**Figura 4.12:** Estructura del cubo cuboPAA

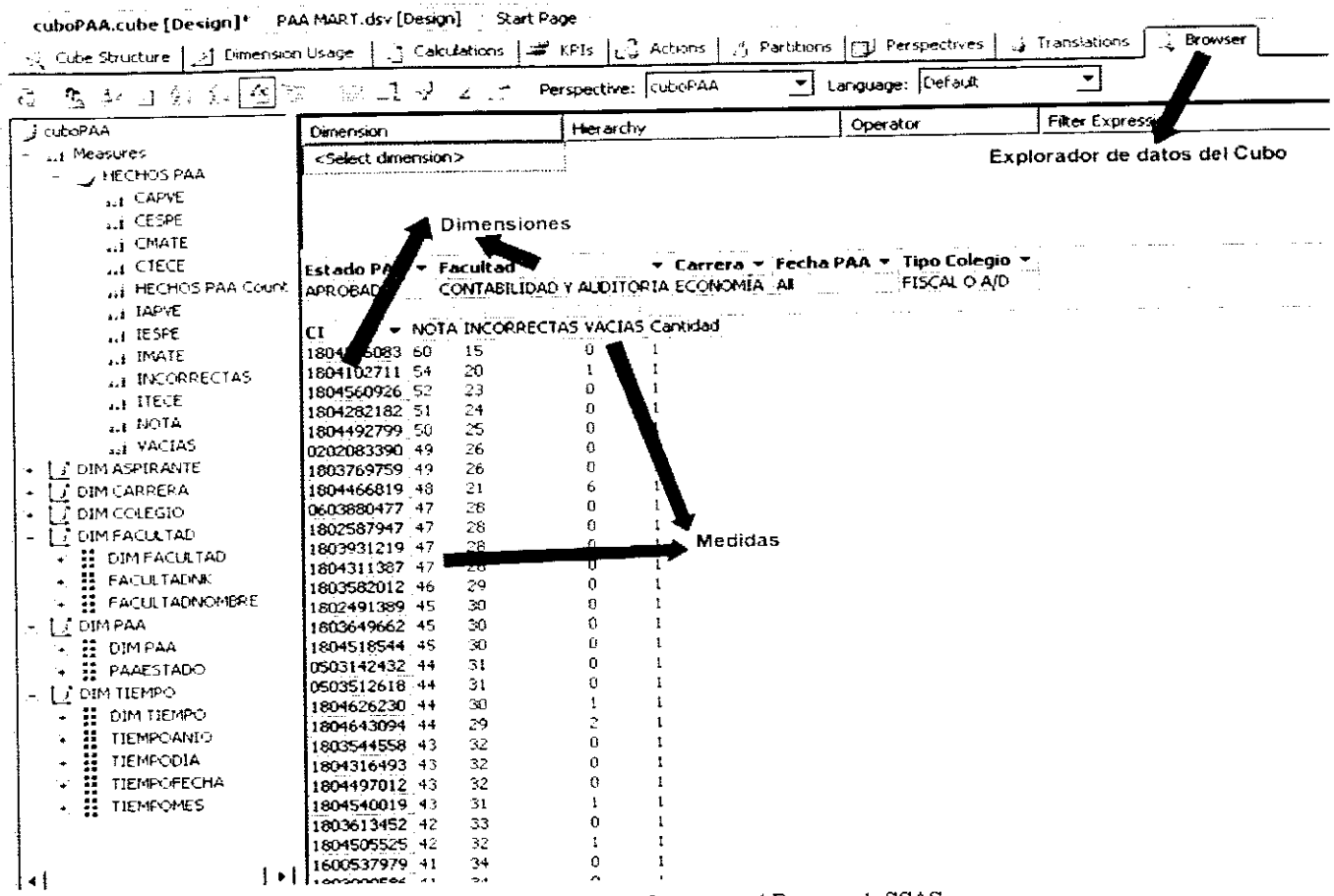


**Figura 4.13:** Ejemplo1 informe con el Browser de SSAS

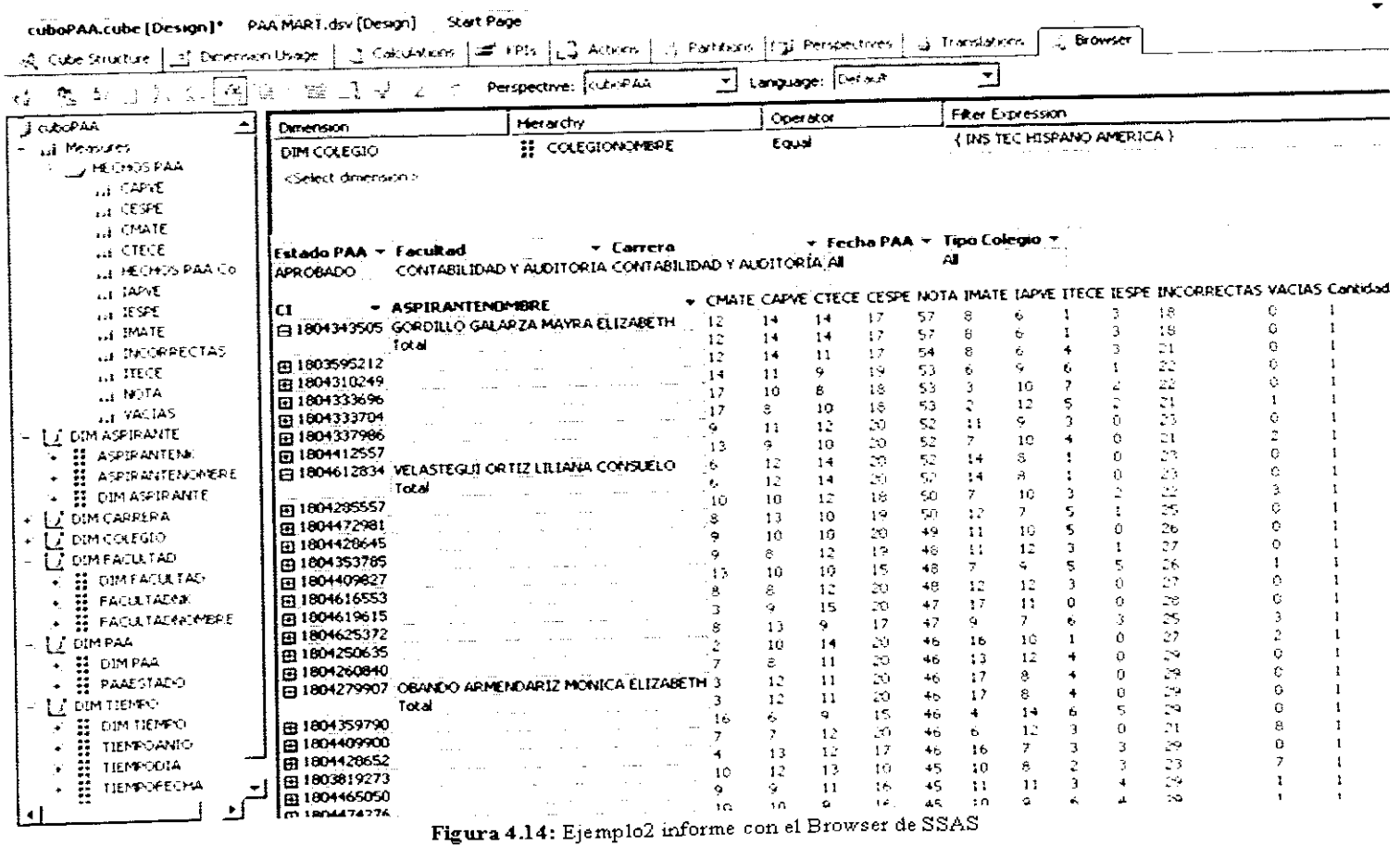


**Figura 4.14:** Ejemplo2 informe con el Browser de SSAS

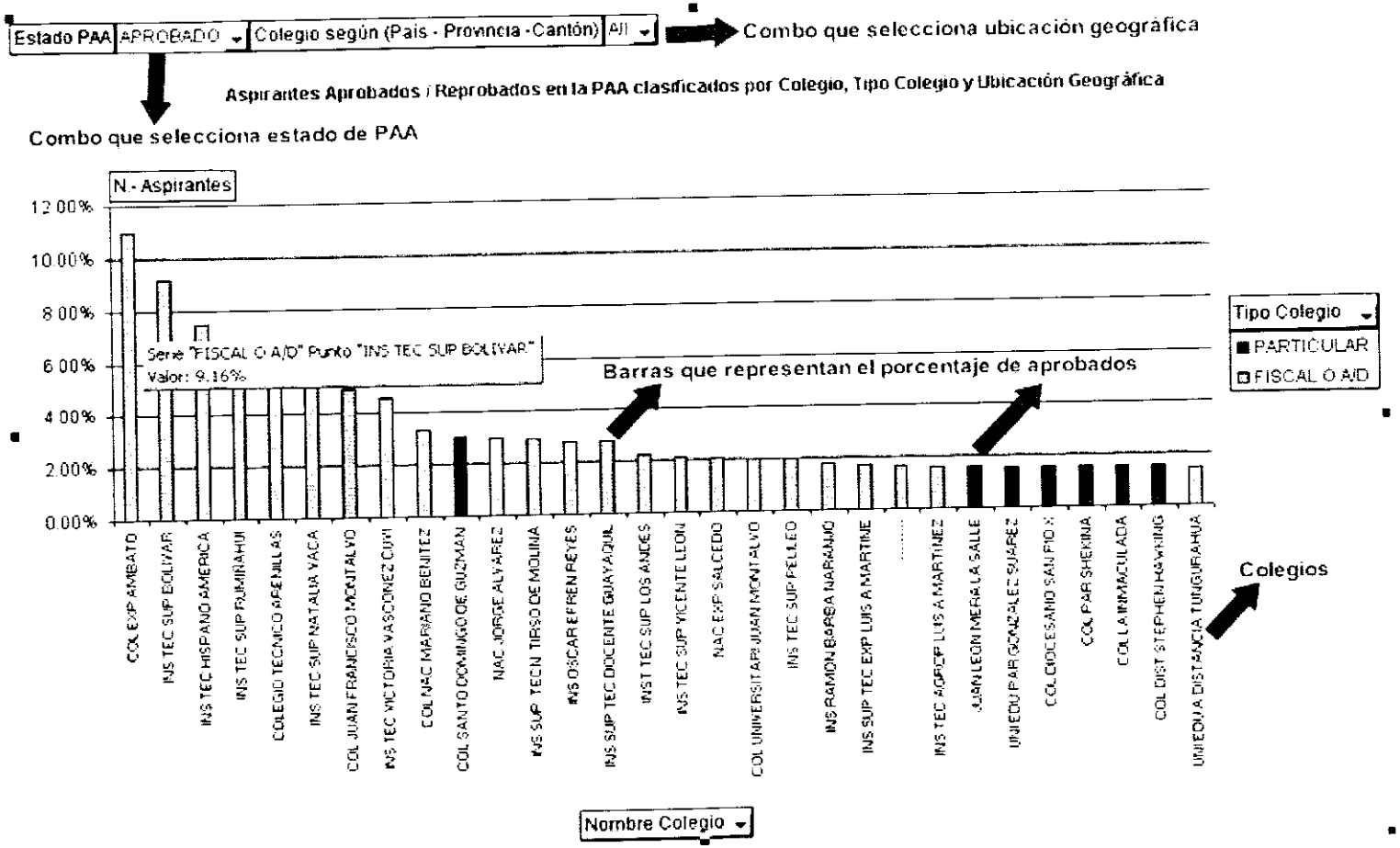


**Figura 4.15:** Ejemplo1 de informe con Excel

| Modalidad | Facultad | N.- Aspirantes |
|---|---|---|
| PRESENCIAL | | |
| | ADMINISTRACION CENTRAL | 117 |
| | CIENCIAS ADMINISTRATIVAS | 801 |
| | CIENCIAS DE LA SALUD | 987 |
| | CIENCIAS HUMANAS Y DE LA EDUCACION | 475 |
| | CONTABILIDAD Y AUDITORIA | 672 |
| | INGENIERIA AGRONOMICA | 77 |
| | INGENIERIA CIVIL | 308 |
| | INGENIERIA EN ALIMENTOS | 147 |
| | INGENIERIA EN SISTEMAS | 387 |
| | JURISPRUDENCIA Y CIENCIAS SOCIALES | 447 |
| Total PRESENCIAL | | 4001 |
| SEMIPRESENCIAL | | |
| | CIENCIAS ADMINISTRATIVAS | 71 |
| | CIENCIAS HUMANAS Y DE LA EDUCACION | 437 |
| | CONTABILIDAD Y AUDITORIA | 121 |
| | JURISPRUDENCIA Y CIENCIAS SOCIALES | 12 |
| Total SEMIPRESENCIAL | | 644 |
| Total general | | 4645 |

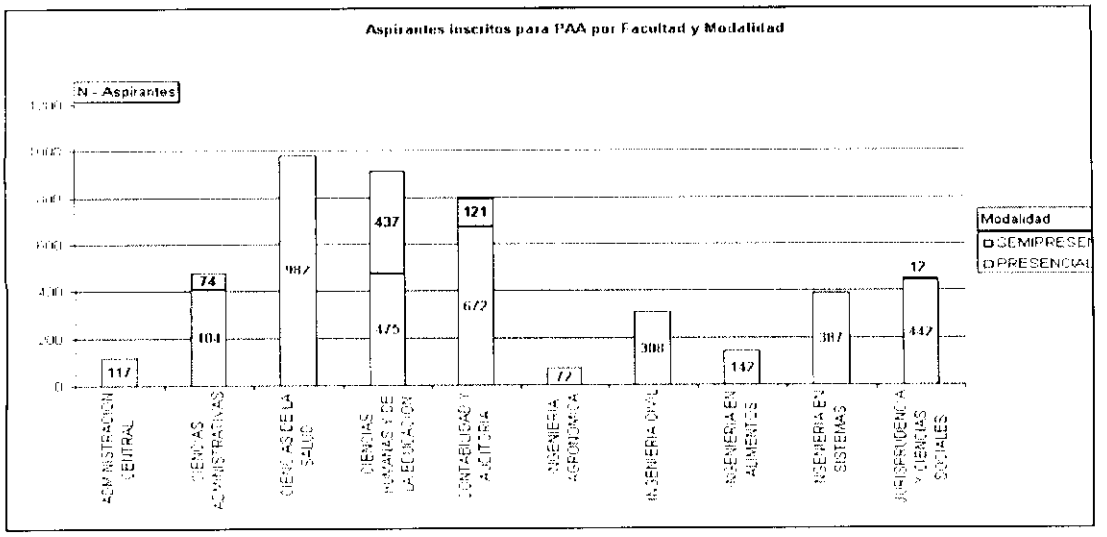


**Figura 4.16:** Ejemplo2 de informe con Excel

| País | Provincia | N. Aspirantes |
|---|---|---|
| ECUADOR | | |
| | AZUAY | 6 |
| | BOLIVAR | 107 |
| | CAÑAR | 2 |
| | CARCHI | 9 |
| | CHIMBORAZO | 78 |
| | COTOPAXI | 682 |
| | EL ORO | 135 |
| | ESMERALDAS | 16 |
| | GALAPAGOS | 1 |
| | GUAYAS | 21 |
| | IMBABURA | 6 |
| | LOJA | 1 |
| | LOS RIOS | 7 |
| | MANABI | 6 |
| | MORONA SANTIAGO | 4 |
| | NAPO | 27 |
| | ORELLANA | 11 |
| | PASTAZA | 111 |
| | PICHINCHA | 199 |
| | SUCUMBIOS | 16 |
| | TUNGURAHUA | 3139 |
| | ZAMORA CHINCHIPE | 3 |
| Total ECUADOR | | 4564 |
| Total general | | 4564 |

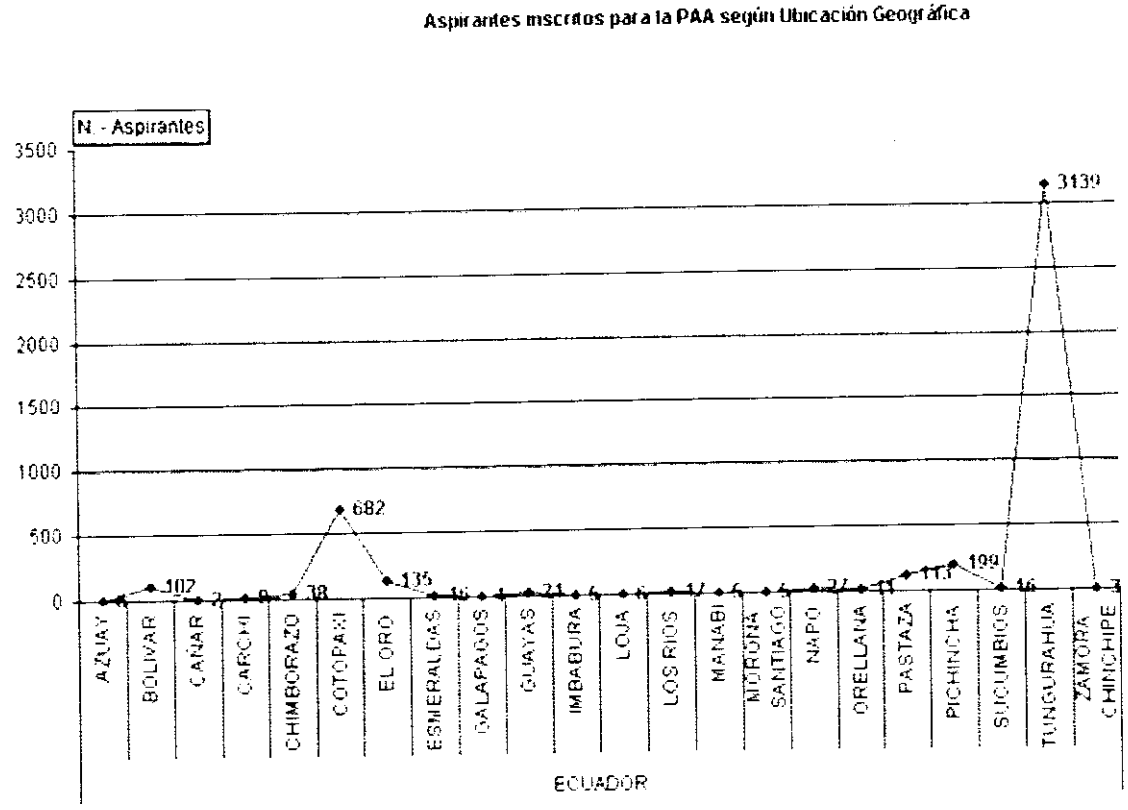


**Figura 4.17:** Ejemplo3 de informe con Excel

# CAPÍTULO V

# 5. CONCLUSIONES Y RECOMENDACIONES

## 5.1. Conclusiones

- El Data Warehouse se constituye en una infraestructura apropiada para los directivos de una organización, con ella se puede acceder a datos de diferentes fuentes de información que luego pueden ser utilizados en tareas de auto evaluación que permitan analizar las diferentes áreas de dicha organización y tomar las mejores decisiones para su permanente progreso.
- El Data Mart desarrollado en este proyecto facilita la mejor información que apoye a la toma de decisiones tanto al Sr. Rector, Vicerrectores: Académico y Administrativo, así como a la Dirección de Sistemas Informáticos y Redes de Comunicación (DISIR)
- La decisión acerca de usar un Data Warehouse o un Data Mart como herramienta estructural para el análisis de los requerimientos de una organización, no depende de una guía específica, sino más bien de las características particulares y de las necesidades de cada organización. En nuestro estudio se optó por la realización de un Data Mart.
- Disponer de este tipo de herramienta no es tarea sencilla, debido a que requiere de un alto compromiso por parte de diferentes unidades de la Universidad.
- Para el desarrollo de un Data Warehouse existen varias metodologías que se pueden utilizar, sin embargo, para el prototipo propuesto se usó la metodología de Sakhr Youness debido a su fácil comprensión y a su corto

tiempo de desarrollo.

- La tarea que provocó una mayor demanda en este proyecto fue el proceso de extracción, transformación y carga (ETL).
- El hecho de que los diferentes sistemas de información de la Universidad Técnica de Ambato trabajan sobre bases de datos de Microsoft SQL Server 2005, facilitó en gran medida el desarrollo de este Data Mart, además que se utilizaron herramientas propias de Microsoft como SSIS y SSAS.
- Una vez que se desarrolló el cubo de información de la PAA, se puede usar Microsoft Excel como herramienta para mostrar varios informes o reportes.

## 5.2. Recomendaciones

- Es recomendable no continuar en el proceso de Data Warehousing mientras no se tenga claro cual es el negocio de la organización.
- Se recomienda con la continuidad de este proyecto, pues se ha logrado realizar la parte inicial que corresponde a la Prueba de Aptitud Académica, lo siguiente por hacer debe estar dirigido a quienes ya son estudiantes de la Universidad.
- Previo al desarrollo de un Data Warehouse, resulta importante realizar un estudio que permita esquematizar una estrategia que se pueda utilizar.
- Se debe determinar si nuestro proyecto encaja o no en los lineamientos de una metodología de desarrollo específica.
- Cuando el modelo de datos en estrella de un Data Warehouse tiene en sus tablas de dimensiones muchos atributos, se recomienda utilizar el modelo en copo de nieve, que permite desnormalizar dichas tablas, optimizando así los procesos de consultas.
- Se sugiere que el área de Stage o preparación siempre debe ser considerada al momento de realizar los procesos de extracción, transformación y carga.
- De ser posible, se recomienda el uso de herramientas especializadas para las actividades de extracción, transformación, carga y explotación de la información.
- Si el campo de acción de un Data Warehouse está orientado a un tema específico de la organización, es recomendable dividirlo en uno o más Data Mart.

## 5.3. Demostración de la Hipótesis

Basado en las necesidades que tiene la Dirección de Sistemas Informáticos y Redes de Comunicación (DISIR) y en general la Universidad Técnica de Ambato, se procedió con el desarrollo del prototipo inicial denominado Caso de Estudio: Desarrollo de un Sistema de Apoyo a la Toma de Decisiones que Gestione la Información de la Prueba de Aptitud Académica de la Universidad Técnica de Ambato.

Luego de haber instalado la base de datos PAA_MART con todas sus tablas de dimensiones, hechos y de stage e instalado los diferentes paquetes de información así como el denominado cuboMART, se procedió a la ejecución de diferentes consultas (varias de las mismas están visibles en el Capítulo IV y son las Figuras: 4.13, 4.14, 4.15, 4.16 y 4.17) con lo que se logró conseguir varios informes de los aspirantes que rindieron la Prueba de Aptitud Académica. Varios de ellos son: quienes rindieron la prueba según su colegio de procedencia, ubicación geográfica, tipo de colegio y estado de la prueba (es decir, aprobado o reprobado). Otro informe que se logró probar es la cantidad de aspirantes inscritos a la prueba y distribuidos en las diferentes facultades y carreras de la Universidad, filtrado además por la modalidad de estudio (presencial o semipresencial). Cada reporte dio sus resultados en cortos intervalos de tiempo, logrando la obtención de varias estadísticas que anteriormente se conseguían incluso en semanas posteriores a la recepción de la Prueba de Aptitud Académica, ya que muchas de ellas se las hacía de manera manual.

Por otra parte haciendo uso del cálculo proposicional, se conoce según Rolando Saenz **[24]** que una proposición en matemáticas tiene el mismo significado que en un lenguaje cualquiera (por ejemplo el lenguaje castellano). Una proposición es una afirmación o enunciado de la cual se puede decir si es verdadera o falsa.

Con ellas se pueden utilizar las denominadas reglas de inferencia que constituyen un razonamiento lógico que suele tomar la forma de la proposición condicional de la forma H→T a partir de la cual, conociendo una hipótesis (premisas) se busca llegar a una tesis (conclusión).

La condicional o implicación es aquella operación que establece entre dos enunciados una relación de causa – efecto. La regla de separación o Modus Ponendo Pones que constituye uno de los medios de demostración más usado, significa "afirmando afirmo" y en un condicional establece, que si el antecedente (primer término, en este caso P) se afirma, necesariamente se afirma el consecuente (segundo término, en este caso Q) y su representación puede ser expresada de la siguiente manera:

Proposición 1: P→Q

Proposición 2: P

---

Conclusión : Q

Es así que:

Aplicando lo antes indicado se puede separar a la hipótesis planteada en dos variables: una Independiente (a la que llamo H) y una Dependiente (a la que llamo T),

En donde:

H= El desarrollo de un sistema de apoyo a la toma de decisiones

T= Mejorará la gestión de la información de la Prueba de Aptitud Académica de la Universidad Técnica de Ambato

Se tiene que según la regla de separación:

Proposición 1: H→T

Proposición 2: H

---

Conclusión : T

Lo que significa que: con el desarrollo de un sistema de apoyo a la toma de decisiones se logra una mejor gestión de la información de la Prueba de Aptitud Académica de la Universidad Técnica de Ambato. Esto se garantiza con la presentación de varios informes que se construyeron a partir de la información proporcionada por el cubo de información desarrollado.

# BIBLIOGRAFÍA


[1]: LAUDON, Kenneth, LAUDON, Jane: (2004) **"Sistemas de Información Gerencial – Administración de la Empresa Digital"**, Octava Edición, Edit. Pearson Educación de México, S.A. de C.V., México.

[2]: DIAZ, José Luis: (2006) **"Data Warehouse: funcionalidad y servicios"**, http://www.tecnomarkets.com.

[3]: IMHOFF, Claudia, GALEMMO, Nicholas, GEIGER, Jonathan: (2003) **"Mastering Data Warehouse Design – Relational and Dimensional Techniques"**, Edit. Wiley Publishing, Inc., United States of America.

[4]: KIMBALL, Ralph, CASERTA, Joe: (2004) **"The Data Warehouse ETL ToolKit – Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data"**, Edit. Wiley Publishing, Inc., United States of America.

[5]: DOS SANTOS, Romina: (2005) **"Bases de Datos Multiplataforma como soporte para la Inteligencia de Negocios"**, Trabajo de Adscripción en la Universidad Nacional del Nordeste – Facultad de Ciencias Exactas, Naturales y Agrimensura, Argentina.

[6]: HARJINDER, Gill, PRAKASH, Rao: (1996) **"DATA WAREHOUSING – La integración de información para la mejor toma de decisiones"**, Edit. Prentice Hall Hispanoamericana, S.A., México.

[7]: INMON, William: (2005) **"Building the Data Warehouse"**, Cuarta Edición, Edit. Wiley Publishing, Inc., United States of America.

[8]: HOBBS, Lilian, HILLSON, Susan, LAWANDE, Shilpa, SMITH, Pete: (2005) **"Oracle Data Base 10g - Data Warehousing"**, Edit. Elsevier Digital Press, United States of America.

**[9]**: SIMSION, Graeme, WITT, Graham: (2005) **"Data Modeling Essentials"**, Tercera Edición, Edit. Morgan Kauffman Publishers y Elsevier Digital Press, United States of America.

**[10]**: MICROSOFT, Press: (2000) **"SQL Server 7.0 Data Warehousing Training Kit"**, Edit. Microsoft Press, United States of America.

**[11]**: FRANCO, Michel, EDS – Institut Prométhéus: (1997) **"El Data Warehouse – El Data Mining"**, Primera Edición, Edit. Gestión 2000, S.A., Barcelona.

**[12]**: WIKIPEDIA: (2005) **"ETL"**, http://es.wikipedia.org/wiki/ETL

**[13]**: KIMBALL, Ralph, ROSS, Margy: (2002) **"The Data Warehouse ToolKit"**, Segunda Edición, Edit. Wiley Publishing, Inc., United States of America.

**[14]**: MICHAEL, Corey, MICHAEL, Abbey: (1997) **"ORACLE Data Warehousing"**, Primera Edición, Edit. McGRAW-HILL/INTERAMERICANA DE ESPAÑA, S.A.U., Madrid.

**[15]**: JARKE, Matthias, LENZERINI, Maurizio, VASSILIOU, Yannis, VASSILIADIS, Panos: (2000) **"Fundamentals of Data Warehouses"**, Edit. Springer - Verlag, Berlin Heidelberg.

**[16]**: KIMBALL, Ralph, REEVES, Laura, ROSS, Margy, THORNTHWAITE, Warren: (1998) **"The Data Warehouse Lifecycle ToolKit – Expert Methods for Designing, Developing and Deploying Data Warehouses"**, Edit. Wiley Publishing, Inc., United States of America.

**[17]**: YOUNESS, Sakhr: (2000) **"Professional Data Warehousing with SQL Server 7.0 and OLAP Services"**, Edit. Wrox Press, Ltd., United States of America.

**[18]**: WIKIPEDIA en español: (2005) **"Información Cubos OLAP"**, http://es.wikipedia.org/wiki/Cubos_Olap

[19]: SAS, Institute: (2000) **"Rapid Data Warehousing Methodology"**, http://www.sas.com

[20]: H. CONSEJO UNIVERSITARIO DE LA UNIVERSIDAD TÉCNICA DE AMBATO: (2000) **"Estatuto de la Universidad Técnica de Ambato"**

[21]: MSDN, Microsoft: (2007) **"Integration Services"**,
http://www. msdn.microsoft.com/es-ec/library/ms141026.com.aspx

[22]: MSDN, Microsoft: (2007) **"Analysis Services"**,
http://www. msdn.microsoft.com/es-es/library/ms176117.com.aspx

[23]: MICROSOFT, Libros en pantalla: (2005) **"Integration Services"**

[24]: SAENZ, Rolando: (1988) **"Fundamentos de Matemática – Introducción al Cálculo"**, Universidad Central del Ecuador.

# GLOSARIO DE TÉRMINOS

**Área de Stage** — Área de preparación, a la que se puede cargar información previa a la carga de un Data Warehouse.

**Cubo de información** — Estructura que facilita y agilita la consulta de información histórica ofreciendo la posibilidad de navegar y analizar los datos.

**Data Mart** — Base de datos orientada a un tema específico, puesto a disposición de los usuarios en un contexto de decisión descentralizado.

**Data Mining** — Conjunto de tecnologías avanzadas susceptibles de analizar la información de un Data Warehouse para obtener sus tendencias, para segmentar la información o para encontrar correlaciones en los datos.

**Data Warehouse** — "Almacén de datos". Base de datos específica del mundo de la decisión destinada principalmente a analizar las palancas de negocio potenciales.

**DBMS** — Data Base Management System, sistema de administración de base de datos o simplemente sistema de bases de datos.

| | |
|---|---|
| **Data Warehousing** | Proceso de desarrollo de un proyecto de Data Warehouse. |
| **Dimensión** | Eje de análisis asociado a los indicadores; corresponde normalmente a los temas de interés del Data Warehouse. |
| **DISIR** | Dirección de Sistemas Informáticos y Redes de Comunicación de la Universidad Técnica de Ambato. |
| **DSS** | Decision Support Systems, sistemas para el apoyo de toma de decisiones. |
| **ETL** | Extract, transform and load; extraer, transformar y llamar o cargar. |
| **Hechos** | Colección de datos relacionados con temas específicos de una organización. |
| **Medidas** | Características cualitativas o cuantitativas de aspectos del negocio que las organizaciones desean analizar. |
| **Metadato** | Información que describe un dato. |
| **Modelo de datos** | Esquema de una base. El modelo describe las tablas, los atributos, las claves, las restricciones de integridad. |

| | |
|---|---|
| **Modelo en copo** | Técnica de modelado dimensional, derivada del modelado en estrella. |
| **Modelo en estrella** | Técnica de modelado dimensional, consistente en distinguir físicamente las tablas de hechos de las tablas de dimensiones. |
| **Modelo relacional** | Técnica de modelado consistente en descomponer una base de datos en entidades y relaciones correlacionando estas entidades. |
| **OLAP** | On Line Analytical Processing, procesamiento analítico en línea; permite formular consultas más sofisticadas, y después visualizar los resultados correspondientes. |
| **OLTP On Line Transactionnel Processing** | Tipo de entorno de tratamiento de la información en el que debe darse una respuesta en un tiempo aceptable y consistente. |
| **PAA** | Prueba de Aptitud Académica de la Universidad Técnica de Ambato. |
| **PAA_MART** | Data Mart propuesto como herramienta de apoyo a la toma de decisiones de la información de la PAA de la |

Universidad Técnica de Ambato.

| | |
|---|---|
| **SSAS** | Microsoft SQL Server 2005 Analysis Services, proporciona funciones de procesamiento analítico en línea (OLAP) y minería de datos para soluciones Business Intelligence. |
| **SSIS** | Microsoft SQL Server 2005 Integration Services, plataforma de alto rendimiento que permite la integración de datos. |
| **UDM** | Unified Dimensional Model, proporciona una única vista empresarial de datos relacionales. |
| **UTAm@tico** | Macro sistema de información de la Universidad Técnica de Ambato, formado por varios subsistemas como el de Gestión Estudiantil. |

# ANEXO

# MANUAL DE USUARIO

El presente documento tiene por finalidad sintetizar los pasos necesarios para poder aprovechar de los beneficios del prototipo propuesto. De manera general van a existir dos posibles tipos de usuarios: quienes tienen un conocimiento técnico de lo que trata la gestión de un DW mediante herramientas de software de Microsoft y quienes únicamente van a hacer uso de los informes finales, que en nuestro caso van a usar Microsoft Excel.

Para los primeros, se necesita tener instalado en el computador Microsoft SQL Server 2005 Enterprise o Developer Edition, que incluya SSIS y SSAS, así como Microsoft Excel. Para el segundo grupo de usuarios, Microsoft Excel es la única herramienta de software que usarán para gestionar la información obtenida.

En este caso se va a indicar las actividades que debe realizar un usuario técnico durante la obtención y gestión de la información de la PAA.

Una vez verificado mediante SQL Server 2005 Management Studio la existencia de las bases de datos utamatico, evalucion y PAA_Mart, se debe localizar la carpeta en la que se encuentran almacenados los paquetes desarrollados por SSIS, y el cubo generado por SSAS. Para el caso de estudio y como práctica, el primer paquete en ejecutarse es el denominado ETL0, dicho paquete tiene por misión blanquear las tablas de dimensiones y de hechos del DM PAA_MART. Se abre el paquete seleccionado, y se lo ejecuta presionando el botón "Start Debugging" de la barra estándar de SQL Server Business Intelligence Development Studio.

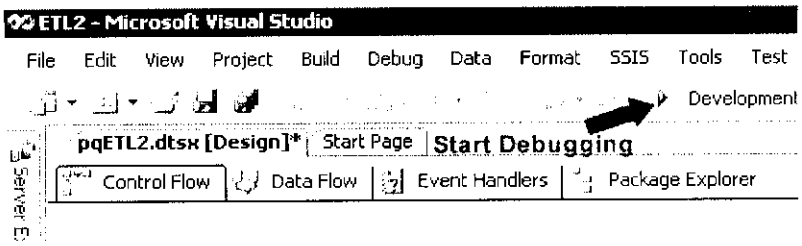


**Figura A1:** Botón Start Debugging de SSIS

Si la acción ha sido correctamente ejecutada por SSIS, se nota la apariencia de los diferentes componentes que toman una tonalidad color verde.

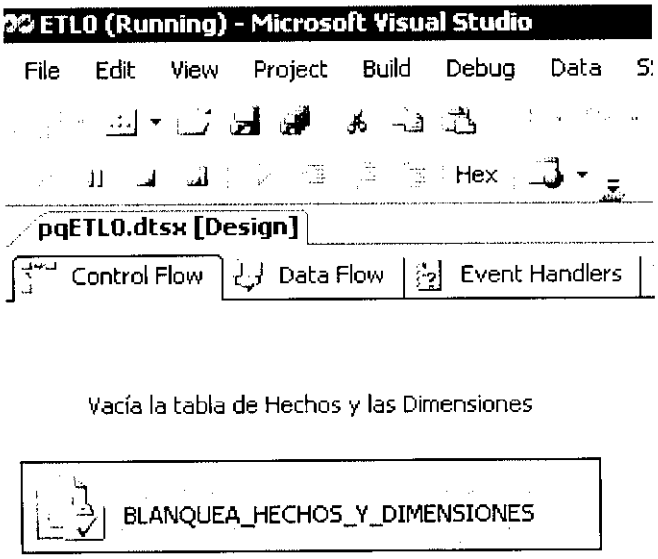


**Figura A2:** Paquete ETL0 ejecutado correctamente

Seguidamente se abre y ejecuta el paquete ETL1 que se encarga de limpiar las tablas de stage o preparación, así como cargar los valores iniciales en las tablas de stage de: STAGE_TIPOCOLEGIOS y STAGE_PAA. De igual forma visualizar que todos los componentes hayan realizado correctamente su tarea respectiva.

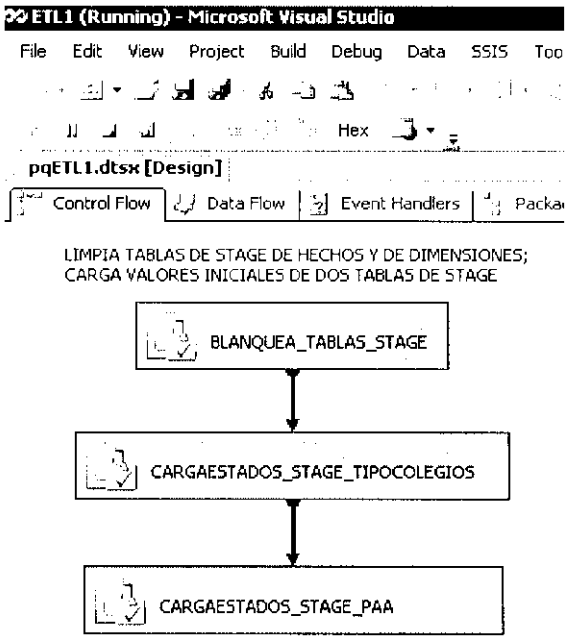


**Figura A3:** Paquete ETL1 ejecutado correctamente

Una vez que ha sido localizado, se ejecuta el paquete ETL2 que realiza la extracción, transformación y carga de la información desde los diferentes orígenes hacia las tablas de stage.

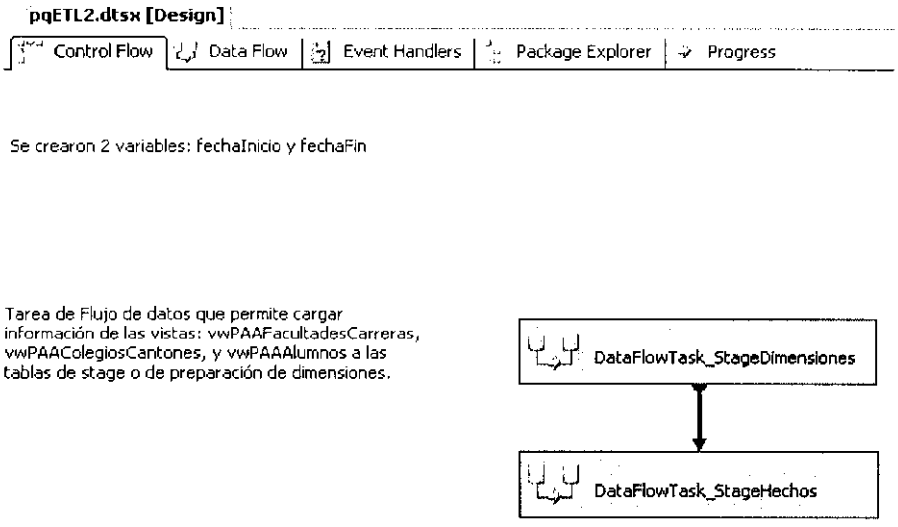


**Figura A4:** Paquete ETL2 ejecutado correctamente

En este caso el paquete también está compuesto por herramientas de flujo de datos que deben ser visualizadas de igual forma con una tonalidad verde, demostrando el cumplimiento correcto de todas las actividades realizadas.

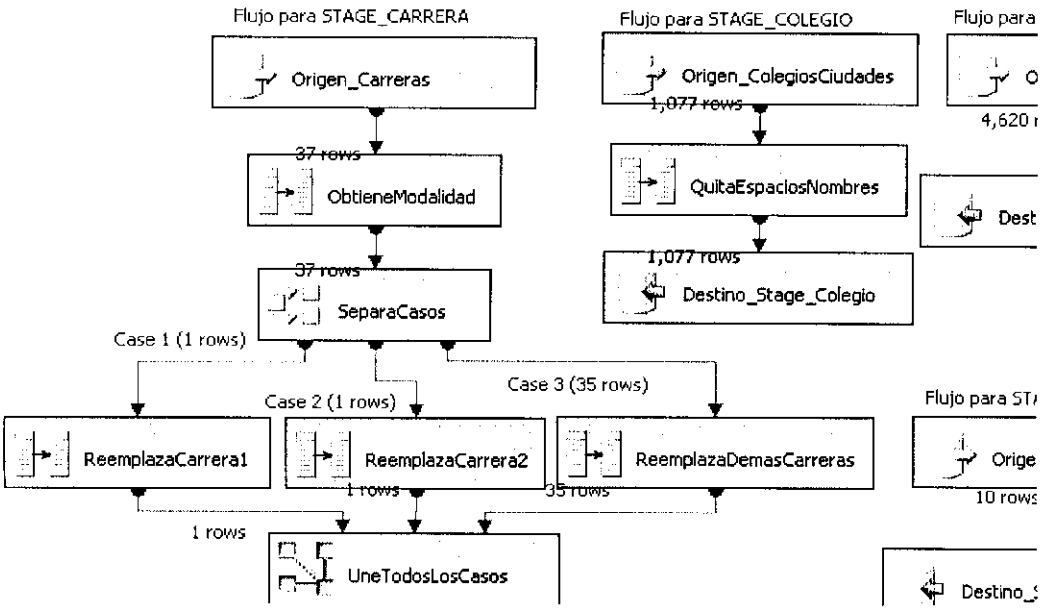


**Figura A5:** Tarea de flujo de datos de ETL2 ejecutado correctamente

El último paquete que se debe ejecutar mediante SSIS es el denominado ETL3. Dicho paquete se encarga de limpiar todas las tablas de dimensiones y la tabla de hechos del DM PAA_MART. Luego de cumplir exitosamente las labores de limpieza, continúa con la carga de la tabla de hechos. No hay que olvidarse de la comprobación visual de que todo esté ejecutado de manera correcta. La siguiente figura es parte de este paquete correctamente ejecutado.

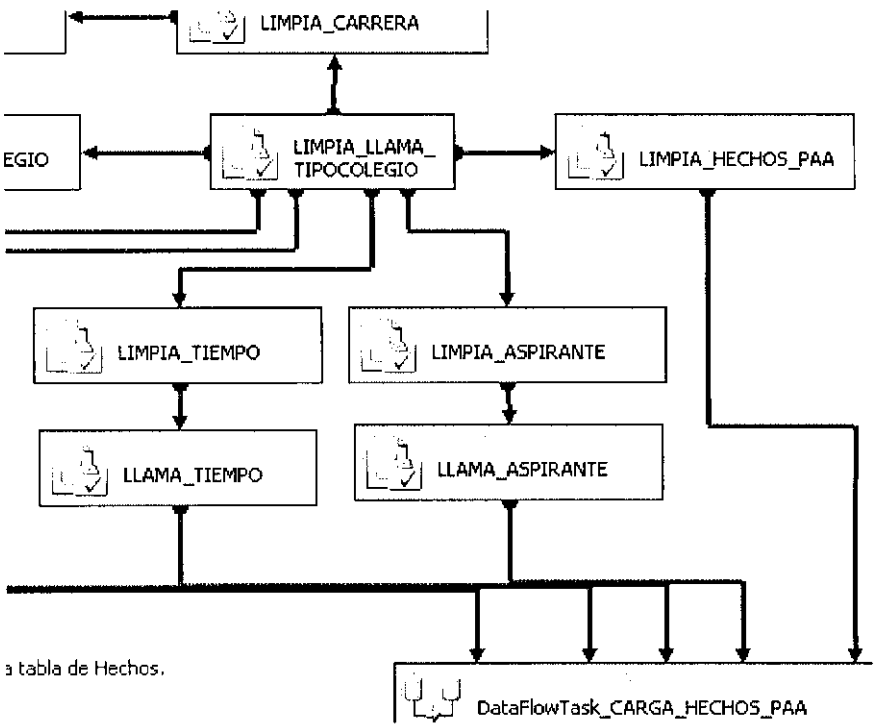


**Figura A6:** Paquete ETL3 ejecutado correctamente

El siguiente paso, constituye localizar y abrir el cubo de información que ha sido desarrollado con SSAS de Microsoft SQL Server 2005. Una vez seleccionada la pestaña o ficha denominada Browser, se puede apreciar todos los elementos que conforman el cubo (dimensiones, medidas, atributos, entre otros) y ciertos espacios o áreas en las que pueden ser arrastrados los elementos que nos permitan analizar la información de una manera personalizada.

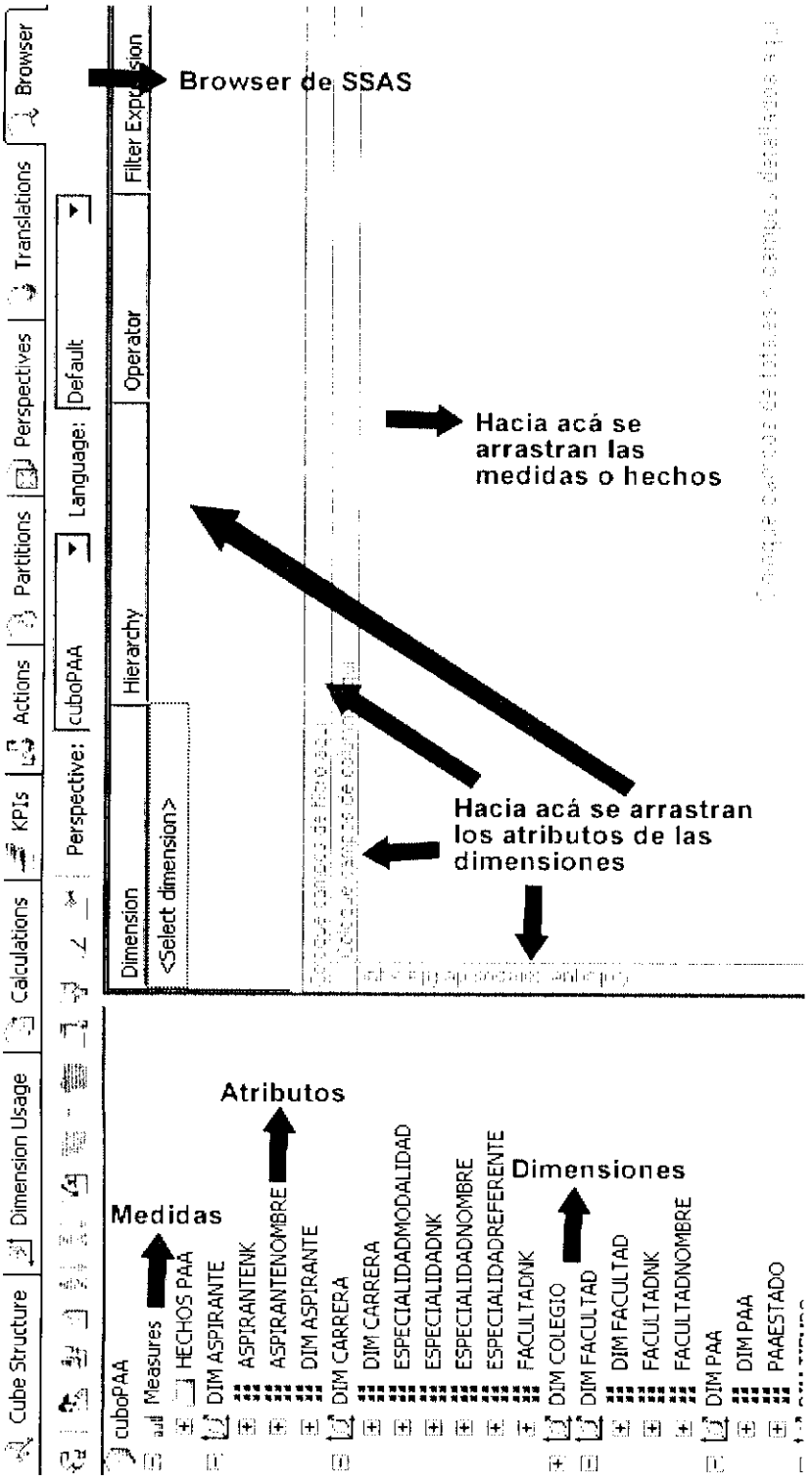


**Figura A7:** Explorador del cubo CuboPAA

Una vez escogidos y arrastrados los elementos deseados, se puede obtener una estructura de consulta como la siguiente:




**Figura A8:** Explorador del cubo CuboPAA con una consulta personalizada

Si se tiene los conocimientos suficientes se puede utilizar únicamente SSAS para visualizar la información proporcionada por el cubo. Sin embargo, otra alternativa de visualización de la información constituye Microsoft Excel. Esta herramienta puede ser usada por los dos tipos de usuarios antes indicados, procediendo de la siguiente forma para realizar gran variedad de informes personalizados; lo primero por hacer es iniciar Microsoft Excel, luego seleccionar de la barra estándar la opción Datos y en ella la tarea de Informe de tablas y gráficos dinámicos, como se muestra a continuación:

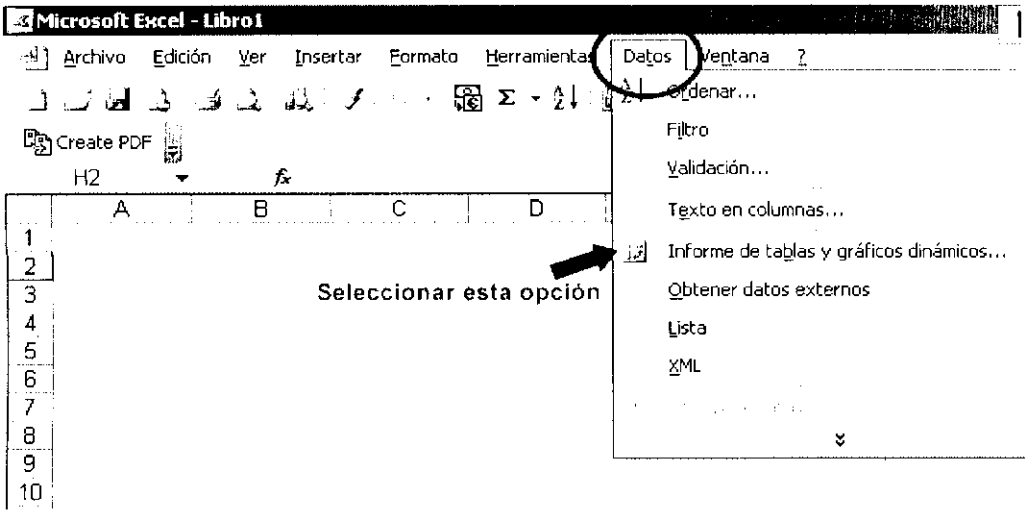


**Figura A9:** Inicio de asistente de Excel para tablas dinámicas

Seguidamente el asistente de Excel se encarga de guiar al usuario los diferentes pasos a realizar para la elaboración de la tabla dinámica que vamos a generar con los datos provenientes del cubo de información.

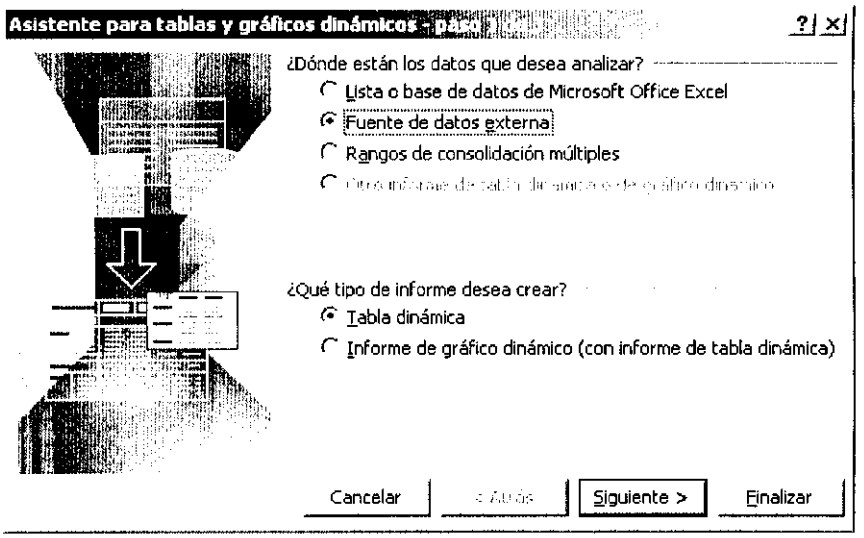


**Figura A10:** Pasos a seguir para obtener los datos del cubo

A continuación se elige el cubo que es nuestro origen de datos:

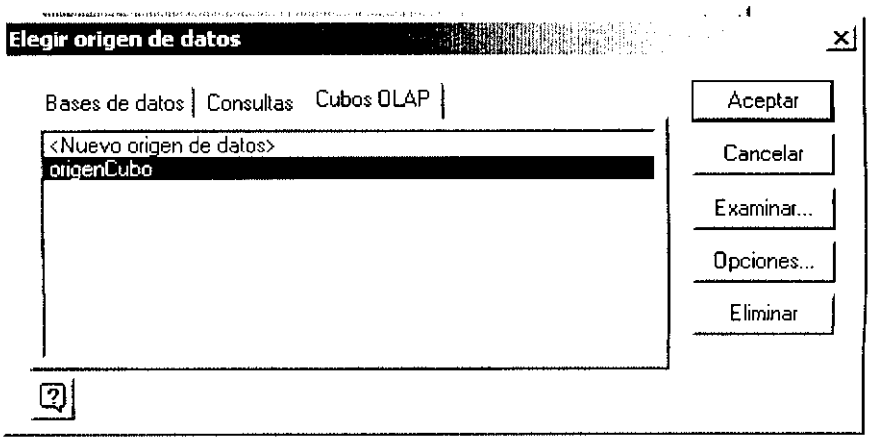


**Figura A11:** Origen de datos

Para el paso final de este asistente, en la sección de diseño se puede arrastrar y soltar los campos (ubicados a la derecha de la interface) a las diferentes secciones que permitan acomodar dichos elementos. La apariencia de esta situación se puede apreciar en la figura que se muestra a continuación:

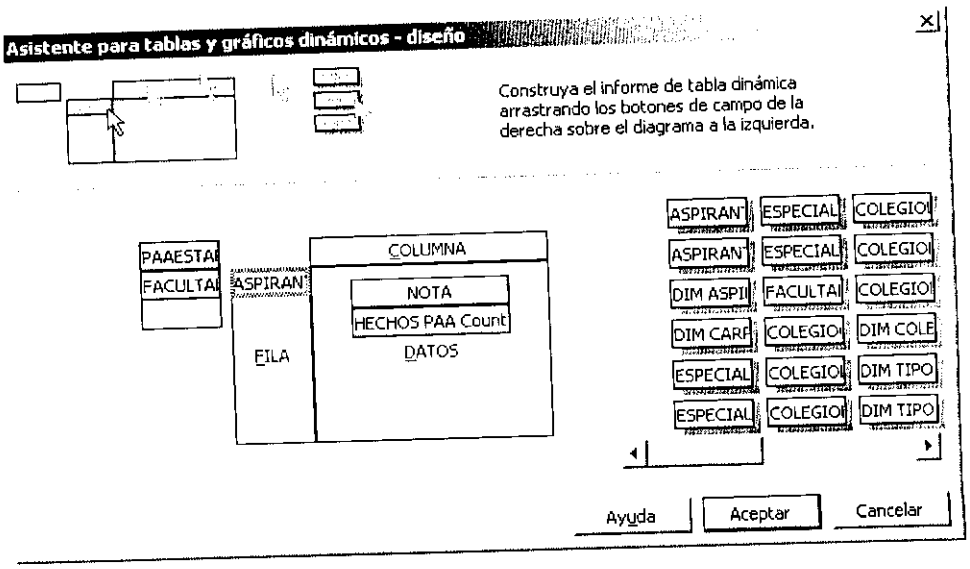


**Figura A12:** Paso de los campos a las áreas de análisis

Una vez presionado el botón de Aceptar, el asistente proporciona una hoja de Microsoft Excel con la información obtenida del cubo de información y que puede igualmente ser formateada y acomodada en un formato deseado.

| Modalidad | Facultad | N.- Aspirantes |
|---|---|---|
| PRESENCIAL | | |
| | ADMINISTRACION CENTRAL | 117 |
| | CIENCIAS ADMINISTRATIVAS | 404 |
| | CIENCIAS DE LA SALUD | 987 |
| | CIENCIAS HUMANAS Y DE LA EDUCACION | 475 |
| | CONTABILIDAD Y AUDITORIA | 677 |
| | INGENIERIA AGRONOMICA | 72 |
| | INGENIERIA CIVIL | 308 |
| | INGENIERIA EN ALIMENTOS | 142 |
| | INGENIERIA EN SISTEMAS | 387 |
| | JURISPRUDENCIA Y CIENCIAS SOCIALES | 442 |
| Total PRESENCIAL | | 4001 |
| SEMIPRESENCIAL | | |
| | CIENCIAS ADMINISTRATIVAS | 74 |
| | CIENCIAS HUMANAS Y DE LA EDUCACION | 137 |
| | CONTABILIDAD Y AUDITORIA | 121 |
| | JURISPRUDENCIA Y CIENCIAS SOCIALES | 12 |
| Total SEMIPRESENCIAL | | 644 |
| Total general | | 4645 |

**Figura A13:** Ejemplo de informe obtenido en Microsoft Excel