

**Universidad de las Ciencias Informáticas**  
**FACULTAD 6**



**Título: Mercado de datos Plan Turquino 2.0 para el Sistema de Información  
de Gobierno**

Trabajo de Diploma para optar por el título de  
Ingeniero en Ciencias Informáticas

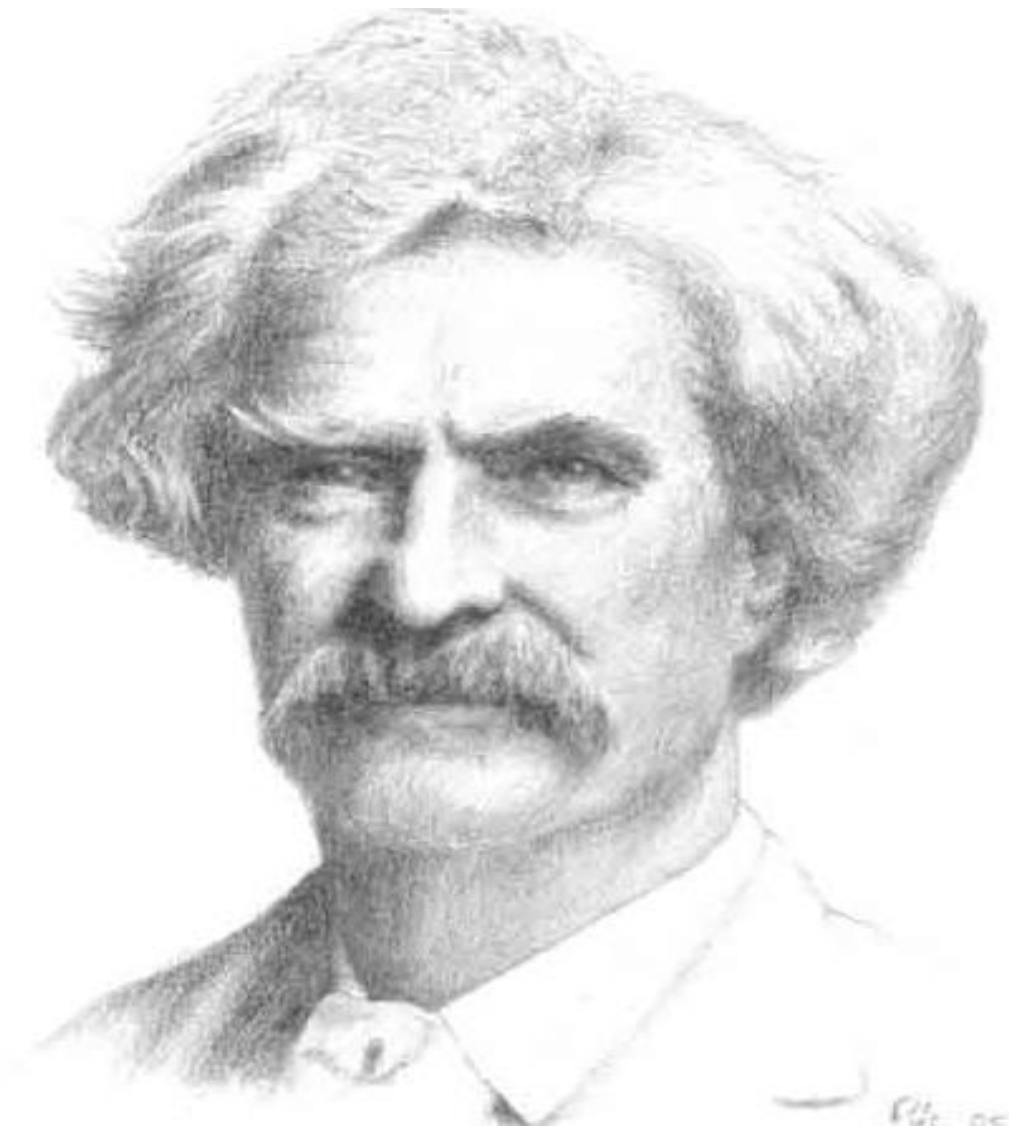
**Autor:**

Yousel Antonio Franco Núñez

**Tutores:**

Ing. Wendy Romalde Ruiz  
Ing. Mario Redonavich Gabey

**LA HABANA**  
**CUBA, JUNIO 2013**



*“El secreto para progresar es empezar por algún lugar. El secreto para empezar por algún lugar es fragmentar tus complejas y abrumadoras tareas de tal manera que queden convertidas en pequeñas tareas que puedas realizar y entonces simplemente comenzar por la primera.”*

*- Mark Twain*

## Declaración de autoría

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los \_\_\_\_ días del mes de \_\_\_\_\_ del año \_\_\_\_\_.

**Yousel Antonio Franco Núñez**

\_\_\_\_\_  
Firma del Autor

**Wendy Romalde Ruíz**

\_\_\_\_\_  
Firma de la Tutora

**Mario Redonavich Gabey**

\_\_\_\_\_  
Firma del Tutor

## Datos de contacto

**Tutora:** Ing. Wendy Romalde Ruiz

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Años de experiencia en el tema: 3

Años de graduado: 2

Correo electrónico: [wromalde@uci.cu](mailto:wromalde@uci.cu)

**Tutor:** Ing. Mario Redonavich Gabey

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Años de experiencia en el tema: 3

Años de graduado: 3

Correo electrónico: [mariorg@uci.c](mailto:mariorg@uci.c)

## *Agradecimientos*

*Agradezco a todas aquellas personas que de una forma u otra estuvieron involucrados en la realización de este trabajo, en especial al apoyo brindado por mi familia, a todos mis amigos que han compartido conmigo estos 5 largos años en la universidad, a mis tutores, por su apoyo incondicional, por el conocimiento que me supieron brindar y por su gran ayuda en la realización de este trabajo, al colectivo de profesores del departamento de almacenes de datos, muy en particular a Fabián, Adaleinis, Leonel (colora 'o), a todos muchas gracias por su apoyo.*

## ***Dedicatoria***

*Dedico esta tesis a mi familia, muy en especial a mi mamita Norvelis Núñez del Toro, por darme siempre su apoyo, confiar en mí y todo su esfuerzo para lograr graduarme finalmente. A mi hermano Jorge Mendoza Núñez, herma, gracias por ser como un padre para mí, por todo el esfuerzo realizado, sabes que te estaré agradecido eternamente y nunca me voy a olvidar de todo lo que has hecho por mí, de veras muchas gracias. A mi querida tía Maritza Guerra del Toro, por ser este el sueño y su deseo de toda la vida, el que me convirtiera en ingeniero, ya el sueño esta hecho realidad. A mis abuelos Elsa del Toro Robles y Mariano Guerra Castellanos que en paz descansen, siempre me apoyaron en todo, dándome sus consejos que me han sido de gran ayuda, lamentablemente mi abue Mariano no pudo ver este sueño hecho realidad, pero sé que donde se encuentre en este momento está muy feliz. Y finalmente a mi hermanita Kalía Mendoza Núñez, la única hembrita de mis hermanos y a quien quiero mucho, gracias a todo los mencionados nuevamente y a todos en general.*

## Resumen

El presente trabajo de diploma surge en el marco de trabajo del proyecto Sistema de Información de Gobierno del Centro de Tecnologías de Gestión de Datos conjuntamente con la Oficina Nacional de Estadísticas e Información. Tiene como objetivo desarrollar un almacén de datos compuesto por varios mercados de datos, entre los que se encuentra el referente al área de Plan Turquino. Tiene como objetivo complementar la información antes recogida a partir de la inclusión de la ficha de datos anual Plan Turquino, en la que se recoge una serie de indicadores de las distintas áreas del sector económico, social y administrativo de las zonas montañosas de nuestro país, incluyendo la Ciénaga de Zapata. Para la inclusión de esta ficha de datos se hace necesario realizar un proceso de análisis y diseño, de extracción, transformación y carga de los datos, además de visualizar los mismos, haciendo uso de las distintas herramientas como el Pentaho Data Integration, Data Cleaner, Pentaho BI\_Server y el sistema gestor de bases de datos PostgreSQL. La implementación de esta nueva versión facilitará el trabajo de los especialistas en el ámbito social del área del Plan Turquino, apoyando así la toma de decisiones.

## PALABRAS CLAVE

Almacén de datos, mercado de datos, Indicadores, ficha de datos, herramientas, toma de decisiones.

## Índice

AGRADECIMIENTOS.....	IV
DEDICATORIA.....	V
RESUMEN .....	VI
INTRODUCCIÓN.....	1
CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA .....	4
INTRODUCCIÓN.....	4
1.1. Caracterización general del Plan Turquino .....	4
1.2. Almacenes de Datos .....	4
1.3. Características de los AD .....	5
1.4. Mercado de Datos .....	5
1.5. Topologías de un almacén de datos.....	6
1.6. Elementos de un almacén de datos .....	6
1.7. Procesamiento Analítico en Línea .....	8
1.8. Metadatos .....	9
1.9. Extracción, Transformación y Carga .....	10
1.10. Inteligencia de negocio .....	10
1.11. Metodología de desarrollo de los AD .....	12
1.12. Herramientas utilizadas en la investigación .....	15
1.13. Conclusiones parciales.....	19
CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS PLAN TURQUINO .....	20



2.1.	Definición del negocio .....	20
2.2.	Especificación de requisitos.....	21
2.2.1.	Requisitos de información .....	21
2.2.2.	Requisitos funcionales .....	22
2.2.3.	Requisitos no funcionales .....	23
2.3.	Casos de uso del sistema.....	24
2.3.1.	Actores del sistema.....	25
2.3.2.	Casos de uso de información .....	25
2.3.3.	Casos de uso funcionales.....	28
2.3.4.	Diagrama de casos de uso.....	28
2.4.	Reglas del negocio.....	29
2.5.	Arquitectura del MD Plan Turquino.....	31
2.5.1.	Diseño del subsistema de almacenamiento.....	32
2.5.2.	Diseño del subsistema de integración.....	36
	Perfilado de los datos .....	36
2.5.3.	Diseño del subsistema de visualización.....	40
2.6.	Políticas de respaldo y seguridad.....	43
2.7.	Conclusiones Parciales.....	43
CAPÍTULO 3: IMPLEMENTACIÓN DEL MERCADO DE DATOS PLAN TURQUINO .....		45
3.1.	Implementación del subsistema de almacenamiento .....	45
3.1.1.	Estándares de codificación.....	45
3.1.2.	Implementación del subsistema de almacenamiento .....	46
3.2.	Implementación del subsistema de integración .....	46
3.2.1.	Implementación de las transformaciones.....	47

3.2.2. Implementación de los trabajos .....	49
3.3. Implementación del subsistema de visualización.....	50
3.3.1. Implementación de los cubos OLAP .....	50
3.3.2. Representación de la arquitectura de la información.....	51
3.3.3. Implementación de los reportes candidatos .....	51
3.4. Conclusiones parciales.....	52
CAPÍTULO 4: PRUEBAS DEL MERCADO DE DATOS PLAN TURQUINO .....	54
4.1. Pruebas aplicadas al MD Plan Turquino 2.0 .....	54
4.2. Herramientas de prueba .....	57
4.3. Resultados de las pruebas .....	58
CONCLUSIONES .....	60
RECOMENDACIONES.....	61
REFERENCIAS BIBLIOGRÁFICAS .....	62
BIBLIOGRAFÍA .....	64
ANEXOS .....	65
GLOSARIO DE TÉRMINOS.....	66

## Índice de tablas

Tabla 1: Descripción de los actores del sistema.....	25
Tabla 2: CU Mostrar información de Población. ....	28
Tabla 3: Hechos y descripción. ....	34
Tabla 4: Matriz BUS.....	35
Tabla 5: Roles y permisos. ....	43
Tabla 6: Estándares de codificación. ....	46
Tabla 7: CP correspondiente al CU "mostrar información de embalses" .....	58

## Índice de figuras

Figura 1: Diagrama de casos de uso. ....	29
Figura 2: Arquitectura del MD. ....	32
Figura 3: Modelo de datos. ....	36
Figura 4: Representación gráfica de los tipos de datos de la fuente. ....	37
Figura 5: Representación del análisis String a una de las pestañas de la fuente. ....	38
Figura 6: Diseño de las transformaciones (hecho centros educativos). ....	39
Figura 7: Diseño de las transformaciones (dimensión tipo de centros educativos). ....	40
Figura 8: Estructura de navegación. ....	41
Figura 9: Diseño de los cubos OLAP. ....	42
Figura 10: Organización de las tablas en la BD del MD. ....	46
Figura 11: Transformación dimensión tipo de centro. ....	48
Figura 12: Transformación hecho vivienda. ....	49
Figura 13: Trabajo para la carga de las dimensiones. ....	49
Figura 14: Trabajo para la carga de los hechos. ....	50
Figura 15: Trabajo principal. ....	50
Figura 16: Implementación del cubo hecho interés. ....	51
Figura 17: Arquitectura de información del MD Plan Turquino. ....	51
Figura 18: Cantidad de embalses del turquino y volumen de embalsado por dpa, año y tipo de embalses. .....	52
Figura 20: Modelo v. ....	55

## Introducción

El amplio desarrollo de la industria del software en el mundo ha sido de vital importancia en la productividad de muchos países, jugando un papel importante el software estadístico, el cual se ha convertido en el principal soporte para tener el control de grandes volúmenes de información. Además, el uso de estas aplicaciones ha permitido elevar la eficiencia y eficacia en cuanto a la toma de decisiones en el desarrollo de la sociedad actual.

Nuestro país no está ajeno a estos procesos de desarrollo. En el 1994 se crea la Oficina Nacional de Estadísticas e Información (ONEI) como resultado de la reorganización de los organismos de la Administración Central del Estado. Su principal objetivo se basa en captar, analizar y difundir los datos recogidos a lo largo y ancho de todo el país; además es responsable de gestionar los principales indicadores relacionados con diferentes áreas del sector económico, social y administrativo.

La Universidad de las Ciencias Informáticas (UCI), específicamente el Centro de Tecnologías de Gestión de Datos (DATEC), conjuntamente con la ONEI se ha dado la tarea de diseñar un almacén de datos compuesto por varios mercados de datos para el Sistema de Información de Gobierno (SIGOB), entre los que se encuentra el correspondiente al área Plan Turquino.

El Plan Turquino es el programa de desarrollo fundado por el Consejo de Estado de Cuba el 2 de junio de 1987 con el propósito de lograr un desarrollo integral y sostenible de las zonas montañosas y de difícil acceso del país, conjugando armónicamente los requerimientos productivos con el desarrollo social, la conservación de la naturaleza, y el fortalecimiento de la defensa del país, e integrando en sus acciones a los organismos e instituciones involucrados en ese proceso. (1)

El proyecto SIGOB nace de la necesidad de centralizar toda la información histórica existente en la ONEI, para lograr un mejor monitoreo y control de los datos que ahí se recogen. Se enfoca en la creación de una herramienta que permita acceder a toda la información, con el objetivo de apoyar la toma de decisiones en las diferentes áreas socioeconómicas.

En un primer instante se crea el mercado de datos (MD) Plan Turquino para SIGOB, con el objetivo de facilitar el trabajo a los especialistas y ayudar a la toma de decisiones, el mismo tenía como fuente de datos varios ficheros access (DBF) correspondientes al modelo 0024. Dicho modelo tenía como objetivo controlar las actividades trazadas en el Plan Turquino para el desarrollo integral de las zonas montañosas del país (2), tanto en los indicadores que caracterizan el desarrollo productivo como los del desarrollo social.

Sin embargo el desarrollo de la primera versión del mercado no cumplió totalmente las expectativas del cliente, ya que la información incluida en el sistema solo contiene indicadores que caracterizan al

desarrollo productivo. El análisis de los indicadores relacionados con el ámbito social al no ser incluidos en la primera versión actualmente se analizan de forma manual por los especialistas del área, tarea que resulta engorrosa debido a la variedad de sectores que se abarcan en la ficha de datos anual y al volumen de información. También la elaboración de informes en el ámbito social se ve afectada en cuanto a tiempo y esfuerzo, dificultándose la entrega al gobierno de información valiosa para la toma de decisiones que influyan en el desarrollo social del Plan Turquino. Producto a lo antes mencionado, actualmente el análisis estadístico a través de este mercado no satisface totalmente las necesidades de información de los especialistas, dificultando así la toma de decisiones correspondientes al desarrollo social del Plan Turquino.

Por la situación anteriormente descrita surge el siguiente **problema de la investigación**: *¿Cómo contribuir a la toma de decisiones en el área del desarrollo social del Plan Turquino del Sistema de Información de Gobierno?*

Definiéndose como **objeto de estudio**: Almacenes de datos y delimitando al **campo de acción**: Mercado de datos para el área Plan Turquino del Sistema de Información de Gobierno, quedando definido como

**objetivo general**: Desarrollar el mercado de datos Plan Turquino 2.0 del Sistema de Información de Gobierno que contribuya a la toma de decisiones en el ámbito social, desglosado en los siguientes **objetivos específicos**:

1. Fundamentar la selección de las metodologías, herramientas y tecnologías a utilizar en el desarrollo del mercado de datos Plan Turquino 2.0.
2. Realizar el análisis y diseño del mercado de datos Plan Turquino 2.0.
3. Implementar el mercado de datos Plan Turquino 2.0.
4. Validar el mercado de datos Plan Turquino 2.0.

Para dar cumplimiento a los objetivos antes planteados se definen las siguientes **tareas de la investigación**:

1. Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.
2. Levantamiento de los requisitos de información, funcionales y no funcionales, teniendo en cuenta las necesidades del cliente.
3. Descripción de los casos de uso del mercado de datos.

4. Definición de la arquitectura del mercado de datos, identificando los subsistemas fundamentales que componen la solución.
5. Definición de los hechos, las medidas y las dimensiones que permitirán conformar el modelo lógico correspondiente al mercado de datos.
6. Diseño del modelo físico del mercado de datos, teniendo en cuenta el modelo lógico diseñado.
7. Diseño del subsistema de integración de datos, quedando definidos los procesos de extracción, transformación y carga de los datos.
8. Diseño del subsistema de visualización, quedando definida la organización de la información en la capa de presentación de datos.
9. Implementación del modelo físico del mercado de datos.
10. Implementación del subsistema de integración de datos.
11. Implementación del subsistema de visualización.
12. Aplicación de los casos de pruebas para avalar el correcto funcionamiento del mercado de datos.

El documento tendrá la siguiente estructura capitular:

**Capítulo 1: Fundamentación teórica.** Este capítulo aborda varios elementos teóricos referentes a los almacenes de datos y mercados de datos. Además se hace referencia a las principales características de las herramientas que serán utilizadas en el transcurso de la investigación, así como, las principales razones que se tuvieron en cuenta para la selección de la metodología a utilizar.

**Capítulo 2: Análisis y diseño del mercado de datos Plan Turquino.** En este capítulo se abordará todo lo referente al proceso de análisis y diseño del mercado de datos. Se definirán los requisitos funcionales, no funcionales y de información, así como, las reglas del negocio; además se identificarán los hechos, medidas y dimensiones. Se definirán los reportes candidatos y el diseño del proceso de integración.

**Capítulo 3: Implementación del mercado de datos Plan Turquino.** En este capítulo se abordará todo lo referente al proceso de integración de datos que contiene los tres tipos de actividades de carácter general, como son: extracción, transformación y carga de los datos en el Mercado de Datos Plan Turquino. Se expondrán los principales elementos relacionados con la implementación del subsistema de visualización, almacenamiento e integración.

**Capítulo 4: Pruebas del mercado de datos Plan Turquino.** Se realizarán las pruebas pertinentes que comprueben la calidad del producto mediante las distintas herramientas, como son los casos de prueba y la lista de chequeo.

## Capítulo 1: Fundamentación teórica

### Introducción

Este capítulo abordará varios elementos teóricos referentes a los almacenes de datos y los mercados de datos. Además se hace referencia a las principales características de las herramientas que serán utilizadas en el transcurso de la investigación así como las principales razones que se tuvieron en cuenta para la selección de la metodología a utilizar.

#### 1.1. Caracterización general del Plan Turquino

A través del Plan Turquino se viene aplicando un conjunto de medidas encaminadas a impulsar el desarrollo económico y social en estas regiones del país. Estas medidas han permitido fortalecer la repoblación forestal en interés de la defensa, la flora y la fauna, creando para ello las condiciones básicas para el asentamiento de la población en estas zonas, para lo que se han invertido cuantiosos recursos tanto en la esfera productiva como en la social. (3)

No obstante los avances acumulados desde 1988 a 1994 se registraron contradicciones en la mayoría de los resultados productivos de estas zonas, básicamente en las producciones agrícolas de alimentos, ganadería y algunas actividades industriales producto a las dificultades por las que atraviesa el país. Desde 1995, debido fundamentalmente a la revitalización en la atención al desarrollo integral de estas zonas, se aprecia una recuperación en varios de sus indicadores. (3)

En los aspectos que miden el desarrollo social se continúa apreciando resultados favorables sostenidos, lo que se corresponde con la intención de continuar mejorando las condiciones de vida y trabajo de los hombres y mujeres de estos territorios. (3)

En la versión anterior del mercado de datos Plan Turquino, las estructuras dimensionales (tablas de hechos y dimensiones) abarcan la forma de organización del Modelo 0024-03, también conocido como indicadores del Plan Turquino. Dicha versión se concluyó satisfactoriamente, pero ha surgido una nueva necesidad de ampliar los datos antes recogidos con la inclusión de la ficha de datos anual del Plan Turquino. En esta nueva versión serán incluidos en el mercado los indicadores que caracterizan el desarrollo social, ya que anteriormente solo se tenían los indicadores que caracterizaban el desarrollo productivo, proceso que indudablemente fortalecerá el análisis estadístico en el área Plan Turquino de la ONEI y de gran necesidad para el país, permitiendo ampliar el universo de información del mercado y almacén en general.

#### 1.2. Almacenes de Datos



Para las grandes empresas, el control de inmensos volúmenes de datos siempre ha sido una gran interrogante por resolver. Por otro lado, el amplio auge que lleva la informatización del mundo moderno es otra de las causas que impulsaron a estas empresas a encontrar una solución a sus problemas. Dichas problemáticas han permitido buscar soluciones inmediatas, ¿resultado?, tener un mayor control y análisis estadístico de la información mediante la utilización de los almacenes de datos (AD).

### Definición de los AD

Un AD es una gran colección de datos que recoge información de múltiples sistemas fuentes, y cuya actividad se centra en la toma de decisiones –es decir, en el análisis de la información– en vez de en su captura (4). Una vez reunidos los datos de los sistemas fuentes se guardan durante mucho tiempo, lo que permite el acceso a datos históricos; así los AD proporcionan al usuario una interfaz consolidada única para los datos, lo que hace más fácil escribir las consultas para las tomas de decisiones. (4)

### 1.3. Características de los AD

Bill Inmon, conocido por muchos como el padre de los AD(s) define a los mismos en torno a las **características** del repositorio de datos en: (4)

- **Organizado en torno a temas:** la información se clasifica en torno a los aspectos que son de interés para la empresa.
- **Integrado:** la integración de los datos consiste en convenciones de nombres, codificaciones existentes, medida uniforme de variables, entre otros.
- **Dependiente del tiempo:** esta dependencia aparece de varias formas:
  - La información representa los datos en un horizonte largo de tiempo.
  - Cada estructura clave contiene (implícita o explícitamente) un elemento de tiempo (día, mes, año, etc.).
  - La información, una vez registrada correctamente, no puede ser actualizada.
- **No volátil:** el AD solo permite cargar nuevos datos y acceder a los ya almacenados, pero no permite borrar ni modificar los datos.

### 1.4. Mercado de Datos

Los MD se comportan como un AD, con la diferencia de que son concretos a un área en específico. Para un mejor entendimiento de lo que es un MD se define de la siguiente manera:

Un MD es un AD limitado a un área concreta de la organización. Muchos expertos definen el AD como un almacén centralizado que alimenta una serie de MD. Los MD se pueden generar obteniendo datos de un AD corporativo central o pueden ser creados independientemente de fuentes de datos independientes. Los MD independientes no son lo más adecuado, ni son recomendables, ya que originan islas de información, siendo esto precisamente lo que los AD intentan evitar. Existen también MD personales, que son subconjuntos de datos extraídos de MD departamentales o de unidades de negocio o pueden ser un AD que responde a los requerimientos de un único usuario o pequeño grupo de usuarios. (5)

Los MD se crean con el objetivo de mejorar el trabajo a los especialistas, facilitando la creación de los AD. Presentan un conjunto de **características** que los identifica, como son: (5)

- Se centran en los requisitos de los usuarios asociados a un departamento o área de negocio concreto.
- Son más sencillos a la hora de utilizarlos y comprender sus datos, debido a que la cantidad de información que contienen es mucho menor que en los AD.

### 1.5. Topologías de un almacén de datos

En un AD se reconocen varios esquemas para el modelado del Almacén de datos, siendo más utilizados tres variantes posibles de modelamiento, que permiten realizar consultas de soporte de decisión: (6)

- **Esquema en estrella:** el esquema en estrella consta de una tabla de hechos central y de varias tablas de dimensiones relacionadas a esta.
- **Esquema copo de nieve:** este esquema representa una extensión del modelo de estrella cuando las dimensiones se organizan en jerarquías de dimensiones.
- **Esquema constelación:** es el producto de la unión de varios esquemas, ya sea en estrella o como de nieve.

Dichos esquemas pueden ser implementados de diversas maneras, requieren que toda la estructura de datos este desnormalizada o semi desnormalizada, para evitar desarrollar uniones (Join) complejas, accediendo a la información con el fin de agilizar la ejecución de consultas. (6)

### 1.6. Elementos de un almacén de datos

#### Tablas de dimensiones

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

mediante los cuales los usuarios podrán filtrar y manipular la información almacenada en la tabla de hechos. (6)

Más detalladamente, cada tabla de dimensión podrá contener los siguientes campos: (6)

- Clave principal o identificador único.
- Datos de referencia primarios: datos que identifican la dimensión. Por ejemplo: nombre del cliente.
- Datos de referencia secundarios: datos que complementan la descripción de la dimensión. Por ejemplo: e-mail del cliente, fax del cliente, etc.

### **Tablas de hechos**

Contienen, precisamente, los hechos que serán utilizados por los analistas de negocio para apoyar el proceso de toma de decisiones. Contienen datos cuantitativos que son filtrados, agrupados y explorados a través de condiciones definidas en las tablas de dimensiones, obteniéndose de este modo una gran capacidad analítica. Los datos presentes en las tablas de hechos constituyen el volumen del AD, y pueden estar compuestos por millones de registros dependiendo de su granularidad y antigüedad de la organización. El registro del hecho posee una clave primaria que está compuesta por las claves primarias de las tablas de dimensiones relacionadas a este. (6)

### **Dimensión Tiempo**

En un AD, la dimensión tiempo es obligatoria ya que no es solo una secuencia cronológica representada de forma numérica, sino que mantiene niveles jerárquicos especiales que inciden notablemente en las actividades de la organización, y la definición de granularidad y jerarquía dependen de la dinámica del negocio que se esté analizando, además posee fechas especiales que inciden notablemente en las actividades de la organización. (6)

### **Jerarquías**

Una jerarquía representa una relación lógica entre dos o más atributos dentro de una misma dimensión. Las jerarquías poseen las siguientes características: (6)

- Pueden existir varias en una misma dimensión.
- Están compuestas por uno o más niveles.

La principal ventaja de manejar jerarquías, reside en poder analizar los datos desde su nivel más general al más detallado y viceversa, al desplazarse por los diferentes niveles. (6)

## Granularidad

La granularidad representa el nivel de detalle al que se desea almacenar la información sobre el negocio que se esté analizando. Por ejemplo, los datos referentes a ventas o compras realizadas por una empresa, pueden registrarse día a día, en cambio, los datos pertinentes a pagos de sueldos o cuotas de socios, podrán almacenarse a nivel de mes. (6)

Mientras mayor sea el nivel de detalle de los datos, se tendrán mayores posibilidades analíticas, ya que los mismos podrán ser resumidos o sumariados. Es decir, los datos que posean granularidad fina (nivel de detalle) podrán ser resumidos hasta obtener una granularidad media o gruesa. No sucede lo mismo en sentido contrario, ya que por ejemplo, los datos almacenados con granularidad media podrán resumirse, pero no tendrán la facultad de ser analizados a nivel de detalle. O sea, si la granularidad con que se guardan los registros es a nivel de día, estos datos podrán sumariarse por semana, mes, semestre y año, en cambio, si estos registros se almacenan a nivel de mes, podrán sumariarse por semestre y año, pero no lo podrán hacer por día y semana. (6)

### 1.7. Procesamiento Analítico en Línea

Para crear el MD de un área de la empresa es preciso encontrar la estructura óptima para el análisis de su información, proceso este de gran importancia para una mejor organización de la información en bases de datos de gran dimensión. Las aplicaciones Procesamiento Analítico en Línea (OLAP), son usadas para organizar grandes bases de datos empresariales.

#### Definición

Funciones que permiten crear bases de datos mucho mejor organizadas, en caso de que sean bases de datos demasiado grandes. Proporcionan la velocidad y la flexibilidad necesaria para dar apoyo al analista en tiempo real. (7)

#### ➤ Sistemas OLAP

Existen tres tipos de modelos OLAP, ellos son:

- **Procesamiento Analítico en Línea Relacional (ROLAP)**

Implementación OLAP que almacena los datos en un motor relacional. Normalmente, los datos son detallados, de esta manera se evitan las agregaciones y además las tablas se encuentran desnormalizadas. Los esquemas sobre los que se trabaja son el esquema de estrella, el esquema copo de nieve y constelación. (7)

- **Procesamiento Analítico en Línea Multidimensional (MOLAP)**

Implementación OLAP que almacena los datos en una base de datos multidimensional. Para optimizar los tiempos de respuesta, se calcula el resumen de la información por adelantado. Algunos sistemas utilizan técnicas de compresión de datos para disminuir el espacio de almacenamiento en disco debido a los valores pre calculados. (7)

- **Procesamiento Analítico en Línea Híbrido (HOLAP)**

Almacena algunos datos en un motor relacional y otros en una base de datos multidimensional. Este tipo de implementación utiliza las dos técnicas expuestas anteriormente. De ahí que es una técnica híbrida. (7) Los datos agregados y precalculados se almacenan en estructuras multidimensionales y los de menor nivel de detalle en estructuras relacionales. Es decir, se utilizará ROLAP para navegar y explorar los datos, y se empleará MOLAP para la realización de tableros.

## 1.8. Metadatos

Los metadatos son datos que describen o dan información de otros datos, que en este caso, existen en la arquitectura del DataWarehousing (almacenamiento de datos) (DW). Brindan información de localización, estructura y significado de los datos, básicamente mapean los mismos.

El concepto de metadatos es análogo al uso de índices para localizar objetos en lugar de datos. Es importante aclarar que existen metadatos también en las bases de datos transaccionales, pero los mismos son transparentes a los usuarios. La gran ventaja que trae aparejada el DW en relación con los metadatos es que los usuarios pueden gestionarlos, exportarlos, importarlos, realizarles mantenimiento e interactuar con ellos, ya sea manual o automáticamente. (6)

Se pueden distinguir tres diferentes tipos de metadatos: (6)

- **Los metadatos de los procesos Extracción, Transformación y Carga (ETL)**, referidos a las diversas fuentes utilizadas, reglas de extracción, transformación, limpieza, depuración y carga de los datos al depósito.
- **Los metadatos operacionales**, que son los que básicamente almacenan todos los contenidos del AD, para que este pueda desempeñar sus tareas.
- **Los metadatos de consulta**, que contienen las reglas para analizar y explotar la información del almacén, tales como drill-up y drill-down. Son estos metadatos los que las herramientas de análisis y consulta emplearán para realizar documentaciones y para navegar por los datos.

## 1.9. Extracción, Transformación y Carga

El término ETL se deriva de sus siglas en inglés Extract-Transform-Load que significan Extraer, Transformar y Cargar. ETL es el proceso que organiza el flujo de los datos entre diferentes sistemas en una organización y aporta los métodos y herramientas necesarias para mover datos desde múltiples fuentes a un almacén de datos, reformatearlos, limpiarlos y cargarlos en otra base de datos. ETL forma parte de la Inteligencia Empresarial (Business Intelligence), también llamado “Gestión de los Datos” (Data Management) (8). La idea es que un proceso ETL lea los datos primarios de sistemas principales, realice transformación, validación, el proceso cualitativo, filtración y al final escriba datos en el almacén y en este momento los datos se encuentran disponibles para analizar por los usuarios. (8)

### Herramientas y aplicaciones ETL del mercado (8)

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

Según **Ralph Kimball** (9) los procesos fundamentales para la integración de los datos en un almacén son:

- **Extracción:** se extraen los datos de las distintas fuentes hacia el área de preparación del AD, para su posterior transformación.
- **Transformación:** luego de ser extraída la información, pasa por procesos de filtrado, limpieza, depurado, homogenización y agrupación de la información. Este proceso es necesario para asegurar la calidad de los datos que serán almacenados en el AD.
- **Carga:** consiste en cargar los datos desde estas fuentes ya transformadas hacia el AD.

## 1.10. Inteligencia de negocio

Las aplicaciones de Business Intelligence (BI) son herramientas de soporte de decisiones que permiten en tiempo real, acceso interactivo, análisis y manipulación de información crítica para la empresa. Estas aplicaciones proporcionan a los usuarios un mayor entendimiento que les permite identificar las oportunidades y los problemas de los negocios. Los usuarios son capaces de acceder y apalancar una vasta cantidad de información y analizar sus relaciones y entender las tendencias que últimamente están apoyando las decisiones de los negocios. (10)

Los **objetivos** de la inteligencia de negocio (BI) son: (11)

- Establecer un ambiente único, empresarial, eficiente, flexible y confiable para la toma de decisiones.
- Integrar las diversas fuentes de información.
- Proveer rapidez, facilidad y flexibilidad de acceso a la información.
- Proveer un acceso intuitivo a la información, con una interfaz de usuario compuesta por gráficos e imágenes.
- Unificar los conceptos de indicadores e información clave a través de toda la organización.

Las herramientas de BI se basan en la utilización de un sistema de información de inteligencia de negocio que se forma con distintos datos extraídos de los datos de producción, con información relacionada con la empresa o sus ámbitos y con datos económicos. Este conjunto de herramientas tienen en común las siguientes **características**: (11)

- **Accesibilidad a la información.** Los datos son la fuente principal de este concepto. Lo primero que deben garantizar este tipo de herramientas y técnicas será el acceso de los usuarios a los datos con independencia de la procedencia de estos.
- **Apoyo en la toma de decisiones.** Se busca ir más allá en la presentación de la información, de manera que los usuarios tengan acceso a herramientas de análisis que les permitan seleccionar y manipular sólo aquellos datos que les interesen.
- **Orientación al usuario final.** Se busca independencia entre los conocimientos técnicos de los usuarios y su capacidad para utilizar estas herramientas.

## Algunas herramientas BI

### Código abierto

- Eclipse BIRT Project: generador de informes para aplicaciones Web de código abierto basado en Eclipse.
- JasperReports.

- LogiReport: aplicación de BI gratuita basada en Web de LogiXML.
- OpenI: aplicación Web simple orientada al reporting OLAP.
- Palo.
- Pentaho BI Server.
- RapidMiner (antes YALE).
- SpagoBI.

### Comerciales

- Microsoft SQL Server - Suite de Herramienta de BI (Analysis Services, Integration Services y reporting Services).
- Bingo Intelligence (<http://www.bingointelligence.com>).
- Business Objects (SAP company) |Business Objects.
- CA Oblicore Guarantee.
- IBM Cognos.
- CyberQuery.
- Microsoft Excel.
- Synerplus.

### 1.11. Metodología de desarrollo de los AD

Existen varias metodologías para el desarrollo de un AD en el mundo, por solo citar algunas se encuentran las propuestas por Bill Inmon y Ralph Kimball.

Las metodologías propuestas por Inmon y Kimball se imponen ante las demás. Kimball define un enfoque ascendente (*Bottom-up*), es decir, comenzando por construir los MD y luego conformar el AD, todo lo contrario a Inmon el cual defiende un enfoque descendente (*Top-down*), el cual plantea construir el AD primeramente y una vez terminado se construyen los MD a partir del AD. La metodología de Kimball se basa en un modelado dimensional, es decir, no normalizado, mientras que Inmon se basa en conceptos bien conocidos de bases de datos relacionales.

En este caso, para definir la metodología de desarrollo a utilizar en la investigación, se tomó como base la propuesta de Kimball por los siguientes elementos: (12)

- Crea los conceptos de hechos y dimensiones, lo que indudablemente es muy eficaz para la toma de decisiones y proporciona mayor agilidad en el proceso de desarrollo.



- Propone ir construyendo el AD a través de la construcción de los MD departamentales, lo que constituye una buena estrategia y coincide con la división lógica de las empresas, entidades, organismos, etc. Además permite ir presentando resultados parciales a los clientes en cortos plazos.
- Existe abundante documentación sobre la misma y se puede consultar la web a través de los servicios que brindan el grupo creador de la metodología.

A pesar de todas las ventajas que ofrece la utilización de la Metodología de Kimball, esta no era totalmente adaptable a las características del centro y de la producción en la UCI, por lo que solo se decidió utilizarla como guía en el proceso de confección de metodología de desarrollo utilizada en la solución. Entre sus principales desventajas se encuentran: (12)

- No tiene definido un criterio que permita estimar los costos de desarrollo de un Almacén de Datos, basándose en las características de la construcción del mismo.
- Presenta un grupo de roles, pero no explica claramente cuáles son las competencias y responsabilidades de cada uno dentro del proyecto. Por la cantidad de roles que propone se necesita de grupos grandes para su desarrollo.
- Propone un gran número de actividades y artefactos que pueden extender los tiempos de desarrollo si se cuenta con pocos recursos humanos, además no se especifica cómo deben realizarse estos artefactos.
- Está estructurada para el desarrollo de proyectos – productos, donde un proyecto desarrolla un producto determinado.
- No establece el análisis de diferentes criterios de diseño en el levantamiento de requisitos que permita la construcción más adecuada del almacén, teniendo en cuenta las metas de la organización, las necesidades de los usuarios y la disponibilidad de las fuentes operaciones.

Por tales motivos se decidió definir una metodología que permita mitigar las desventajas identificadas en la Metodología de Kimball y ajustada a las condiciones y características de producción del centro DATEC y de la UCI.

## Fases

La metodología de desarrollo está dividida en ocho fases. (12)

- **Estudio preliminar y planeación:** Se realiza un estudio minucioso en la entidad cliente. Esto incluye un diagnóstico integral de la organización, con el fin de determinar qué es lo que se desea

construir y qué condiciones existen para el desarrollo y montaje de la misma. Además se llevan a cabo las tareas de planeación del proyecto.

- **Requisitos:** Se realiza el proceso entrevistas al cliente para determinar los requisitos de información. Se hace levantamiento detallado de las fuentes de datos para validar la disponibilidad de la información. Además se definen los requisitos funcionales y no funcionales de la solución y se hace el análisis de los requisitos que dan paso al diseño e implementación.
- **Arquitectura:** Se definen las vistas arquitectónicas de la solución, aspectos como, los subsistemas y componentes, la seguridad, la comunicación y la tecnología a utilizar.
- **Diseño e Implementación:** Se define el diseño de las estructuras de almacenamiento de datos, se diseñan los procesos de integración de datos como, el mapa lógico de datos, los cubos OLAP para la presentación de la información, así como el diseño gráfico de la aplicación definido por el cliente. Después se implementan cada uno de los subsistemas (repositorio de datos, integración de datos, presentación de datos).
- **Prueba:** Se realizan las pruebas que validan la calidad del producto, comenzando por las Pruebas de Unidad llevadas, las Pruebas de Integración y Sistema, hasta llegar a las Pruebas de Aceptación con el cliente final. Esta fase no es la única en la que se realizan pruebas durante el desarrollo del proyecto, en todas las fases hay actividades de aseguramiento de la calidad.
- **Despliegue:** Consta de dos etapas, despliegue piloto, donde se configuran los servidores necesarios y se instalan las herramientas según la arquitectura definida, se cargan una muestra de los datos en un ambiente controlado, con el fin de mostrarle al cliente final el sistema en funcionamiento. Una vez aceptada la solución por el cliente, se realiza la carga histórica de los datos, puede ser en el mismo entorno que el despliegue piloto u otro, todo depende de las condiciones que establezca el cliente. Además se realiza la capacitación y transferencia tecnológica de la solución a los clientes. El resultado fundamental es la solución desplegada en el entorno real y en correcto funcionamiento.
- **Soporte y Mantenimiento:** Comienza cuando la solución está implantada y en explotación, y se ejecuta según el contrato firmado y las condiciones de soporte establecidas. Puede realizarse a través de variados servicios, que pueden ser soporte en línea, vía telefónica, web, correo u otros y el acompañamiento al cliente. Además se realizan las tareas de manteniendo de la aplicación tan necesarias para este tipo de desarrollo y que garantiza el adecuado funcionamiento y crecimiento del almacén de datos.

- **Gestión del proyecto:** Esta fase se ejecuta a lo largo de todo el ciclo de vida del proyecto. Es aquí donde se controla, gestiona y chequea todo el desarrollo, los gastos, las utilidades, los recursos, las adquisiciones, los planes y cronogramas entre otras actividades relacionadas con la gestión de proyectos. Esta fase es la columna vertebral del proyecto y si no se ejecuta de forma continua y correcta el proyecto puede fracasar.

## 1.12. Herramientas utilizadas en la investigación

En el desarrollo de la presente investigación se utilizarán un conjunto de herramientas que se explicarán a continuación.

### Herramientas de modelado

Las herramientas de modelados son útiles en la modelación de proyectos, entidades, creación de flujos de información, etc. Son nombradas herramientas *CASE* (Ingeniería de Software Asistida por Computadora) y con ellas se pueden diseñar los casos de usos, diseñar el modelo lógico, obtener el físico y generar el script de la base de datos. La herramienta usada se explicará a continuación.

#### Visual Paradigm 8.0

Visual Paradigm es una herramienta *CASE*. Propicia un conjunto de ayudas para el desarrollo de programas informáticos, desde la planificación, pasando por el análisis y el diseño, hasta la generación del código fuente de los programas y la documentación (15). Entre las disímiles **características** que tiene se muestran las siguientes: (13)

- Disponibilidad en múltiples plataformas.
- Diseño centrado en casos de uso y enfocado al negocio que generan un software de mayor calidad.
- Capacidades de ingeniería directa e inversa.
- Licencia gratuita y comercial.
- Soporta aplicaciones Web.

Además como **ventajas** tiene que: (13)

- Es fácil de usar, instalar y actualizar.
- Es una herramienta de modelado UML profesional que soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue.

## Herramienta de desarrollo

Las herramientas de desarrollo permiten al usuario la confección de los distintos softwares para empresas o para el consumo propio. Existen varios tipos de herramientas pero en el desarrollo de la investigación se explicarán solo las que vienen a continuación.

### ✚ Sistemas gestores de bases de datos

Hoy en día las grandes empresas y distintas organizaciones presentan problemas a la hora de gestionar grandes cantidades de datos, evidenciándose la necesidad de toma de decisiones más rápidas, como también el aumento de la productividad y mejoras en el desarrollo, presión ejercida por el amplio auge que va teniendo la informatización en nuestros días, para dar solución a estos problemas se han creado distintas herramientas como las que se muestran a continuación.

### ✚ PostgreSQL 9.1

PostgreSQL es el líder en sistemas de bases de datos de código abierto, con una comunidad mundial de miles de usuario contribuyentes, docenas de empresas y organizaciones. El proyecto PostgreSQL se basa en 25 años de ingeniería, a partir de la Universidad de California, Berkeley, y tiene un ritmo sin precedentes de desarrollo en la actualidad. El conjunto de características maduras que tiene PostgreSQL no sólo rivaliza con sistemas de bases de datos propietarias, sino que los supera en características avanzadas, extensibilidad, seguridad y estabilidad. (14)

Esta versión trae muchas mejoras en cuanto a las anteriores, ofreciendo tecnología innovadora, extensibilidad sin igual, y nuevas características, entre las que se encuentran: (14)

- **Replicación sincrónica:** permitiendo alta disponibilidad con consistencia sobre múltiples servidores.
- **Regionalización por columna:** soportando correctamente el ordenamiento por lenguaje en las bases de datos, tablas o columnas.
- **Tablas unlogged:** importante incremento del rendimiento para datos efímeros.
- **Indexamiento de los K vecinos más próximos (K-Nearest-Neighbor):** índices sobre "distancia" para consultas rápidas de ubicación y búsquedas de texto.
- **Nivel de aislamiento serializable a través de "Snapshots":** mantiene consistentes múltiples transacciones concurrentes sin el uso de bloqueos, usando "verdadera serialización".
- **Writeable common table expressions:** ejecuta actualizaciones multi-fases complejas en una simple consulta.

- **Security-enhanced postgres:** despliega seguridad de nivel militar y control de acceso mandatorio.

#### **Ventajas del uso de PostgreSQL (15)**

- Instalación ilimitada.
- Brinda soporte.
- Ahorros considerables en costos de operación.
- Estabilidad y confiabilidad legendaria.
- Extensible.
- Multiplataforma.
- Diseñado para ambientes de alto volumen.

#### **PgAdmin III 1.14**

PgAdmin III es una aplicación gráfica para gestionar el SGBD (Sistema Gestor de Bases de Datos) PostgreSQL, siendo la más completa y popular con licencia open source. Es capaz de gestionar versiones a partir de la PostgreSQL 7.3 ejecutándose en cualquier plataforma. Está diseñado para responder a las necesidades de todos los usuarios, desde escribir consultas sql simples hasta desarrollar bases de datos complejas. (16)

Entre sus principales **características** se tienen: (16)

- Multiplataforma.
- Amplia documentación.
- Acceso a los datos.
- Acceso a todos los objetos de PostgreSQL.
- Diseñado para múltiples versiones de PostgreSQL.

#### **Herramienta de perfilado de datos**

El perfilado de datos es indispensable en el análisis de los datos de las fuentes de origen, ya que en muchas ocasiones estas no se encuentran de una forma más deseadas o entendibles. Es por eso que se realiza con el objetivo de entender más las fuentes de datos, su estructura y calidad. También se utilizará después de la carga de los datos para revisar que tengan la calidad requerida. El perfilado se puede realizar a través de distintas herramientas como el Microsoft excel, las consultas a la BD en el PostgreSQL y visualmente, en la presente investigación se utilizó el DataCleaner por las siguientes razones.

#### **DataCleaner 1.5.4**

DataCleaner es una aplicación de código abierto, simple, fácil de utilizar, diseñado para ayudar a perfilar, comparar, validar y supervisar los datos. Permite también perfilar la base de datos dentro de minutos y es la versión utilizada por el departamento para dichas funciones.

## Herramienta para la implementación de procesos ETL

Las herramientas para los procesos ETL son las encargadas de realizar el proceso que organiza el flujo de los datos entre diferentes sistemas en una organización, herramientas necesarias para mover datos desde múltiples fuentes a un AD, transformarlos, limpiarlos y cargarlos en otra base de datos. Existen una gran cantidad de herramientas como el Talend, Microsoft SQL Server Integration Services, Informatics Power Center, BusinessObjects Data Integrador y el Pentaho Data Integration, el cual fue utilizado en la investigación por las razones siguientes

### Pentaho Data Integration (PDI) 4.2

Con el incremento de la variedad y velocidad de los datos, las organizaciones necesitan formas más rápidas y fáciles de aprovechar los datos y obtener información. PDI proporciona la solución ideal para cualquier tipo de integración de datos, análisis de negocio o proyecto de datos grande. (17)

PDI es un intuitivo y rico diseñador que permite hacer exactamente lo que los desarrolladores de código más expertos pueden lograr en una fracción del tiempo, y sin necesidad de codificar manualmente (17).

Entre sus principales **características** (17) se tiene que es una herramienta multiplataforma, fácil de instalar y configurar. Usa tecnologías estándar como Java, XML, Java Script.

Además esta incluye:

- Amplia fuente de datos de apoyo, incluyendo aplicaciones empaquetadas, más de 30 plataformas de código abierto y propietario de base de datos, archivos planos y documentos de Excel.
- Unificado de ETL, modelado y visualización de entorno de desarrollo para el diseño de aplicaciones de BI.

## Herramientas de BI

Estas herramientas están diseñadas para ayudar o colaborar con la Inteligencia de Negocios. Su trabajo en específico es asistir al análisis y presentación de los datos que se mostrarán en los reportes. Vale resaltar que aunque algunas herramientas poseen funcionalidades de ETL, las herramientas ETL no son consideradas generalmente como herramientas de BI. Las herramientas BI utilizadas se describen a continuación:

### **Pentaho Schema Workbench 3.3**

Pentaho Schema Workbench también conocido como Mondrian Schema Workbench es un diseñador de interfaz que permite crear y probar esquemas OLAP de Mondrian visualmente. El motor Mondrian procesa las solicitudes de MDX con los esquemas ROLAP (18). Es una aplicación creada en java y además multiplataforma, orientada al desarrollador conocedor de la estructura de un esquema de Mondrian. Una de las principales herramientas de BI para el desarrollo de un AD.

### **Pentaho BI Server 3.10.0**

La aplicación más conocida de la Plataforma de BI es el Pentaho BI server. Funciona como un sistema de gestión basado en el informe en la web, el servidor de integración de aplicaciones y el motor de flujo de trabajo ligero (secuencias de acción). Está diseñado para integrarse fácilmente en cualquier proceso de negocio. Proporciona el servidor y plataforma web del usuario final. Podrá interactuar con la solución BI previamente creada con las herramientas anteriormente comentadas. (19)

## **1.13. Conclusiones parciales**

En este capítulo se realizó un estudio detallado de los elementos principales a tener en cuenta en la elaboración de un AD, determinando que:

- ✓ La investigación detallada de las metodologías existentes en el mundo para la confección de un AD, permitió seleccionar la que más se acerca al trabajo realizado en el centro DATEC.
- ✓ Se realizó un estudio de las distintas herramientas y sus principales características, para escoger las más adecuadas teniendo en cuenta las características de la investigación.
- ✓ Se analizaron los principales conceptos y tecnologías para la confección de un AD permitiendo escoger los que se consideraron más adecuados.

## Capítulo 2: Análisis y diseño del mercado de datos Plan Turquino

### Introducción

En este capítulo se abordará todo lo referente al proceso de análisis y diseño del mercado de datos, donde se hará un estudio preliminar del negocio para facilitar una mejor comprensión del desarrollo de la solución. También se definirán los requisitos de información, funcionales, no funcionales, así como, las reglas del negocio. Se realizarán las descripciones de los casos de usos, se desarrollará la matriz BUS, además se identificarán los hechos, medidas y dimensiones. Se definirán los reportes candidatos y el diseño del proceso de integración.

### 2.1. Definición del negocio

La ONEI es la institución rectora de la estadística en nuestro país, en ella se recoge toda la información de los indicadores del sector socioeconómico del país. Estos datos no se encuentran integrados de una manera asequible, por lo que con el objetivo de mejorar el trabajo de los especialistas se hace necesario crear mercados de datos referentes a un área temática específica del almacén. Existen diferentes modelos que guardan esta información según su tipo, con una frecuencia (anual, semestral y trimestral). En el modelo 0024-03 se recoge la información de los datos estadísticos referente al Plan Turquino de la siguiente forma:

- ✓ Las Direcciones de Arquitectura y Urbanismo (DAU) brindan los datos de viviendas y asentamientos poblacionales.
- ✓ Las unidades presupuestadas de Comunes a nivel municipal brindaran la información de las plantas eléctricas.
- ✓ Las Unidades Empresariales de Base de Hidroenergía a nivel provincial pertenecientes a la UNE del MINBAS brindarán la información de las hidroeléctricas pertenecientes al Plan Turquino, desagregada por municipios a la ONEI provincial correspondiente, la que entrega a sus oficinas municipales.
- ✓ Las empresas o establecimientos provinciales de energía eléctrica y las comunicaciones deben ofrecer la información de sus actividades respectivas correspondientes al Plan Turquino, desagregada por municipios a la ONEI provincial correspondiente, la que entrega a sus oficinas municipales.
- ✓ Para el caso del sector no estatal, la información se obtiene de los modelos del Sistema de Información Estadístico Nacional (SIEN), así como, de las empresas del Ministerio de la Agricultura por las que son atendidos.



- ✓ Las informaciones que se brinden referentes al Plan Turquino tienen que estar enmarcadas dentro de los límites establecidos por el Instituto de Planificación Física en el año 1999.

Con la inclusión de la ficha de datos anual del Plan Turquino se complementará la información recogida en la anterior versión del mercado, incluyendo los indicadores que caracterizan el desarrollo social de estas zonas del país. Esta representa un anexo al modelo 0024-03, recogiéndose los datos estadísticos de la misma manera.

## 2.2. Especificación de requisitos

El análisis detallado de requisitos contribuye a una correcta elaboración de un mercado de datos, además de guiar el proceso de desarrollo hacia la construcción de un sistema correcto. Una buena descripción de los requisitos posibilita llegar a acuerdos y resultados de las especificaciones que debe cumplir el sistema. En dicha fase se definen los requisitos de información, funcionales y no funcionales del sistema, partiendo de las necesidades de los clientes. Se trabaja conjuntamente con los usuarios finales permitiendo que estas especificaciones de la aplicación sean descritas por el personal de mayor conocimiento de los procesos que se llevan a cabo en la organización. De todo el correcto cumplimiento de estos procesos de análisis depende la correcta implementación del sistema.

### 2.2.1. Requisitos de información

Los requisitos de información (RI) permiten describir toda la información que se debe almacenar en el sistema para satisfacer las necesidades de los usuarios finales, además identifican los conceptos relevantes sobre los que se debe almacenar la información y los datos específicos de interés. A continuación se muestran los identificados en el análisis de la investigación:

**RI1.** Obtener la cantidad total de Km<sup>2</sup> del país por división político administrativa (dpa), tema y año.

**RI2.** Obtener la cantidad de Km<sup>2</sup> de turquino por dpa, tema y año.

**RI3.** Obtener la cantidad total de habitantes del país por dpa, tema y año.

**RI4.** Obtener la cantidad de habitantes de turquino por dpa, tema y año.

**RI5.** Obtener la cantidad de habitantes varones por dpa, tema y año.

**RI6.** Obtener la cantidad de habitantes hembras por dpa, tema y año.

**RI7.** Obtener la cantidad de habitantes ancianos por dpa, tema y año.

**RI8.** Obtener la cantidad de habitantes niños por dpa, tema y año.

**RI9.** Obtener la cantidad total de consejos populares por dpa, tema y año.

**RI10.** Obtener la cantidad de consejos populares mixtos por dpa, tema y año.

**RI11.** Obtener la cantidad total de asentamientos poblacionales por dpa, tema y año.

**RI12.** Obtener la cantidad de asentamientos poblacionales conectados telefónicamente por dpa, tema y año.

**RI13.** Obtener la cantidad de asentamientos poblacionales electrificados por dpa, tema y año.

**RI14.** Obtener la cantidad de asentamientos poblacionales con acueducto por dpa, tema y año.

**RI15.** Obtener el total de entidades existentes por dpa, tema y año.

**RI16.** Obtener el total de entidades existentes electrificadas por dpa, tema y año.

**RI17.** Obtener el porcentaje que representa el total de viviendas o centros electrificados del total existente por dpa, tema y año.

**RI18.** Obtener la cantidad de entidades existentes electrificadas por el SEN por dpa, tema y año.

**RI19.** Obtener el porcentaje entidades existentes electrificadas por el SEN por dpa, tema y año.

**RI20.** Obtener la cantidad entidades existentes electrificadas por mini hidroeléctricas por dpa, tema y año.

El resto de los requisitos de información se encuentran detallados en el artefacto " DATEC-SIGOB\_PT-0113\_Especificación de requisitos de software v2.0 ".

### **2.2.2. Requisitos funcionales**

Los requisitos funcionales (RF), permiten describir los casos de uso en los cuales los actores utilizan los servicios proporcionados por el sistema, es decir, son las funcionalidades que brindara el sistema. A continuación se muestran los identificados:

**RF1.** Autenticar usuario.

**RF2.** Adicionar rol.

**RF3.** Eliminar rol.

**RF4.** Mostrar rol.

**RF5.** Adicionar usuario.

**RF6.** Eliminar usuario.

**RF7.** Mostrar usuario.

**RF8.** Adicionar reporte.

**RF9.** Mostrar reporte.

**RF10.** Eliminar reporte

**RF11.** Modificar reporte.

**RF12.** Realizar extracción de los datos de los ficheros fuente.

**RF13.** Realizar transformación y carga de los datos al mercado plan turquino.

**RF14.** Obtener información almacenada de los últimos 5 años.

**RF15.** Obtener los datos de la fuente, en un formato ".xls".

**RF16.** Crear metadatos asociados a la organización.

**RF17.** Exportar datos a PDF y a Excel.

### 2.2.3. Requisitos no funcionales

Los requisitos no funcionales (RNF) no son más que las propiedades o cualidades que el sistema debe cumplir, determinan como se comportará el producto una vez finalizado. A continuación se muestran los identificados durante la investigación.

#### ✓ **Requisitos de usabilidad**

**RNF1.** Los usuarios deben recibir capacitación de las herramientas utilizadas.

#### ✓ **Requisitos de fiabilidad**

##### **Disponibilidad**

**RNF2.** El sistema debe tener disponibilidad al 100% entre las 8:00 am y las 5:00 pm en los días laborables (lunes a viernes).

**RNF3.** Asegurar que el sistema sea capaz de recuperarse ante un fallo

##### **Exactitud**

**RNF4.** La salida de los datos debe ser 100% exacta, ya que de ahí dependen las decisiones que se tomen basándose en los mismos.

#### ✓ **Requisitos de eficiencia**

##### **Tiempo de respuesta**

**RNF5.** El sistema debe tener un tiempo de respuesta aproximado de 1 minuto, no debe excederse de 2 minutos.

##### **Capacidad**

**RNF6.** Durante la integración solo se conectará un usuario que deberá monitorear el proceso de integración.

**RNF7.** El sistema deberá permitir varios usuarios (mín. de 5) conectados sin que esto afecte la respuesta de las consultas.

#### ✓ **Restricciones de diseño**

**RNF8.** Lograr similitud en los datos presentes en el negocio

**RNF9.** El SGBD que se utilizará será PostgreSQL versión 9.1

#### ✓ **Requisitos para la documentación de usuarios en línea y ayuda del sistema.**

**RNF10.** Se elaborará un manual de usuario que guiará la ejecución del usuario teniendo en cuenta cada funcionalidad.

#### ✓ **Componentes comprados**

**RNF11.** La herramienta CASE Visual Paradigm no es un software libre pero se cuenta con una licencia adquirida previamente por la UCI.

✓ **Interfaces de usuario**

**RNF12.** El sistema debe tener una interfaz amigable y sencilla de utilizar, teniendo en cuenta que los usuarios finales no son personas adiestradas en el campo de la informática.

**RNF13.** Las interfaces de salida no deben tener otra información que no sea concerniente a lo que deben mostrar.

✓ **Interfaces hardware**

**RNF14.** Para el proceso de transformación es necesaria una memoria de 512 MB como mínimo.

**RNF15.** Para el proceso de visualización e inteligencia de negocio se necesita una memoria de 1 GB como mínimo, para garantizar el correcto funcionamiento del sistema cuando es accedido por varios usuarios simultáneamente.

**RNF16.** Se necesita un mínimo de 60 GB para el almacenamiento de la información.

**RNF17.** Se debe garantizar al menos una impresora, para imprimir los reportes de salida.

✓ **Interfaces de software**

**RNF18.** El lenguaje para realizar las consultas y la programación dentro del repositorio será el SQL.

**RNF19.** Se debe disponer de un navegador común, preferentemente asociado al sistema operativo, para garantizar la visualización de las interfaces web de los reportes.

**RNF20.** La base de datos se acoplará con el gestor PostgreSQL 9.1

✓ **Interfaces de comunicación**

**RNF21.** El sistema debe estar conectado a un dispositivo de red.

**RNF22.** La comunicación entre el almacén de datos y la base de datos de la integración será a través del protocolo TCP/IP.

✓ **Requisitos de licencia**

**RNF23.** De las herramientas utilizadas, Visual Paradigm 8.0 es la única herramienta que no es libre, pero se cuenta con su licencia previamente adquirida por la universidad.

### 2.3. Casos de uso del sistema

Un caso de uso (CU) es una secuencia de interacciones que se desarrollarán entre un sistema y sus actores en respuesta a un evento que inicia un actor principal sobre el propio sistema. Es además la técnica para la captura de requisitos potenciales de un nuevo sistema o una actualización de software. Los diagramas de casos de uso sirven para especificar la comunicación y el comportamiento de un sistema

mediante su interacción con los usuarios y/u otros sistemas. O lo que es igual, un diagrama que muestra la relación entre los actores y los casos de uso en un sistema. (20)

### 2.3.1. Actores del sistema

Actor	Descripción
Administrador	El administrador interactúa con el sistema para realizar todas las operaciones de administración de roles, usuarios y reportes.
Administrador ETL	El administrador es el encargado de realizar los procesos de extracción, transformación y carga de los datos.
Analista	El analista es el encargado de analizar y consultar la información.

Tabla 1: Descripción de los actores del sistema.

### 2.3.2. Casos de uso de información

Los casos de uso informativos describen los requisitos de información agrupados según el tema de análisis. A continuación se describe cada caso de uso de información (CUI) definido para el MD Plan Turquino:

**CUI1. Mostrar información de Población:** muestra los reportes de los indicadores seleccionados de población.

**CUI2. Mostrar información de Electricidad:** muestra los reportes de los indicadores seleccionados de electricidad.

**CUI3. Mostrar información de Viviendas:** muestra los reportes de los indicadores seleccionados de viviendas.

**CUI4. Mostrar información de Transporte:** muestra los reportes de los indicadores seleccionados de transporte.

**CUI5. Mostrar información de Salud:** muestra los reportes de los indicadores seleccionados de matrícula inicial.

**CUI6. Mostrar información de Educación:** muestra los reportes de los indicadores seleccionados de educación.

**CUI7. Mostrar información de Centros:** muestra los reportes de los indicadores seleccionados de centros.

**CUI8. Mostrar información de Embalses:** muestra los reportes de los indicadores seleccionados de embalses.

**CUI9. Mostrar información de Cuencas:** muestra los reportes de los indicadores seleccionados de cuencas.

**CUI10. Mostrar información de Áreas Protegidas:** muestra los reportes de los indicadores seleccionados de áreas protegidas.

**CUI11. Mostrar información de Comunicación:** muestra los reportes de los indicadores seleccionados de comunicación.

**CUI12. Mostrar información de Economía:** muestra los reportes de los indicadores seleccionados de economía.

En la tabla 2 se presenta la especificación del CUI "Mostrar información de población", el resto de los casos de usos se encuentran detallados en el artefacto "DATEC-SIGOB\_PT0114\_Especificación de casos de uso v2.0".

<b>Objetivo</b>	Visualizar los reportes de los indicadores seleccionados de población.
<b>Actores</b>	Analista , Administrador
<b>Resumen</b>	El caso de uso inicia cuando el actor desea hacer un análisis de la información relacionada con la población sin obviar ningún detalle. El actor selecciona el reporte que desea ver, el sistema muestra la información contenida en él. El caso de uso finaliza cuando el actor termina el análisis de la información relacionada con la población.
<b>Complejidad</b>	Alta
<b>Prioridad</b>	Crítico
<b>Precondiciones</b>	El usuario se autenticó correctamente. Los datos correspondientes fueron cargados en el mercado de datos. Los reportes relacionados la información de la población fueron creados
<b>Postcondicio</b>	Los reportes correspondientes al caso de uso fueron consultados.

nes		
<b>Flujo de eventos</b>		
<b>Flujo básico Visualizar los reportes de los indicadores seleccionados de población</b>		
	<b>Actor</b>	<b>Sistema</b>
1.	Selecciona el área de información a la población	
2.		Muestra las áreas de información de la población
3.	Selecciona la información que desea consultar.	
4.		Muestra los reportes contenidos en él.
5.	Selecciona el reporte que desea analizar	
6.		Muestra la información contenida en el reporte seleccionado y brinda opciones al actor para visualizar los reportes durante su análisis. Ir al CU Visualizar reportes. Finaliza el CU.
<b>Flujos alternos</b>		
2ª. Los datos son incorrectos.		
	<b>Actor</b>	<b>Sistema</b>
		Muestra un mensaje de error. Vuelve al paso 1 del flujo básico.
<b>Relaciones</b>	<b>CU Incluidos</b>	No aplica
	<b>CU Extendidos</b>	Mostrar visualizaciones sobre los reportes: Paso 6 del Flujo Básico. Mostrar visualizaciones sobre los reportes en el CU Presentar información relacionada con la población.
<b>Requisitos no funcionales</b>	RNF1, RNF2, RNF3, RNF4, RNF5, RNF6, RNF7, RNF8, RNF9, RNF10, RNF11, RNF12, RNF13, RNF14, RNF15, RNF16, RNF17,	

	RNF18, RNF19, RNF20, RNF21, RNF22, RNF23
<b>Asuntos pendientes</b>	

Tabla 2: CU Mostrar información de Población.

### 2.3.3. Casos de uso funcionales

Los casos de usos funcionales (CUF) se basan en la gestión de los roles, reportes, usuarios y la autenticación de los usuarios, además de las consultas en la base de datos, a continuación se describen algunos:

**CUF1. Autenticar usuario:** se realiza la autenticación de los usuarios en el sistema.

**CUF2. Gestionar roles:** permite adicionar, modificar, eliminar y mostrar roles.

**CUF3. Gestionar usuarios:** permite adicionar, modificar, eliminar y mostrar usuarios.

**CUF4. Gestionar reportes:** permite adicionar, modificar, eliminar y mostrar reportes.

**CUF5. Extraer los datos de la fuente:** se realiza la extracción de los datos de las fuentes.

**CUF6. Realizar la transformación y carga de los datos:** se realiza la transformación y carga de los datos necesarios para la construcción del Mercado de datos Plan Turquino.

### 2.3.4. Diagrama de casos de uso

El diagrama de casos de uso (CU) se realiza con el objetivo de lograr una mejor comprensión del funcionamiento del sistema, en el se evidencian las relaciones entre los casos de uso y sus actores, proporcionando una visión general del funcionamiento de cada usuario que interactúa con el sistema.



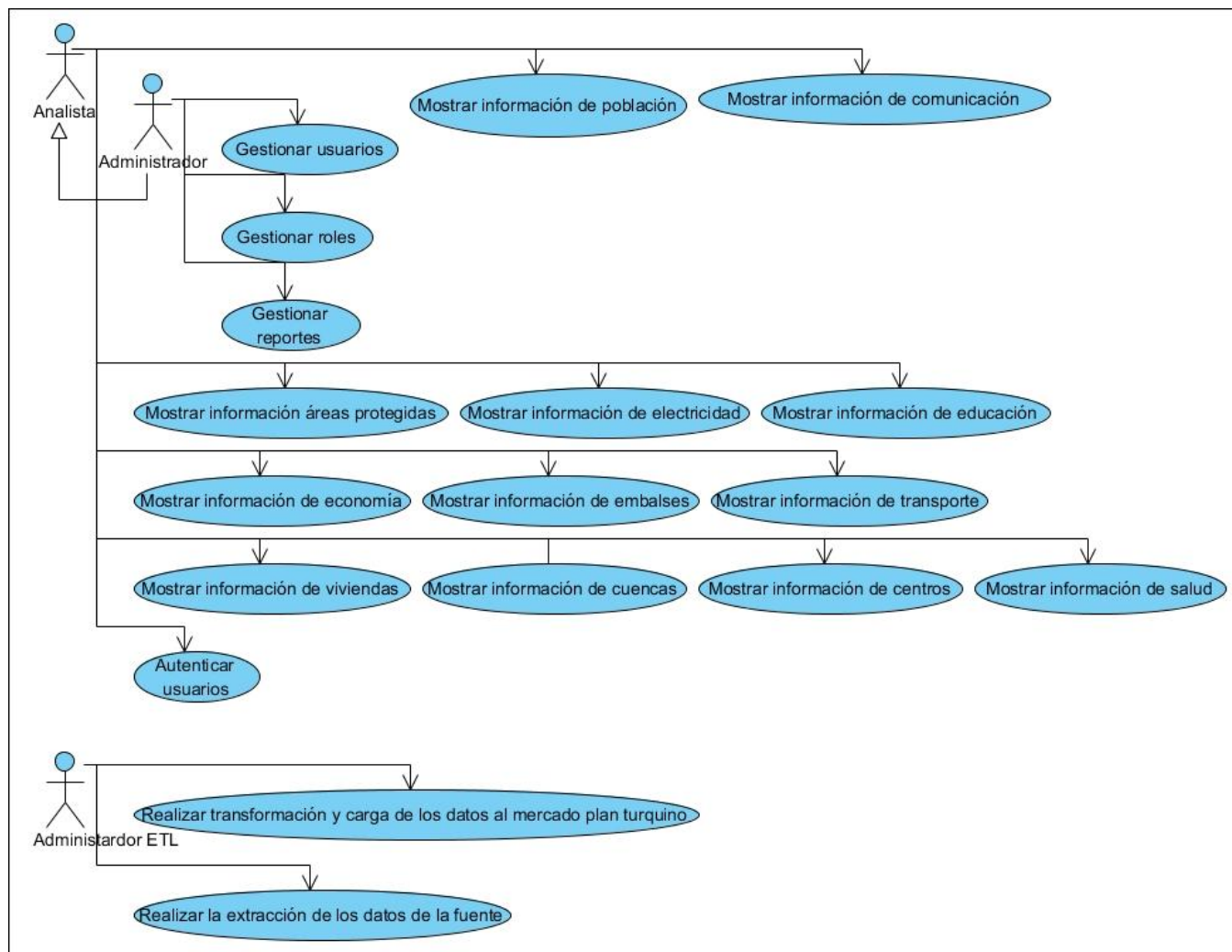


Figura 1: Diagrama de casos de uso.

## 2.4. Reglas del negocio

Las Reglas del Negocio (RN) se basan en la extracción de información originada de las políticas, reglas y regulaciones del negocio y de la descripción del flujo. Estas se clasifican en:

- ❖ **Reglas de variables:** Son las reglas que definen las variables calculables que son objeto de análisis. Ejemplo: porcentaje de  $x = x*y/100$ .
- ❖ **Reglas de almacenamiento:** Son las reglas que definen características específicas del almacenamiento de alguna variable. Ejemplo: tipo de datos, cantidad de caracteres, entre otros.
- ❖ **Reglas de transformación:** Son las reglas que implican la transformación de alguna variable durante los procesos de integración de datos. Ejemplo: femenino o masculino = F o M.

- ❖ **Reglas de visualización:** Son las reglas que implican alguna condición para la visualización de alguna variable. Ejemplo: las variables de tipo float deben visualizarse solo con dos espacios después de la coma.

En la investigación se identificaron las siguientes reglas:

### Reglas de transformación

**RN1.** Los valores en la fuente representados por el carácter "-" serán cambiados por valores nulos.

### Reglas de variables

**RN2.** El porcentaje que representa el total de viviendas existentes electrificadas del total de viviendas existentes se calcula de la forma:  $(\text{total de viviendas existentes electrificadas} / \text{total de viviendas existentes}) * 100$ .

**RN3.** El porcentaje que representa las viviendas existentes electrificadas por el SEN del total de viviendas existentes electrificadas se calcula de la forma:  $(\text{viviendas existentes electrificadas por el SEN} / \text{total de viviendas existentes}) * 100$ .

**RN4.** El porcentaje que representa las viviendas existentes electrificadas por minihidroeléctricas del total de viviendas existentes electrificadas se calcula de la forma:  $(\text{viviendas existentes electrificadas por minihidroeléctricas} / \text{total de viviendas existentes}) * 100$ .

**RN5.** El porcentaje que representa las viviendas existentes electrificadas por plantas eléctricas del total de viviendas existentes electrificadas se calcula de la forma:  $(\text{viviendas existentes electrificadas por plantas eléctricas} / \text{total de viviendas existentes}) * 100$ .

**RN6.** El porcentaje que representa las viviendas existentes electrificadas por celdas fotovoltaicas del total de viviendas existentes electrificadas se calcula de la forma:  $(\text{viviendas existentes electrificadas por celdas fotovoltaicas} / \text{total de viviendas existentes}) * 100$ .

**RN7.** El porcentaje que representa las viviendas existentes electrificadas por otras formas del total de viviendas existentes electrificadas se calcula de la forma:  $(\text{viviendas existentes electrificadas por otras formas} / \text{total de viviendas existentes}) * 100$ .

**RN8.** Los porcentajes que se recogen en la pestaña de centros e instalaciones existentes para el total, por el SEN, minihidroeléctricas, por plantas eléctricas, celdas fotovoltaicas y por otras formas se calcula de la misma manera que para la pestaña de viviendas existentes electrificadas.

**RN9.** El porcentaje de electrificación se calcula de la siguiente manera:  $(\text{total de viviendas existentes electrificadas} + \text{total de centros e instalaciones existentes electrificados}) / (\text{total de viviendas existentes} + \text{total de centros e instalaciones existentes}) * 100$ .

**RN10.** El porcentaje de mortalidad de menores de 1 año se calcula de la siguiente forma: (defunciones menores de 1 año/nacidos vivos)\*1000.

**RN11.** El porcentaje de mortalidad de menores de 5 años se calcula de la siguiente forma: (defunciones menores de 5 años/nacidos vivos)\*1000.

**RN12.** El porcentaje de mortalidad materna se calcula de la siguiente forma: (mortalidad materna/nacidos vivos)\*100000.

**RN13.** El porcentaje de retención de educación general y superior se calcula de la siguiente forma:  $((\text{matricula inicial} + \text{matriculas altas}) - \text{matriculas bajas}) / (\text{matricula inicial} + \text{matriculas altas}) * 100$ .

**RN14.** El porcentaje de promoción de educación general y superior se calcula de la siguiente forma:  $(\text{aprobados}) / (\text{matricula inicial} + \text{matriculas altas}) * 100$ .

**RN15.** El índice de alumnos por computadora se calcula de la siguiente manera:  $(\text{matricula facultad de montaña} + \text{matricula de unidades docentes}) / (\text{cantidad de computadoras de facultad de montaña} + \text{cantidad de computadora de unidades docentes})$ .

## 2.5. Arquitectura del MD Plan Turquino

La arquitectura del MD Plan Turquino quedó estructurada de manera tal que se compone por la fuente de datos (0000073172-FICHA DE DATOS ANUAL DEL PLAN TURQUINO) y tres subsistemas bases, cuales son:

- **Subsistema de integración:** encargado de la extracción, limpieza e integración de la información para su posterior carga al MD.
- **Subsistema de almacenamiento:** encargado de almacenar todos los datos del MD en las diferentes tablas de hechos y dimensiones definidas.
- **Subsistema de visualización:** encargado de consultar los datos almacenados en el almacén, con el objetivo de mostrarlos a los usuarios finales en los distintos gráficos y reportes, contribuyendo a la toma de decisiones.

En la figura 2 se muestra como quedó compuesta la arquitectura del MD plan turquino:

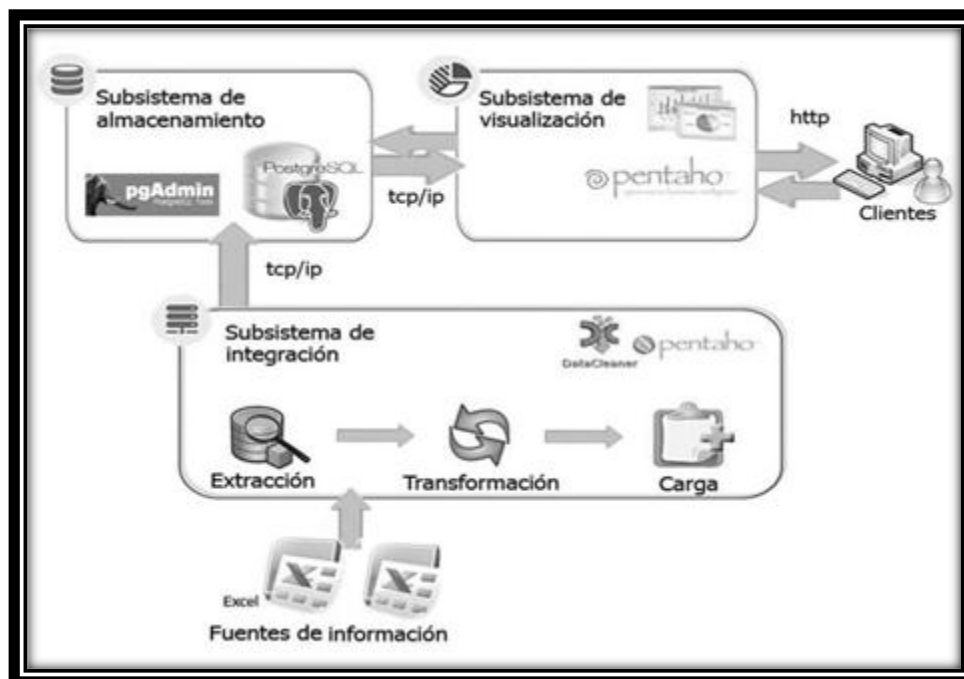


Figura 2: Arquitectura del MD.

### 2.5.1. Diseño del subsistema de almacenamiento

El modelo dimensional de los datos se realiza para el correcto funcionamiento de la solución, el cual contiene las tablas de hechos identificadas en el negocio, las dimensiones y las relaciones que existen entre estos. Para el proceso de almacenamiento se realiza un Procesamiento Analítico Relacional en Línea (ROLAP), ya que el SGBD (PostgreSQL) utilizado no soporta el almacenamiento multidimensional de los datos.

#### Dimensiones

A continuación se muestran las dimensiones identificadas en el estudio de la investigación del MD Plan Turquino.

**Dimensión tema:** esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a extensión territorial y población residente, Consejos Populares y Asentamientos poblacionales, Estado de electrificación, Viales existentes y transporte de pasajeros total.

**Dimensión año:** esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo al año en que se recogió la información de la fuente.

**Dimensión DPA:** esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a la división político administrativa del país.

**jerarquía: dpa**

**Dimensión tipos de centros:** esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los tipos de centros (culturales, deportivos, productivos, etc.) del país.

**jerarquía: centros**

**Dimensión tipos de centros educacionales:** esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los centros educacionales (escuelas primarias, de oficio, facultades de montaña, etc.) del país.

**Dimensión tipos de embalses:** esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los tipos de embalses que se encuentran en esas regiones del país.

**Dimensión tipo de interés:** esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los intereses (nacional, provincial y local).

**Dimensión tipo de comunicación:** esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los tipos de comunicaciones existentes en esas áreas del turquino.

**Dimensión tipo de producción:** esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los tipos de producción mercantil (industria, construcciones, agropecuario, etc.).

## Hechos

A continuación se muestran los hechos identificados en la investigación:

hechos	Descripción
<b>hecho centros educacionales</b>	En este hecho se almacena toda la información relacionada con indicadores referente a los centros educacionales (escuelas primarias, secundarias, preuniversitarios, etc.) del turquino
<b>hecho electricidad y salud</b>	En este hecho se almacena toda la información relacionada con indicadores como salud, electricidad, etc.
<b>hecho temas</b>	En este hecho se almacena toda la información relacionada con indicadores como población, consejos populares y electricidad.
<b>hecho comunicación</b>	En este hecho se almacena toda la información relacionada con indicadores relacionados con los tipos de comunicación existentes en el área del plan turquino
<b>hecho interés</b>	En este hecho se almacena toda la información relacionada con

	indicadores como tipo de interés (nacional, provincial y local)
<b>hecho centro</b>	En este hecho se almacena toda la información relacionada con indicadores relacionados a los tipos de centros existentes en el turquino, dígase culturales, deportivos, centros de producción agropecuarias, etc.
<b>hecho embalses</b>	En este hecho se almacena toda la información relacionada con indicadores relacionados a los embalses existentes en dichas áreas
<b>hecho vivienda</b>	En este hecho se almacena toda la información relacionada con indicadores referentes a la situación de las viviendas en el área plan turquino
<b>hecho producción mercantil</b>	En este hecho se almacena toda la información relacionada con indicadores de la producción mercantil, dígase industria, construcciones, agropecuario, etc.

**Tabla 3: Hechos y descripción.**

### Matriz BUS o matriz dimensional

La matriz bus representa la relación existente entre las dimensiones y los hechos del MD Plan Turquino. Es un medio para verificar que no exista solapamiento entre los hechos. Las columnas de la matriz representan los hechos identificados en el MD y las filas las dimensiones utilizadas. Las celdas marcadas con una X indican que la fila de dimensión está relacionada con la columna del hecho.

D1: dpa

D2: año

D3: tipo\_interés

D4: tipo\_centro

D5: tipo\_centro\_educacional

D6: tipo\_producción

D7: tipo\_embalse

D8: tipo\_comunicacion

D9: tipo\_vivienda

D10: tema

hechos/dimensiones	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
--------------------	----	----	----	----	----	----	----	----	----	-----

<b>hech_centros_educacionales</b>	X	X			X					
<b>hech_electricidad_salud</b>	X	X								
<b>hech_temas</b>	X	X								X
<b>hech_comunicacion</b>	X	X						X		
<b>hech_interes</b>	X	X	X							
<b>hech_centro</b>	X	X		X						
<b>hech_embalses</b>	X	X					X			
<b>hech_vivienda</b>	X	X							X	
<b>hech_produccion_mercantil</b>	X	X				X				

Tabla 4: Matriz BUS.

### Estructura del modelo de datos

Una vez realizado la definición de las dimensiones, hechos y medidas se procede a estructurar el modelo de datos del MD Plan Turquino. Su objetivo es que los datos queden representados de manera lógica, mostrando las relaciones existentes entre los hechos y las dimensiones. A continuación se muestra el modelo estructurado.

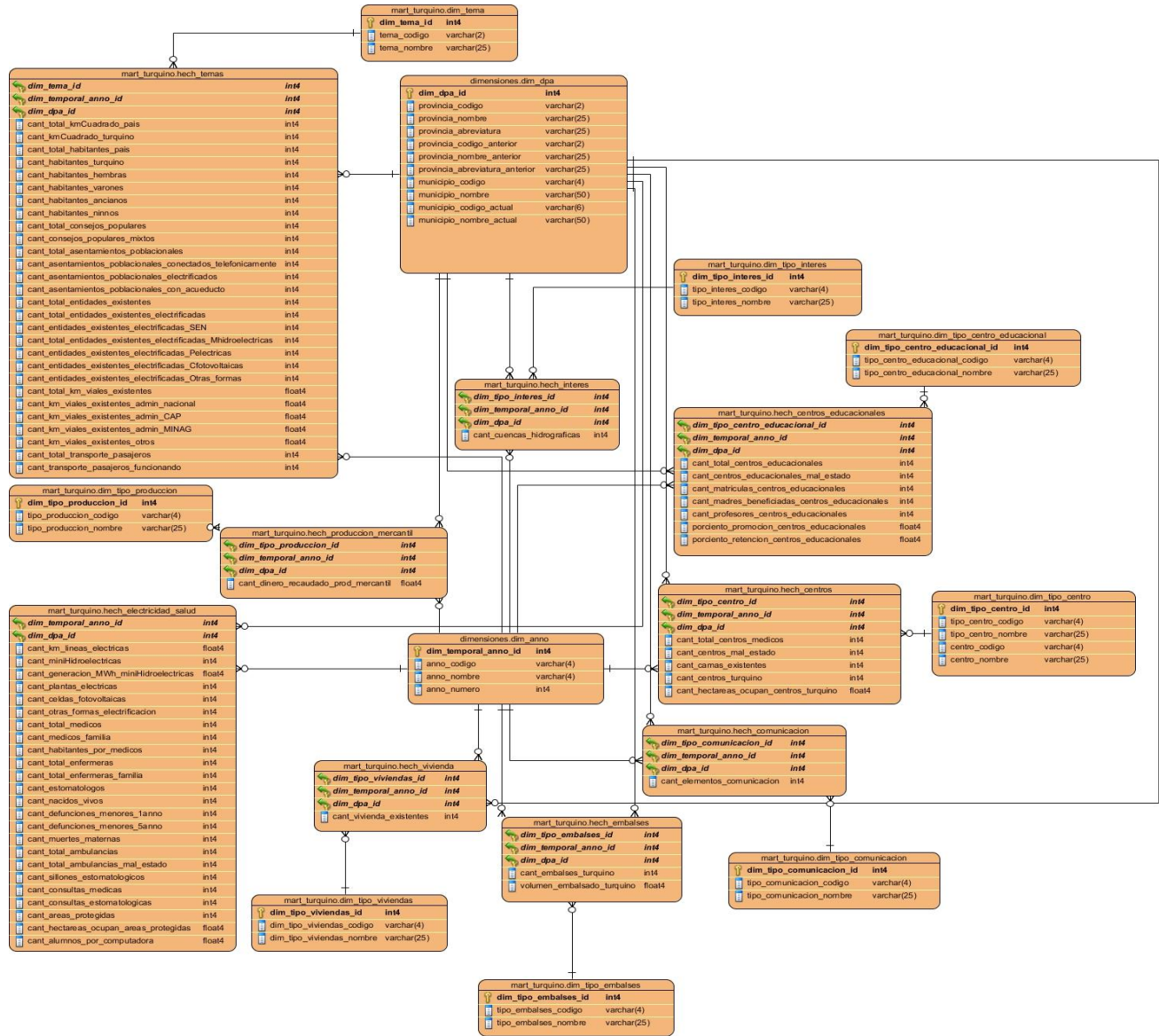


Figura 3: Modelo de datos.

### 2.5.2. Diseño del subsistema de integración

Un buen diseño del subsistema de integración es la base fundamental para la posterior implementación del mismo. Para ello es imprescindible realizar un buen perfilado de los datos para así entender cómo se encuentra el estado de la fuente y realizar una integración de mayor calidad.

#### Perfilado de los datos

El perfilado de los datos se realiza con el objetivo de saber el estado en que se encuentra la fuente, como antes se menciona, además de saber la cantidad de datos nulos, duplicados y con tipo de datos erróneos.

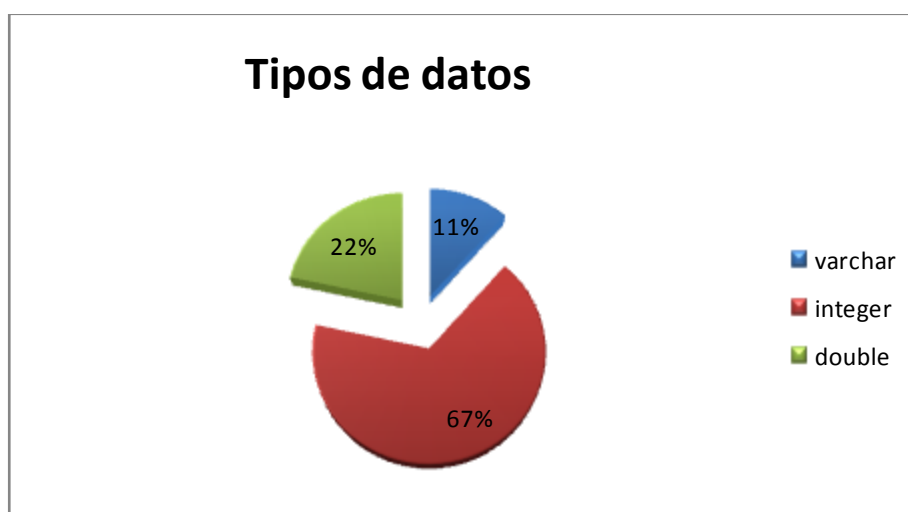


Es aquí donde se definen las distintas reglas de transformación que serán implementadas luego en el proceso ETL.

Al realizar el análisis de los datos contenidos en la fuente se arrojan las siguientes conclusiones:

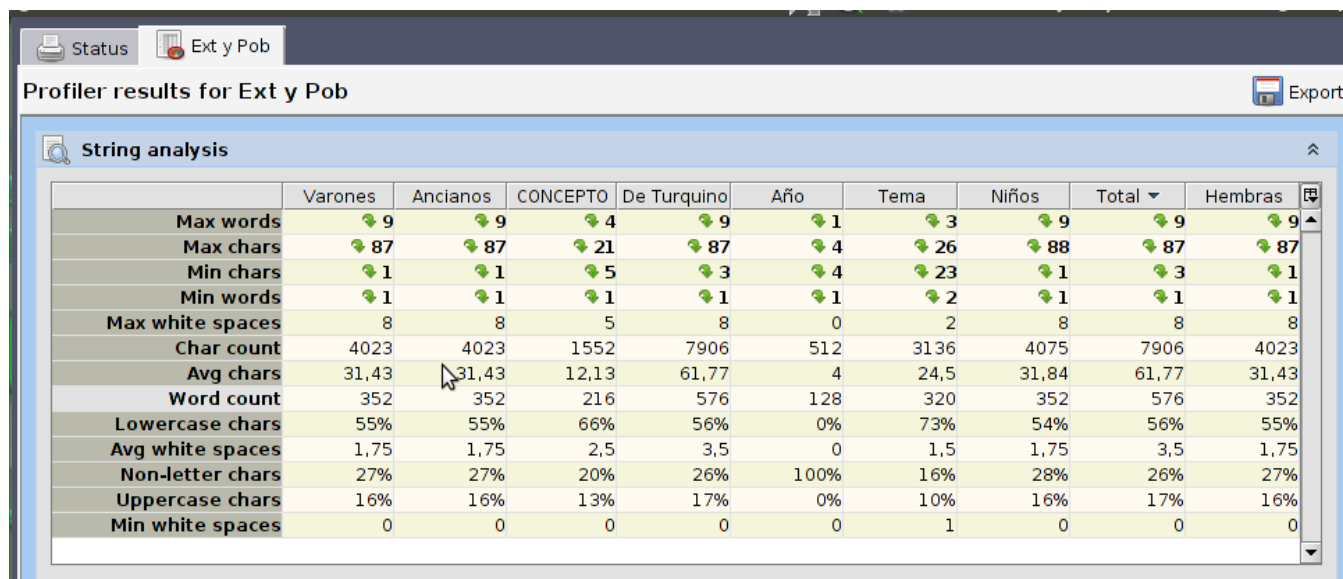
- Los tipos de datos identificados fueron: double, integer y varchar de los cuales el más utilizado es el integer.
- No existen valores negativos.
- No existen valores nulos.
- Los campos integer y double no muestran valor 0, en ese caso se sustituye por el carácter “\_”.

En la siguiente figura se muestra el porcentaje que representa cada tipo de dato del total identificado:



**Figura 4: Representación gráfica de los tipos de datos de la fuente.**

Seguidamente se muestran los resultados del análisis String de la pestaña Ext. y Pob.



	Varones	Ancianos	CONCEPTO	De Turquino	Año	Tema	Niños	Total	Hembras
Max words	9	9	4	9	1	3	9	9	9
Max chars	87	87	21	87	4	26	88	87	87
Min chars	1	1	5	3	4	23	1	3	1
Min words	1	1	1	1	1	2	1	1	1
Max white spaces	8	8	5	8	0	2	8	8	8
Char count	4023	4023	1552	7906	512	3136	4075	7906	4023
Avg chars	31,43	31,43	12,13	61,77	4	24,5	31,84	61,77	31,43
Word count	352	352	216	576	128	320	352	576	352
Lowercase chars	55%	55%	66%	56%	0%	73%	54%	56%	55%
Avg white spaces	1,75	1,75	2,5	3,5	0	1,5	1,75	3,5	1,75
Non-letter chars	27%	27%	20%	26%	100%	16%	28%	26%	27%
Uppercase chars	16%	16%	13%	17%	0%	10%	16%	17%	16%
Min white spaces	0	0	0	0	0	1	0	0	0

Figura 5: Representación del análisis String a una de las pestañas de la fuente.

Para más información del perfilado de los datos de la presente investigación consultar el artefacto " Perfil de los Datos " del expediente de proyecto del MD Plan Turquino.

## Diseño de las transformaciones

Finalizado el diseño del subsistema de almacenamiento y conocida la estructura de los datos se procede a realizar el diseño de las transformaciones del MD Plan Turquino. En las siguientes figuras se muestra como quedó diseñado el hecho centros educacionales y la dimensión tipo de centros educacionales respectivamente (Figuras 6-7).



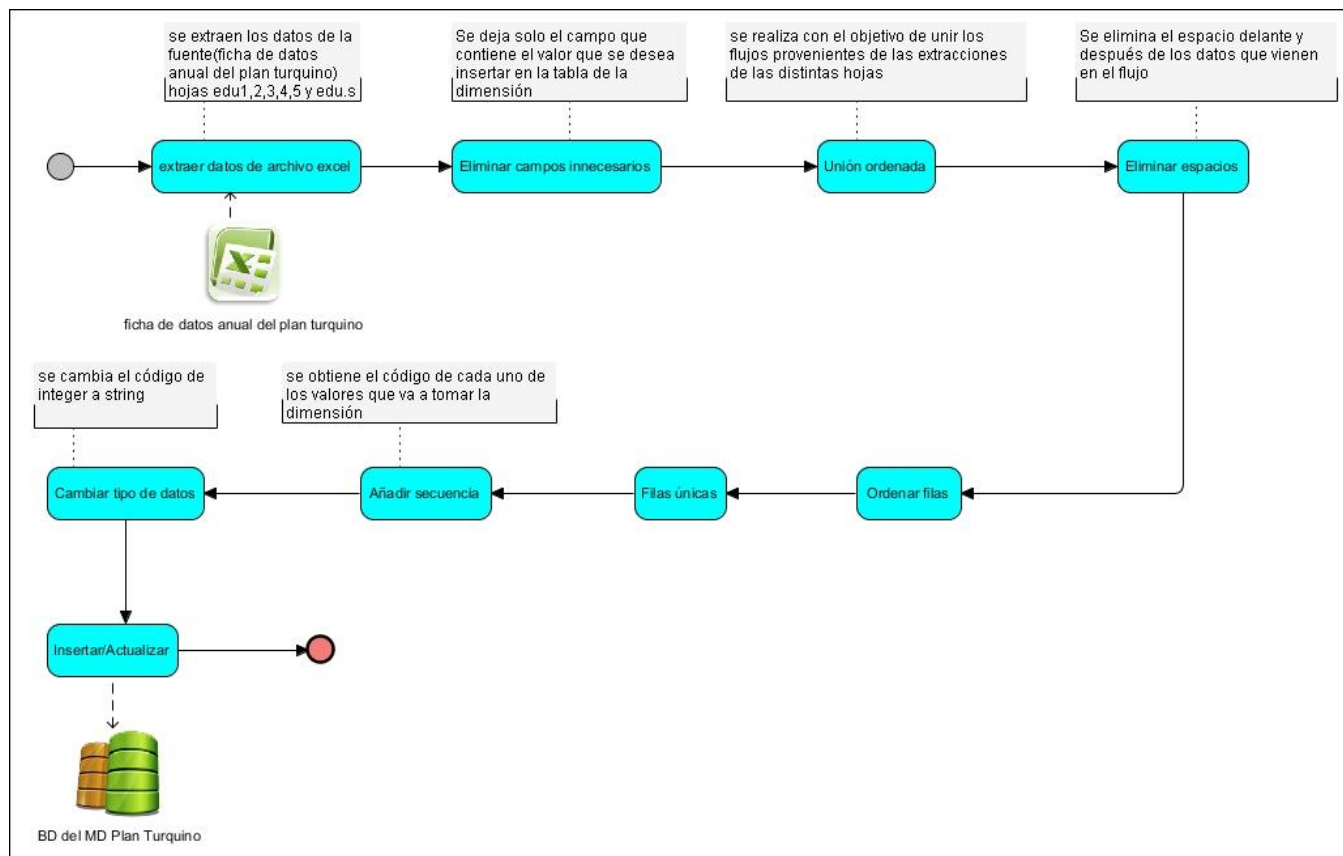


Figura 7: Diseño de las transformaciones (dimensión tipo de centros educacionales).

### 2.5.3. Diseño del subsistema de visualización

El diseño del subsistema de visualización comprende la realización de los cubos OLAP, además de los distintos reportes que contribuyan a la toma de decisiones de los usuarios finales del MD. En estos cubos se definen las dimensiones, medidas y las distintas jerarquías que posean las dimensiones para su posterior publicación en el Pentaho BI Server, el cual proporciona el servidor y plataforma web al usuario final. Este podrá interactuar con la solución BI previamente creada.

### Arquitectura de la información

La arquitectura de la información o mapa de navegación está compuesta según las necesidades de los usuarios finales, por el Área de Análisis General (A.A.G) SIGOB, dentro de la cual se encuentra el Área de Análisis (A.A) Plan turquino, donde están comprendidos doce Libros de Trabajo (L.T) que incluyen los distintos reportes. En la figura 8 se muestra como queda estructurada la navegación para el MD Plan Turquino.



Figura 8: Estructura de navegación.

Seguidamente se describe la información que contienen los distintos L.T:

**L.T Electricidad:** contiene los reportes con los que se logra realizar un análisis general de los datos en función de los indicadores que se ven reflejados en el tema de análisis Electricidad.

**L.T Presas:** contiene los reportes con los que se logra realizar un análisis general de los datos en función de los indicadores que se ven reflejados en el tema de análisis Presas.

**L.T Población:** contiene los reportes con los que se logra realizar un análisis general de los datos en función de los indicadores que se ven reflejados en el tema de análisis Población.

**L.T Centros:** contiene los reportes con los que se logra realizar un análisis general de los datos en función de los indicadores que se ven reflejados en el tema de análisis Centros.

**L.T Viviendas:** contiene los reportes con los que se logra realizar un análisis general de los datos en función de los indicadores que se ven reflejados en el tema de análisis Viviendas.

**L.T Educación:** contiene los reportes con los que se logra realizar un análisis general de los datos en función de los indicadores que se ven reflejados en el tema de análisis Educación.

**L.T Comunicaciones:** contiene los reportes con los que se logra realizar un análisis general de los datos en función de los indicadores que se ven reflejados en el tema de análisis Comunicaciones.

**L.T Cuencas:** contiene los reportes con los que se logra realizar un análisis general de los datos en función de los indicadores que se ven reflejados en el tema de análisis Cuencas.

**L.T Transporte:** contiene los reportes con los que se logra realizar un análisis general de los datos en función de los indicadores que se ven reflejados en el tema de análisis Transporte.

**L.T Áreas Protegidas:** contiene los reportes con los que se logra realizar un análisis general de los datos en función de los indicadores que se ven reflejados en el tema de análisis Áreas Protegidas.

**L.T Economía:** contiene los reportes con los que se logra realizar un análisis general de los datos en función de los indicadores que se ven reflejados en el tema de análisis Economía.

**L.T Salud:** contiene los reportes con los que se logra realizar un análisis general de los datos en función de los indicadores que se ven reflejados en el tema de análisis Salud.

### Diseño de los cubos OLAP

El diseño de los cubos se realizó con la herramienta Schema Workbench, en la cual se definieron los cubos que representan los distintos hechos, además las dimensiones con sus jerarquías, los niveles que contienen las mismas y las medidas contenidas dentro de los hechos. También permitió añadir medidas calculables, dando cumplimiento a las reglas del negocio especificadas en la fase de análisis. En la figura 9 se muestra el diseño de los cubos del MD Plan Turquino:



Figura 9: Diseño de los cubos OLAP y las dimensiones.

## Roles y permisos

Para el acceso al MD se define un usuario por cada uno de los roles existentes en el sistema, con el objetivo de garantizar que cada usuario opera en el sistema según los permisos que se le definan al rol. En la siguiente tabla se muestran los roles con sus distintos permisos.

Roles	Aplicación		Base de datos	
	Lectura	Escritura	Lectura	Escritura
Administrador	X	X	X	X
r_etl			X	X
r_bi	X			

Tabla 5: Roles y permisos.

### 2.6. Políticas de respaldo y seguridad

Con el objetivo de garantizar la persistencia de la información, se establece una política de respaldo y recuperación que comprende tres elementos esenciales:

- ✓ **Periodicidad de las salvadas:** las salvadas de toda la información contenida en la base de datos (BD) del área de Plan Turquino se realizan anualmente, quedando definido así por la organización.
- ✓ **Tablas involucradas:** las tablas que se involucran en la realización son las 9 tablas de hechos identificadas en el proceso de análisis y las 10 dimensiones relacionadas.
- ✓ **Backups o salvadas existentes:** los backups son realizados anualmente debido a que es la frecuencia en que se realizarán cambios en el MD.

### 2.7. Conclusiones Parciales

Después de realizar un profundo análisis y diseño del MD Plan turquino se llegaron a las conclusiones siguientes:

- Las necesidades de información identificadas fueron la base para definir los 80 requisitos de información y los 17 requisitos funcionales, agrupados en 12 casos de uso de información y 6 funcionales respectivamente. Además de establecerse las reglas del negocio que servirán de apoyo para la realización de las transformaciones.
- A través del diseño del subsistema de almacenamiento se determinaron los elementos que formarán parte del modelo físico de datos, identificando las tablas de hechos, dimensiones y las medidas.

- El diseño del subsistema de integración realizado servirá de apoyo para la implementación de los procesos de ETL.
- Luego de diseñado el subsistema de visualización quedó definida la estructura de navegación compuesta por el A.A Plan turquino, 12 L.T y 17 reportes. También se diseñaron 15 cubos OLAP.



## Capítulo 3: Implementación del mercado de datos Plan Turquino

### Introducción

En este capítulo se abordará todo lo referente al proceso de integración de datos, que contiene los tres tipos de actividades de carácter general como son: extracción, transformación y carga en el mercado de datos. Se expondrán los principales elementos relacionados con la implementación del subsistema de visualización, modelo de datos, cubos multidimensionales y reportes candidatos, además de la implementación del subsistema de almacenamiento.

### 3.1. Implementación del subsistema de almacenamiento

En la implementación del subsistema de almacenamiento se realiza el desarrollo de la estructura física del MD, además se definen todos los estándares de codificación que van a poseer las estructuras del MD, para facilitar la comprensión por parte del cliente.

#### 3.1.1. Estándares de codificación

Con el objetivo de organizar como se va a denominar la estructura del AD, se formaliza un modelo, norma, patrón o un estándar de codificación. Esta acción permite a los desarrolladores entender cada una de las estructuras de los mercados. En la siguiente tabla se muestran como quedaron definidos estos estándares.

Estructura	Descripción	Ejemplo
Tablas de hechos	Todas las tablas de hechos tendrán una cadena que demuestra que son hechos y el concepto que describen.	hech_<concepto>
Tablas de dimensiones	Todas las tablas de dimensiones tendrán una cadena que demuestra que son dimensiones y el concepto que describen.	dim_<concepto>
Llaves	Todas las llaves primarias tendrán una cadena que demuestra que son llaves primarias y el nombre de la tabla a la	pk_<nombre_tabla>

primarias	que pertenecen.	
-----------	-----------------	--

Tabla 6: Estándares de codificación.

### 3.1.2. Implementación del subsistema de almacenamiento

Para la organización de las tablas en la BD del MD plan turquino quedaron definidos tres esquemas:

- El esquema **dimensiones** en el cual se encuentran las dimensiones comunes (DPA y Año).
- El esquema **mart\_turquino** donde se encuentran 17 tablas, 9 hechos y 8 dimensiones.
- El esquema **metadatos** en el cual se encuentran 6 tablas.

Formando esto un total de 25 tablas en la BD (figura 6).

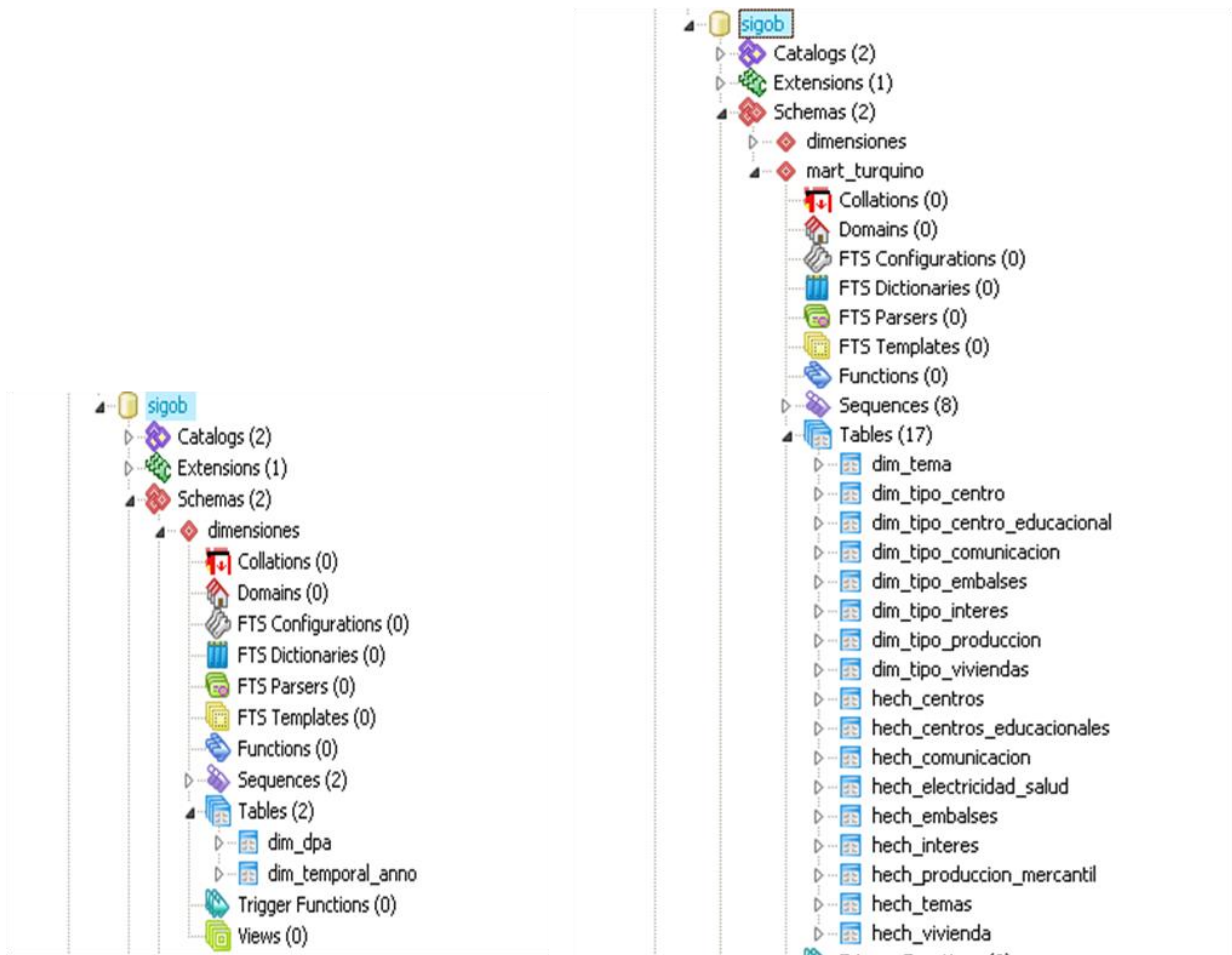


Figura 10: Organización de las tablas en la BD del MD.

### 3.2. Implementación del subsistema de integración

Para la implementación del subsistema de integración se lleva a cabo el proceso de integración de los datos, el cual consta con tres etapas fundamentales, extracción, transformación y carga de los datos. Antes de realizar este proceso es necesario hacer una limpieza de los datos de la fuente, permitiendo identificar los posibles errores de la fuente y así poder corregirlos. Una vez que los datos son extraídos, corregidos y transformados, se procede a la carga de los mismos, poblando las tablas contenidas en la BD del MD Plan Turquino.

### 3.2.1. Implementación de las transformaciones

Los elementos principales de los procesos ETL son las transformaciones y los trabajos, los mismos están compuestos por una serie de pasos que quedan entrelazados entre sí mediante los saltos, a través de los cuales se efectúa el flujo de los datos. En la presente investigación se realizó una transformación para cada una de las cargas de las tablas pertenecientes al esquema `mart_turquino`. Para las dimensiones, las transformaciones se realizaron a partir de la carga de los indicadores de cada una de las hojas de la fuente de datos, estos contienen los campos que son necesarios para poblar la BD para luego pasar a realizar las transformaciones de cada uno de los hechos.

Para realizar la carga de la dimensión `tipo_centro` (figura 7), el primer paso es la extracción de los datos de la fuente (INFO BASE DATOS 11), luego en el primer flujo de datos el componente **S/R** elimina los campos innecesarios del flujo, dejando solo los campos que contienen el valores que va a tener la dimensión. En el componente **Normalización de filas** se procede a tomar los valores de la dimensión que se encuentran en forma de columna y hacer una columna nueva con esos valores ahora en forma de filas. El componente **Filtrar filas** permite separar en dos flujos los centros deportivos de los centros de producción, ambos valores que se cargan de la misma hoja, luego de tener ya solo los valores de las tuplas que van a formar parte de la dimensión se procede a la unión de esos flujos en un mismo flujo y en las mismas columnas, para lo que se utiliza el componente **Unión Ordenada**. Seguidamente, luego de tener toda la información en un mismo flujo de datos se procede a la validación de los datos, validando que no vengan valores nulos en los campos y los tipos de datos, luego con el componente **Replace in string** se procede a eliminar el espacio que existe delante y detrás de los valores que toman los campos. El componente **Ordenar Filas** permite ordenar los valores de las tuplas ya sea de manera ascendente o viceversa, para que con el componente **Filas Únicas** obtener un solo valor para las tuplas que se repiten luego de organizadas en el **Ordenar Filas**. En el **Mapeo de Valores** se le cambian el valor de una tupla que venga por el flujo de datos por el otro deseado según la necesidad. El componente **Añadir Secuencia** permite obtener en una variable una secuencia de números de acuerdo a la cantidad de tuplas que vengan en los campos del flujo para luego en el **Insertar Actualizar** realizar la conexión a la BD igualando

los campos que hay en la tabla de la BD con los que vienen en el flujo de datos correspondidamente y de esta manera poblar dicha tabla.

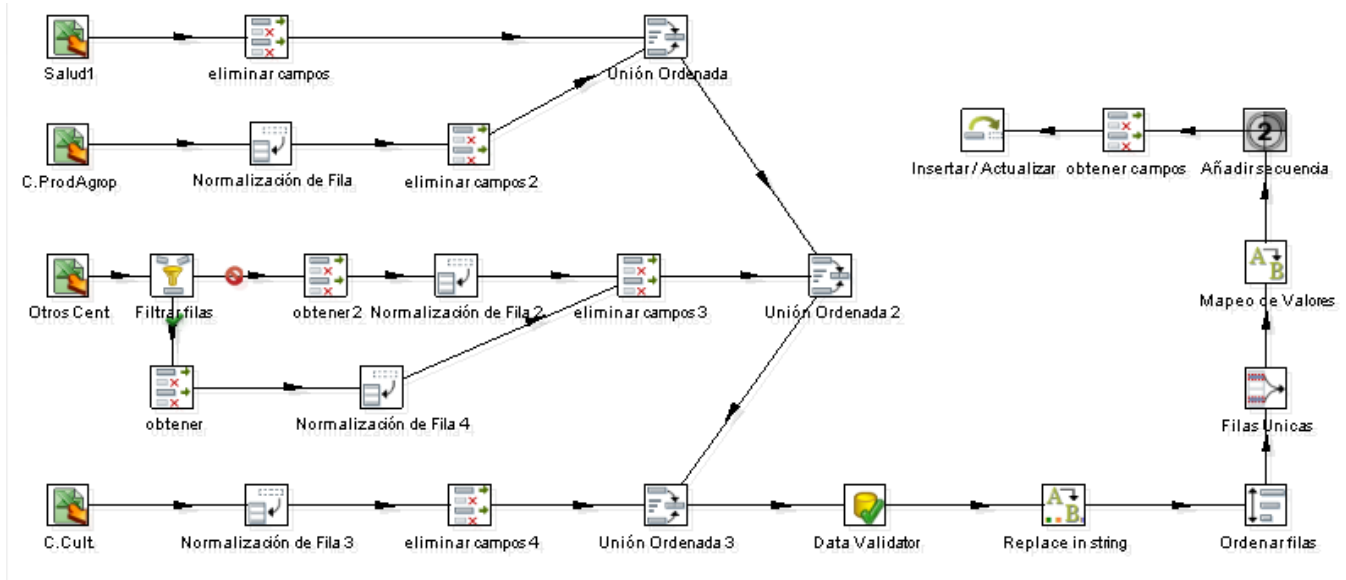


Figura 11: Transformación dimensión tipo de centro.

La siguiente figura es un ejemplo de transformación para la carga del hecho viviendas (Figura 8). En él se extraen los datos de la pestaña cuencas hidrográficas, se validan los datos. En el componente **Nulo si**, se hacen nulas las tuplas por las que venga el valor "-". En el componente **script** se ejecuta un código para obtener dos campos nuevos a partir del campo concepto, donde se tendrán dos columnas, en una el valor de las provincias y en la otra de los municipios dependiendo de ese valor de la provincia, luego se filtra en la columna municipio todos los valores que corresponden al nombre de una provincia y son mandados al paso " **no hace nada** ", se eliminan los espacios delante y detrás de los valores en los campos provincia y municipios se hace un mapeo de los valores de los nombres de dos municipios para igualarlos a como se llaman en la dim\_dpa para la búsquedas de sus id. Se procede a buscar los id (llaves primarias de las dimensiones) de las distintas dimensiones relacionadas con el hecho, se validan que estos id no vengan nulos y su tipo de dato y al final se procede a poblar la BD en la tabla correspondiente a este hecho.

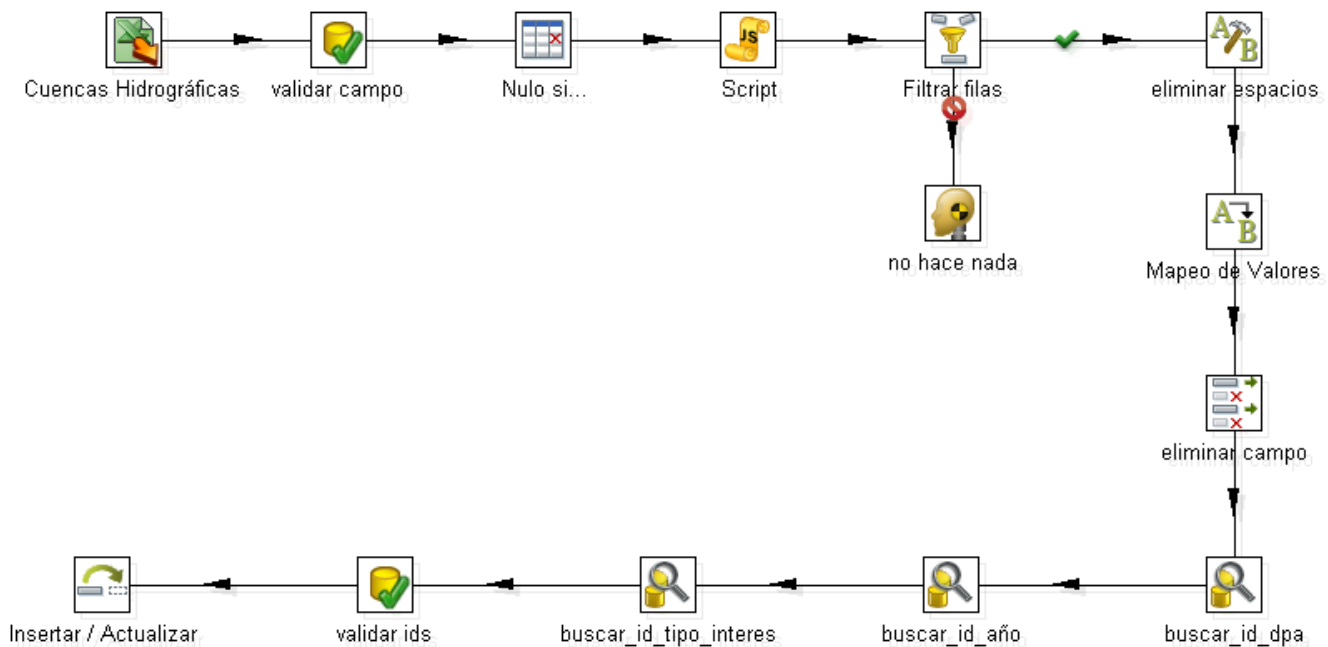


Figura 12: Transformación hecho vivienda.

### 3.2.2. Implementación de los trabajos

En la presente investigación se realizaron tres trabajos, uno para ejecutar las transformaciones para la carga de las 8 dimensiones (Figura 9), otro para ejecutar las transformaciones para la carga de los 9 hechos (Figura 10) y un trabajo general que ejecuta los trabajos anteriores (Figura 11).

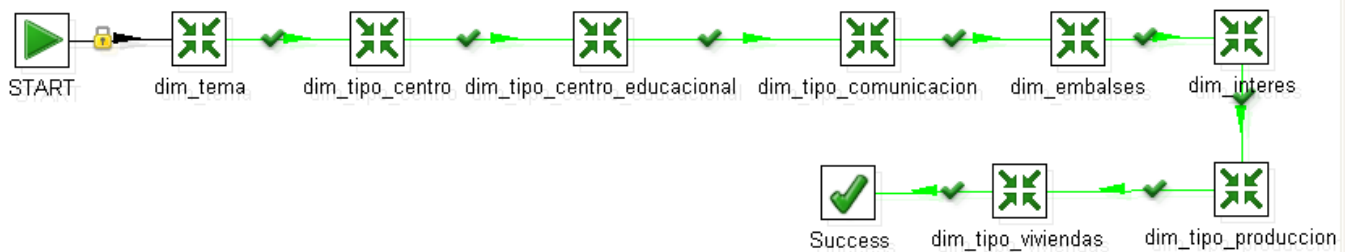


Figura 13: Trabajo para la carga de las dimensiones.

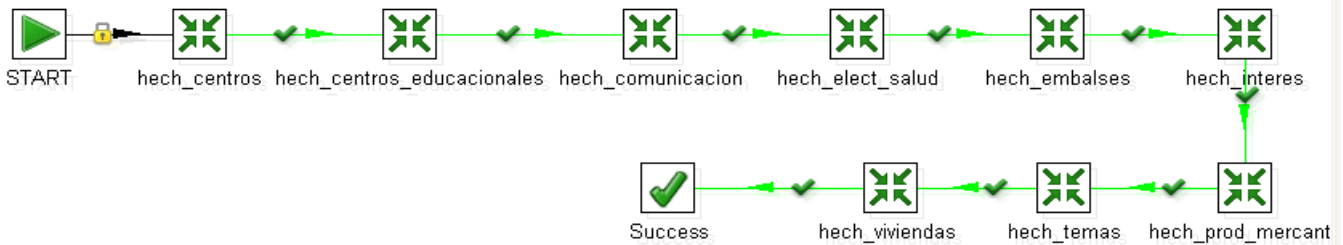


Figura 14: Trabajo para la carga de los hechos.

En el trabajo principal se comienza por la transformación **cargar configuración** en la cual se cargan las variables de conexión a la BD, luego se procede a **verificar la existencia de la fuente de datos**, en caso que no exista se crea un archivo txt que va a contener una breve descripción del error ocurrido. En el caso de que exista el fichero se procede al trabajo **carga dimensiones** y una vez concluido el mismo se **cargan los hechos**.

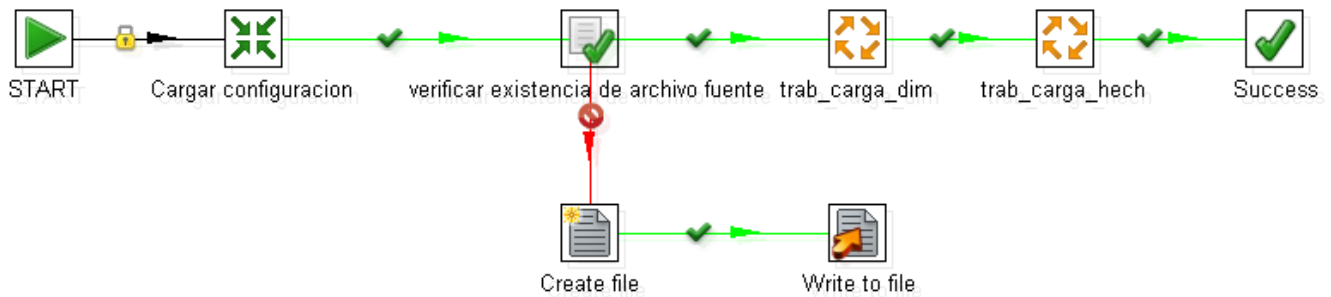


Figura 15: Trabajo principal.

### 3.3. Implementación del subsistema de visualización

Luego de realizada la carga de los datos se procede a la implementación de los cubos OLAP y la creación de los reportes candidatos, acciones correspondientes a la implementación del subsistema de visualización.

#### 3.3.1. Implementación de los cubos OLAP

En la implementación de los cubos OLAP se definieron todas las dimensiones, medidas y jerarquías que conforman el esquema Sigob. También se creó un cubo por cada una de las tablas de hechos. En la figura 16 se puede observar la implementación del cubo hecho interés:



Figura 16: Implementación del cubo hecho interés.

### 3.3.2. Representación de la arquitectura de la información

En la siguiente figura se hace una muestra de cómo queda representada la arquitectura de información del MD Plan Turquino detalladamente, donde se muestran todos los L.T que contienen los reportes que brindarán la información necesitada a los usuarios finales.



Figura 17: Arquitectura de información del MD Plan Turquino.

### 3.3.3. Implementación de los reportes candidatos

Los reportes candidatos o tablas de salida como también se les conoce contienen los valores de interés para el cliente. Dichos reportes fueron confeccionados luego de analizar la fuente de datos que recoge

toda esta información correspondiente al MD Plan Turquino. Fueron implementados en la herramienta Pentaho BI Server mediante consultas MDX, consultas muy similares a las de SQL, con la diferencia que en vez de utilizar los términos de tablas y campos utiliza el de hechos, dimensiones y medidas. A continuación se muestra la estructura de un reporte y la consulta MDX.

En la siguiente figura se muestra como queda conformado el reporte cantidad de embalses del turquino y volumen de embalsado mediante las distintas dimensiones relacionadas con el hecho. La dimensión dpa que contiene los nombres de las provincias y municipios de donde se recoge la información reflejada, a demás del año en que se realizó y los tipos de embales con que se cuenta.

	Año			
	2011			
	Medidas			
	Cantidad de embalses del turquino		Volumen de embalsado del turquino	
	Tipos de embalses		Tipos de embalses	
DPA	● Micropresas	● Presas	● Micropresas	● Presas
+ Pinar del Río	20	5	181.521.504	79.400.024
+ Artemisa	3	2	270.000	128.000.000
+ Villa Clara		1		108.244.000
+ Sancti Spiritus	2		400	
+ Ciego de Ávila	12	3	1.100.000	70.400.400
+ Holguín		1		141.000.000
+ Granma	1		155.400	
+ Santiago de Cuba	20	7	1.041.891,562	274,8
+ Guantánamo	3	4	2.800.000	23.400.000

Figura 18: Reporte cantidad de embalses del turquino y volumen de embalsado por dpa, año y tipo de embalses.

### 3.4. Conclusiones parciales

Concluida la etapa de implementación se puede arrojar que:

- En las transformaciones realizadas para las dimensiones y los hechos fueron implementadas satisfactoriamente las reglas de negocio identificadas en la etapa de análisis, así como, las reglas de transformación identificadas en el perfilado de los datos.
- Se crearon en la BD un total de 25 tablas: 10 de dimensiones, 9 hechos, organizadas en 3 esquemas, quedando implementado el subsistema de almacenamiento.



- Se diseñaron 9 cubos OLAP, una AA y 12 libros de trabajo, así como, 17 reportes candidatos, quedando así implementado el subsistema de visualización.
- Se realizó la carga de los datos de la fuente, quedando poblada la BD.

## Capítulo 4: Pruebas del mercado de datos Plan Turquino

### Introducción

En este capítulo se realizarán las validaciones al sistema y las pruebas pertinentes que comprueben la calidad del producto, como son los casos de prueba (CP), con el fin de comprobar si los resultados obtenidos a partir de la entrada de datos satisfacen los resultados esperados.

#### 4.1. Pruebas aplicadas al MD Plan Turquino 2.0

La prueba del software es un elemento crítico para la garantía de calidad del software y representa una revisión final de las especificaciones, del diseño y la codificación. La prueba requiere que se descarten las ideas preconcebidas sobre la “corrección” del software que se acaba de desarrollar y se supere cualquier conflicto de intereses que aparezca cuando se detecten errores. (21)

El modelo V es un método para comprobar el buen funcionamiento de los almacenes de datos, proviene del principio que establece que los procedimientos utilizados para probar si la aplicación cumple las especificaciones del cliente, ya deben haberse creado en la fase de diseño. Es una representación de dos cascadas enfrentadas y relacionadas, con su vértice en la codificación como punto en común. Propone una cascada a la izquierda, con las actividades relacionadas al desarrollo y una a la derecha con las actividades del aseguramiento de la calidad del software. Mediante este modelo se describe a un nivel muy alto de abstracción las fases del ciclo de desarrollo en las que se involucra la prueba. (22)

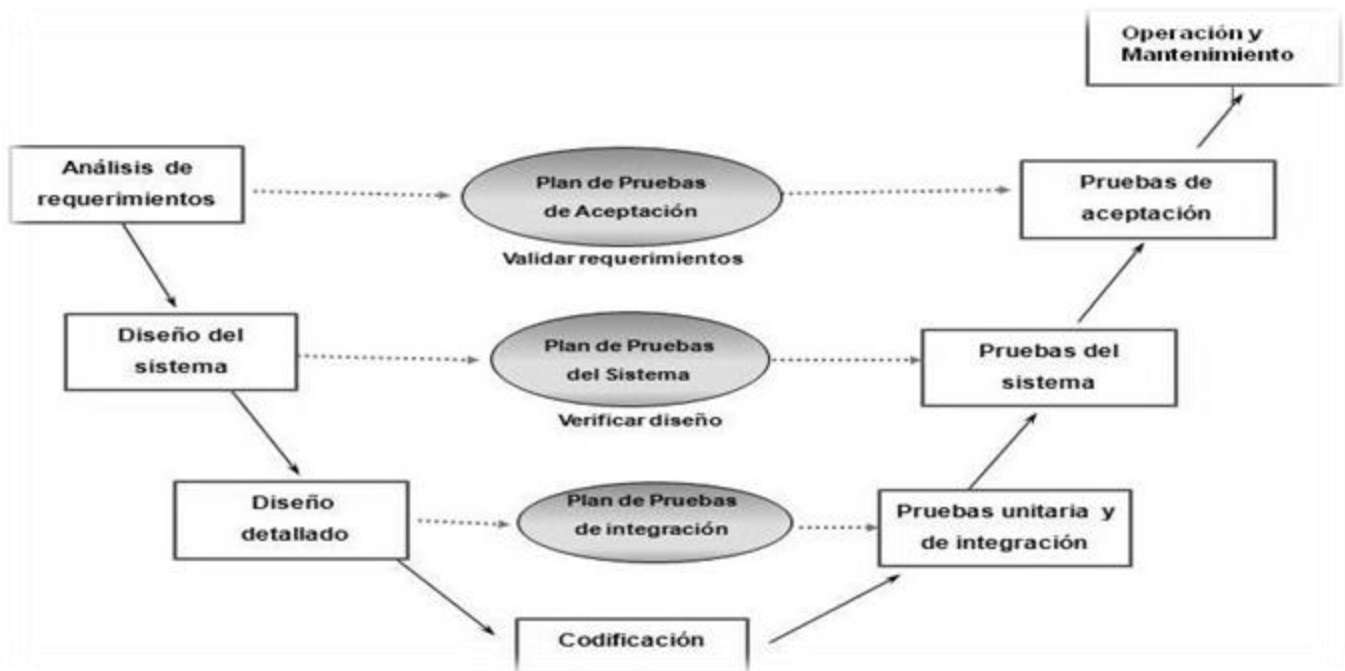


Figura 19: Modelo v.

### Pruebas unitarias

La prueba de unidad es la primera fase de las pruebas dinámicas y se realizan sobre cada módulo del software de manera independiente. (23)

**Objetivo:** comprobar que el módulo, entendido como una unidad funcional de un programa independiente, está correctamente codificado. En estas pruebas cada módulo será probado por separado y lo hará, generalmente, la persona que lo creó. En general, un módulo se entiende como un componente software que cumple las siguientes características: (23)

- Debe ser un bloque básico de construcción de programas.
- Debe implementar una función independiente simple.
- Podrá ser probado al cien por cien por separado.
- No deberá tener más de 500 líneas de código.

### Pruebas de integración

Aún cuando los módulos de un programa funcionen bien por separado es necesario probarlos conjuntamente: un módulo puede tener un efecto adverso o inadvertido sobre otro módulo; las subfunciones, cuando se combinan, pueden no producir la función principal deseada; la imprecisión

aceptada individualmente puede crecer hasta niveles inaceptables al combinar los módulos; los datos pueden perderse o malinterpretarse entre interfaces, etc. Por lo tanto, es necesario probar el software ensamblando todos los módulos probados previamente. Este es el objetivo de la pruebas de integración. A menudo hay una tendencia a intentar una integración no incremental; es decir, a combinar todos los módulos y probar todo el programa en su conjunto. El resultado puede ser un poco caótico con un gran conjunto de fallos y la consiguiente dificultad para identificar el módulo (o módulos) que los provocó. En contra, se puede aplicar la integración incremental en la que el programa se prueba en pequeñas porciones en las que los fallos son más fáciles de detectar. Existen dos tipos de integración incremental, la denominada ascendente y descendente. Veamos los pasos a seguir para cada caso: Integración incremental ascendente: (23)

- Se combinan los módulos de bajo nivel en grupos que realicen una subfunción específica.
- Se escribe un controlador (un programa de control de la prueba) para coordinar la entrada y salida de los casos de prueba.
- Se prueba el grup.
- Se eliminan los controladores y se combinan los grupos moviéndose hacia arriba por la estructura del programa.

### **Pruebas del sistema**

Este tipo de pruebas tiene como propósito ejercitar profundamente el sistema para verificar que se han integrado adecuadamente todos los elementos del sistema (hardware, otro software, etc.) y que realizan las funciones adecuadas. Concretamente se debe comprobar que: (23)

- Se cumplen los requisitos funcionales establecidos.
- El funcionamiento y rendimiento de las interfaces hardware, software y de usuario.
- La adecuación de la documentación de usuario.
- Rendimiento y respuesta en condiciones límite y de sobrecarga.

Para la generación de casos de prueba del sistema se utilizan técnicas de caja negra. Este tipo de pruebas se suelen hacer inicialmente en el entorno del desarrollador, denominadas Pruebas Alfa, y seguidamente en el entorno del cliente denominadas Pruebas Beta. (23)

### **Pruebas de aceptación**

A la hora de realizar estas pruebas, el producto está listo para implantarse en el entorno del cliente. El usuario debe ser el que realice las pruebas, ayudado por personas del equipo de pruebas, siendo deseable, que sea el mismo usuario quien aporte los casos de prueba. Estas pruebas se caracterizan por: (23)

- Participación activa del usuario, que debe ejecutar los casos de prueba ayudado por miembros del equipo de pruebas.
- Están enfocadas a probar los requisitos de usuario, o mejor dicho a demostrar que no se cumplen los requisitos, los criterios de aceptación o el contrato. Si no se consigue demostrar esto el cliente deberá aceptar el producto.
- Corresponden a la fase final del proceso de desarrollo de software.

Es muy recomendable que las pruebas de aceptación se realicen en el entorno en que se va a explotar el sistema (incluido el personal que lo maneje). En caso de un producto de interés general, se realizan pruebas con varios usuarios que reportarán sus valoraciones sobre el producto. Para la generación de casos de prueba de aceptación se utilizan técnicas de caja negra. (23)

## 4.2. Herramientas de prueba

### Casos de prueba

El propósito de los casos de prueba (CP) es la comprobación de la calidad que posee el software desarrollado, en ellos se identifican los fallos en la implementación de la solución, los resultados esperados y el cumplimiento de las especificaciones del sistema. En la presente investigación se desarrollaron 12 CP, uno por cada caso de uso de información identificado en la etapa de análisis. En la siguiente tabla se muestra el caso de prueba correspondiente al CU "Mostrar información de embalses", donde se aprecia el reporte del L.T Presas.

Escenario	Descripción	Variables de entrada	Variables de salida	Respuesta del sistema	Flujo central

EC 1.1 Cantidad de embalses y volumen embalsado del turquino	Muestra la cantidad de embalses y el volumen de embalsado de presas y micropresas del turquino	-dpa - año - tipos de embalses	- cantidad de embalses del turquino - volumen de embalsado del turquino	Se muestra la información correspondiente al escenario	1. Se abre la aplicación 2. Se autentica 3. Se entra al sistema 4. En la parte superior izquierda se selecciona el Área de Análisis General SIGOB. 5. Se selecciona el Área de Análisis Plan Turquino 6. Se selecciona el Libro de Trabajo L. T Presas 7. En la parte inferior izquierda, se selecciona el reporte que corresponde al escenario especificado. 8. Se visualiza el reporte en el área de trabajo
---	--	--------------------------------------	--	--	--

Tabla 7: CP correspondiente al CU "mostrar información de embalses".

### 4.3. Resultados de las pruebas

#### Pruebas unitarias y de integración

Una vez terminada la implementación se realizaron un conjunto de pruebas de unidad e integración a los distintos componentes de los subsistemas arrojando un total de 7 no conformidades (NC), las cuales se muestran a continuación:

- La estructura de carpetas del repositorio de ETL no está correcta.
- Los metadatos relacionados con la gestión del cambio en la fuente no han sido implementados.
- No utiliza variables de entorno para la navegación por los directorios, grave error pues en el despliegue puede presentar serios problemas para los desarrolladores.
- Errores lógicos en la implementación de las transformaciones.
- Agregar los archivos .properties para los nombres de los reportes.
- Debe arreglar el nombre de las medidas.
- Arreglar la conexión, ponerla genérica, actualizar los reportes con esa conexión.

Las mismas fueron resueltas satisfactoriamente en el tiempo establecido.

## Pruebas del sistema

En las pruebas realizadas por el grupo de calidad del departamento de almacenes de datos de DATEC fueron identificadas un total de 42 NC, estas se basaban principalmente en errores relacionados con los nombres de las columnas de la fuente de datos y abuso del uso de mayúsculas. Una vez resueltas las NC detectadas se realizó una segunda iteración donde fue verificado el cumplimiento de la solución a dichas NC, arrojando como resultado que las mismas fueron resueltas satisfactoriamente.

### 4.4. Conclusiones parciales

Terminada la implementación del mercado de datos Plan Turquino y quedando aplicadas las pruebas de calidad a la aplicación con el objetivo de comprobar la calidad de los datos integrados se arribaron a las siguientes conclusiones:

- ✓ Se demostró mediante las distintas pruebas aplicadas a la aplicación la identificación de un conjunto de NC que finalmente fueron resueltas garantizando el funcionamiento del MD.
- ✓ Se establece que las herramientas de prueba utilizadas permitieron identificar la calidad de la aplicación desarrollada y verificar el cumplimiento de los objetivos definidos por el cliente.

## Conclusiones

El estudio de los distintos temas relacionados con el desarrollo de los almacenes de datos, proporcionó la elaboración del presente trabajo, el cual arrojó como resultado el “Mercado de datos Plan turquino 2.0 para el Sistema de Información de Gobierno”. Los siguientes resultados demuestran el cumplimiento de los objetivos propuestos en la investigación:

- ✓ Se logró mediante una investigación detallada la selección de la metodología, herramientas y tecnologías a utilizar en el desarrollo de la solución, permitiendo sentar las bases en el proceso de construcción del MD.
- ✓ Se realizó el análisis y diseño del mercado de datos Plan turquino, identificando los requisitos de información, funcionales, no funcionales y las reglas del negocio. Además, se diseñaron los subsistemas de almacenamiento, integración y visualización, sirviendo de base para la implementación del mercado de datos.
- ✓ La implementación de los subsistemas de almacenamiento, integración y de visualización posibilitaron la obtención de un mercado de datos correctamente poblado, con información disponible para ser consultada por parte de los usuarios, brindando apoyo al proceso de toma de decisiones.
- ✓ Se demostró mediante las pruebas que el sistema cumple con las necesidades del cliente.



## Recomendaciones

- Se recomienda que se incluya la información de la ficha de datos en el almacén de datos del SIGE ya que solo se encuentra la información del modelo 0024-03.

## Referencias bibliográficas

1. **Ecured.** [En línea] 2012. [http://www.ecured.cu/index.php/Plan\\_Turquino](http://www.ecured.cu/index.php/Plan_Turquino).
2. **Modelo 0024-03.** *Informe de la actividad del Plan Turquino*. Ciudad de la Habana : s.n., 2008.
3. **ONEI.** *Oficina Nacional de Estadísticas e Información*. [En línea] 2008. <http://www.onei.cu>.
4. **Velasco, R. H.** (24 de abril de 2004). Recuperado el 2013, de <http://www2.rhernando.net/modules/tutorials/viewexttutorial.php?tid=40&PHPSESSID=59d1af89fcb1a6d6581d3f6744dbfc8b>.
5. **Synerplus.** [En línea] <http://www.synerplus.es/Informacion-Tecnica/Data-Mart/309.html>.
6. **DataPRIX.** (s.f.). Recuperado el 2013, de <http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager>.
7. **Ecured.** [En línea] 2012. <http://www.ecured.cu/index.php/OLAP>.
8. **etl-tools.info.** (s.f.). Recuperado el 2013, de [http://etl-tools.info/es/bi/proceso\\_etl.htm](http://etl-tools.info/es/bi/proceso_etl.htm).
9. **Kimball, Ralph.** *The Data Warehouse ETL Toolkit*. s.l.: Wiley.
10. [Online] [Cited: mayo 28, 2013.] <http://www.gestiopolis.com/recursos5/docs/ger/buconce.htm>.
11. **Ecured.** [En línea] 2012. [http://www.ecured.cu/index.php/Plataforma\\_Pentaho](http://www.ecured.cu/index.php/Plataforma_Pentaho).
12. **González Hernández, Yanisbel.** *PROPUESTA DE METODOLOGIA DE DESARROLLO DE ALMACENES DE DATOS PARA DATEC*. Ciudad de la Habana : s.n., 2011.
13. **Pressman, Roger S.** *Ingeniería de Software, un enfoque práctico*. Quinta edición. S.l.: McGraw-Hill Companies, 2002. ISBN: 8448132149.
14. **Sitio Oficial de PostgreSQL.** [En línea] 2012. <http://www.postgresql.org/about/press/presskit91/es/>.
15. **Lgs.** [En línea] 2012. [http://www.lgs.com.ve/pres/PresentacionES\\_PSQL.pdf](http://www.lgs.com.ve/pres/PresentacionES_PSQL.pdf).
16. **Guía Ubuntu.** [En línea] 2012. [http://www.guia-ubuntu.org/index.php?title=PgAdmin\\_III](http://www.guia-ubuntu.org/index.php?title=PgAdmin_III).
17. **Sitio Oficial de Pentaho.** [En línea] 2012. <http://www.pentaho.com>.
18. **Sitio oficial de Pentaho.** [En línea] 2012. <http://mondrian.pentaho.com/documentation/workbench.php>.
19. **Comunidad Pentaho.** [En línea] 2012. [http://community.pentaho.com/projects/bi\\_platform](http://community.pentaho.com/projects/bi_platform).
20. **ECURED.** [En línea] 2013. [Citado el: 28 de mayo de 2013.] [http://www.ecured.cu/index.php/Caso\\_de\\_uso](http://www.ecured.cu/index.php/Caso_de_uso).
21. **Gonzalez, Y., & De Cuadra, F.** (2013). *Entorno virtual de aprendizaje*. Obtenido de [http://eva.uci.cu/file.php/158/Documentos/Recursos\\_bibliograficos/Libros\\_y\\_articulos\\_UD\\_2/Comun/Calidad\\_del\\_software\\_I\\_Yolanda\\_Gonzalez\\_Fernando\\_de\\_Cuadra\\_.pdf](http://eva.uci.cu/file.php/158/Documentos/Recursos_bibliograficos/Libros_y_articulos_UD_2/Comun/Calidad_del_software_I_Yolanda_Gonzalez_Fernando_de_Cuadra_.pdf).

22. **CICLO DE VIDA DEL SOFTWARE.** [En línea] (2008). [Citado el: 9 de mayo de 2013.]. Disponible en:  
< <http://es.kioskea.net/contents/genie-logiciel/cycle-de-vie.php3> />.
23. **Slideshare.** [En línea] <http://www.slideshare.net/LGasperin/estrategias-de-pruebas-de-software>.

## Bibliografía

1. **Ecured. (s.f.)**. Obtenido de [http://www.ecured.cu/index.php/Metodología\\_Hefesto](http://www.ecured.cu/index.php/Metodología_Hefesto)
2. **González Hernández, Yanisbel**. *PROPUESTA DE METODOLOGIA DE DESARROLLO DE ALMACENES DE DATOS PARA DATEC*. Ciudad de la Habana : s.n., 2011.
3. **ONEI**. *Oficina Nacional de Estadísticas e Información*. [En línea] 2008. <http://www.onei.cu>.
4. **BibliotecaUCI**. (s.f.). Obtenido de <http://biblioteca.uci.cu>.
5. **Inmon, W. H.** *Building the Data Warehouse*. s.l.: Wiley Publishing. ISBN: 0-471-08130-2.
6. **Kimball, Ralph**. *The Data Warehouse ETL Toolkit*. s.l.: Wiley.
7. **Base de Datos Oracle**. [En línea] Manuel de la Herrán Gascón, 1999-2004. <http://www.redcientifica.com/oracle/c0001p0005.html>
8. **Dataprix**. [En línea] <http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager#x1-480003.4.4.5>.
9. **El Rincón del BI**. [En línea] Roberto Espinosa Milla. <http://churriwifi.wordpress.com/2010/07/04/17-3-preparando-el-analisis-dimensional-definicion-de-cubos-utilizando-schema-workbench/>.
10. **Inmon, William H.** *"What is a Data Warehouse?"*. [ed.] Prism. 1995.
11. **Kimball, Ralph**. *"The Data Warehouse Lifecycle"*. 2002.
12. **Sánchez, Leopoldo Zenaido Zepeda**. *Metodología para el Diseño*. Valencia: s.n., Marzo de 2008.

## Anexos

## Glosario de términos

- ❖ **A.A:** Área de Análisis.
- ❖ **A.A.G:** Área de Análisis General.
- ❖ **BI:** Bussines Intelligence.
- ❖ **CU:** Caso de Uso.
- ❖ **CASE:** Ingeniería de Software Asistida por Computadora.
- ❖ **CAP:** Consejo de la Administración Provincial.
- ❖ **DATEC:** Centro de Tecnologías de Gestión de Datos.
- ❖ **CALISOFT:** Calidad del Software.
- ❖ **DPA:** División Política Administrativa.
- ❖ **Estadísticas:** Ciencia que se ocupa de la recogida y ordenación de datos numéricos, para obtener a partir de ellos conclusiones basadas en el cálculo de probabilidades.
- ❖ **ETL:** Extraction, Transformation, Load.
- ❖ **L.T:** Libro de Trabajo.
- ❖ **MINAG:** Ministerio de la Agricultura.
- ❖ **MICONS:** Ministerio de la Construcción.
- ❖ **MITRANS:** Ministro del Transporte.
- ❖ **Plan Turquino:** Se crea en Cuba con el fin de lograr un desarrollo integral y sostenible de las zonas montañosas.
- ❖ **SIGOB:** Sistema de Información de Gobierno.
- ❖ **SGBD:** Sistema de Gestor de Base de Datos.
- ❖ **Tupla:** Llamadas también registros o filas. Contienen la información relativa a una única entidad.
- ❖ **TCP/IP:** Protocolo de control de transmisión/Protocolo de Internet. Protocolo utilizado para asignar una dirección única para cada computadora en la red, dividir los mensajes en paquetes, enrutar los datos en la red y detectar errores en dichos datos.
- ❖ **T.S:** Tabla de Salida.
- ❖ **UCI:** Universidad de las Ciencias Informáticas.