

**Universidad de las Ciencias Informáticas**

**Facultad 6**



**Título: Sistema de Información de Gobierno. Mercado de Datos Medio ambiente.**

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

**Autores:**

**Arleidys Pérez Hernández**

**David Macias Garcias**

**Tutores:**

**Ing. Yosbel Rodríguez Rodríguez**

**Ing. Yunier Santana Aldana**

**Ciudad de La Habana, junio 2011**



*"El mundo camina hacia la era electrónica... Todo indica que esta ciencia se constituirá en algo así como una medida del desarrollo; quien la domine será un país de vanguardia. Vamos a volcar nuestros esfuerzos en este sentido con audacia revolucionaria"*

A handwritten signature in black ink, which appears to be 'Che'.

## 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 firmamos la presente a los \_\_\_\_ días del mes de \_\_\_\_\_ del año \_\_\_\_\_.

\_\_\_\_\_

Firma del Autor

\_\_\_\_\_

Firma del Autor

\_\_\_\_\_

Firma del Tutor

\_\_\_\_\_

Firma del Tutor

**Tutor:** Ing. Yunier Santana Aldana

**Especialidad de graduación:** Ingeniería en Ciencias Informáticas

**Categoría docente:** Instructor en Adiestramiento

**Categoría Científica:** Ingeniero

**Correo electrónico:** ysaldana@uci.cu

**Tutor:** Ing. Yosbel Rodríguez Rodríguez

**Especialidad de graduación:** Ingeniería en Ciencias Informáticas

**Categoría docente:** Instructor en Adiestramiento

**Categoría científica:** Ingeniero

**Correo electrónico:** yrdguezro@uci.cu

### *De ambos:*

A **Fidel** y a la **Revolución** por habernos dado la oportunidad de estudiar en la UCI y formarnos profesionalmente.

A los **profesores** que de una forma u otra contribuyeron en nuestro desarrollo profesional y otros que llegaron a ser mis amigos, les agradezco por haberme preparado para el futuro.

A nuestros tutores Tito y Yosbel que nos han aclarado cuantas dudas han surgido a lo largo del camino, gracias por ser nuestros guías.

A mi **compañera/o de tesis** por tanto sacrificio y dedicación para la realización del trabajo.

A todas las personas que hemos conocido en la UCI por haber compartido con nosotros esta etapa tan linda, principalmente a nuestros **compañeros del aula** por apoyarnos en las diferentes asignaturas. Sepan que siempre estarán en nuestros corazones como el más grato recuerdo de nuestras vidas como estudiantes.

A **todos** los que han contribuido a la realización de este trabajo de diploma, de corazón, muchas gracias.

### *De Arleidy:*

Ocultar lo que siento hoy, dejaría en mí palabras guardadas que quiero compartir con personas que quizás nunca vuelva a ver. Así que aprovecho para decirle, a las amistades que he perdido por un mal entendido o desacuerdo y a aquellas personas que no supieron valorarme como yo hubiese querido y lograron hacerme sentir mal, que no les guardo rencor ni quisiera que lo hicieran conmigo, porque solo llevaré de ustedes los buenos momentos compartidos y les pido perdón por no haber sido más tolerante. Y les agradezco porque me ayudaron a hacerme fuerte y a enfrentar la vida de manera diferente.

Le agradezco:

A la Revolución y a su máximo líder Fidel Castro porque la admiración que siento hacia su persona provocó que desde mi primer año en la Enseñanza Media Superior añorara estudiar en esta Universidad tan solo por la satisfacción de que él haya sido su creador.

A todos los profesores de la facultad, en especial a los que han formado parte de este tribunal, mis tutores, mi oponente y a los profesores de la Línea de Almacenes de Datos que han contribuido a mejorar la calidad de este

*trabajo. Sepan que de cada uno de ustedes me llevaré un consejo, una palabra o un simple gesto que me harán recordarlos por siempre.*

*A mis compañeros de 5to año por todas las cosas que hemos compartido juntos y los felicito además porque al igual que yo, están viendo cumplidos su sueño de ser ingenieros. A mis compañeras de apartamento por soportar mis malos humores y mis payasadas que solo fueron fruto de picardías para ocultar lo mucho que las aprecio a todas. Agradezco a todas las amistades que tengo en esta escuela que muchas de ellas seguirán aquí y que sé, darán su mejor esfuerzo en aras de crecerse, cultivarse y enriquecerse con las experiencias vividas en la UCI. Mis mejores deseos de que la vida les sonría a todos.*

*A los muchachos del 111101 y 111102 en especial a Frank, Eric, José y Eduardo por la amistad tan linda que hemos tenido este curso, por contar conmigo y hacerme partícipe de sus problemas. Solo lamento que haya sido tan poco el momento que compartimos y espero poder estar aquí con ustedes dentro de 4 años.*

*A Aylenis porque aunque a veces me dice cosas que no quisiera escuchar ha sido la persona que más cerca ha estado de mí todo este curso y me ha hecho apreciarla y valorarla como una gran amiga.*

*A Enrique que a pesar de que seamos como perro y gato cuando pasamos mucho tiempo junto ha sido el hombro al que he ido a llorar cuando algo me lastima o me preocupa.*

*A Taimí que más que mi amiga es mi hermana y siempre ha estado ahí incluso cuando nos separaba todo el Mar Caribe. Porque es la única persona que me ha acompañado a llorar chateando una al lado de la otra en cualquier laboratorio.*

*A mi hermano por haberme acompañado a crecer y ser el apoyo que he encontrado desde mis primeros pasitos hasta el día de hoy.*

*A mi mamá por haber sido tan tolerante conmigo, por haberme enseñado desde bien pequeña a ser una mujer independiente. Por educarme, por escucharme ininterrumpidamente largas horas de historias para acabar diciéndome que emborracho más que una Guayabita del Pinar. Por haberme permitido ser su mejor amiga y por ser mi eterno ejemplo de fortaleza y decisión en la vida.*

*A mi papá que quizás no sea un hombre perfecto, pero si ha sabido ser el padre perfecto. Porque mi amor por ti es incondicional y nació desde esos días en los que no me dormía hasta verte llegar cargándome y diciéndome: "Cerrando los ojitos, arriba a dormir". Porque cuando alguien me pregunta quién ha sido el hombre que ha*

*marcado mi vida, ese sin el cual no quisiera o pudiera vivir, no dudo nunca en decir que el hombre de mi vida eres tú.*

*A mí porque después de tantas cuentas sacadas creo que hoy la modestia puede quedar a un lado para darme a mí misma las gracias por no dejarme vencer ante las dificultades, por reponerme y levantarme después de haber caído. Porque hoy veo el fruto de tantos años de esfuerzo, de estar alejada de las personas que más quiero. Porque después de estos 5 años me voy con miles de nuevos conocimientos, con una manera de pensar diferente y con muchas ganas de vivir.*

### *De David*

*Agradecer al final de camino es realmente difícil porque para agradecer a todos los involucrados en mi resultado como estudiante en esta escuela necesito más que una hoja es por eso que :*

*Agradezco a todos los que de una forma u otra aportaron su granito de arena, mis amigos, mis compañeros de aula durante estos 5 años, y a mis profesores.*

*Agradezco especialmente a mi querida madre Santa Eulalia y a mi padre Cristino quienes me apoyaron, guiaron y aconsejaron para hacer de mí un hombre, un estudiante y una mejor persona cada día.*

*Agradezco a mi hermano, mi cuñada y mi sobrina quienes estuvieron ahí para darme su apoyo siempre que lo necesité y cuando no lo necesite también.*

*A mis amigos Roberto, Ariorkis, Gustavo, Yever, Dayelin, Soto, Aramis quienes no me dejaron fuera de una fiesta, de un cuento, de un partido de futbol, y menos de su agradecimiento, así que aquí está el suyo.*

*Agradezco Dunieska la chica que más dolores de cabeza me provocó, por su apoyo, tiempo y dedicación para ti mis agradecimientos especiales.*

*Agradezco a mi familia y a la familia de mi familia, y a la familia de mis amigos que me brindaron apoyo, ayuda y confianza cuando más lo necesité o cuando menos lo esperé.*

*Agradezco a una familia de la que yo no formo parte pero sin embargo me siento parte de ella, a la familia Cardero: Alina, Roberto, Carlito, Carlitin, El tío, el colorao, los abuelos, y a todos y todas sus miembros por abrir las puertas y dejarme entrar y formar parte de ella en sus alegrías y sus tristezas. A todos muchas gracias....*

*De Arleidys:*

A la **Revolución** que nos ha dado todo.

A **Fidel** por haber hecho esta Revolución.

A mi **mamá**, mi **papá** y mi **hermano** que son mi mayor posesión y los seres más importantes para mí en esta vida y en las venideras, si es que existe la reencarnación.

*De David:*

A mi **Santa** y **Cristino**, mis queridos padres, por hacer de mi todo lo que soy.

A **Cristian** mi hermano, **Yaremis** su esposa y su linda hija **Anabel** por su apoyo y guía en la vida.

A quienes son como mis padres, **Alina** y **Roberto** y a su hijo **chichi** que es como un hermano por su dedicación y cariño hacia mí.

A mi **familia** y mis **amigos** por su apoyo incondicional en todo momento.

La Oficina Nacional de Estadísticas en colaboración con el Centro de Tecnologías de Gestión de Datos de la Universidad de la Ciencias Informáticas, trabajan en conjunto para mejorar el proceso de análisis y gestión de la información que maneja esta entidad. En el presente trabajo se muestra la implementación de un Mercado de datos para el área de Medio ambiente, donde se utilizó la Metodología para el diseño de Kimball. Apoyándose en la herramienta Visual Paradigm para el modelado y PostgreSQL como Sistema Gestor de Base de Datos. Para los procesos de Extracción, Transformación y Carga el Pentaho Data Integration y el DataCleaner y para los procesos de Inteligencia de Negocio el Pentaho Schema Workbench, Pentaho BI Server, Mondrian y el Apache Tomcat. Este Mercado de datos contribuye a que el trabajo de los especialistas sea más eficiente en el proceso para la toma de decisiones, por lo que posibilita una mejor organización en esta entidad sobre los datos que se manejan.

**Palabras claves:** Mercado de Datos, toma de decisiones, BI.

---

|                                                                                |    |
|--------------------------------------------------------------------------------|----|
| INTRODUCCIÓN .....                                                             | 1  |
| CAPÍTULO 1. FUNDAMENTACIÓN TEÓRICA .....                                       | 5  |
| Introducción .....                                                             | 5  |
| 1.1 Medio ambiente .....                                                       | 5  |
| 1.2 Almacén de datos .....                                                     | 6  |
| 1.2.1 Modelo multidimensional .....                                            | 8  |
| 1.2.2 Mercados de datos .....                                                  | 10 |
| 1.2.3 Metadatos .....                                                          | 11 |
| 1.2.4 Modos de Almacenamiento de Datos .....                                   | 12 |
| 1.2.5 Metodologías para el desarrollo .....                                    | 15 |
| 1.3 Integración de datos .....                                                 | 16 |
| 1.3.1 Replicación de datos .....                                               | 17 |
| 1.3.2 Integración de Información Empresarial: .....                            | 17 |
| 1.3.3 Extracción, Transformación y Carga de datos (ETL).....                   | 17 |
| 1.4 Inteligencia de Negocios. ....                                             | 18 |
| 1.4.1 Ejemplo del mundo real de aplicaciones de Inteligencia de Negocios ..... | 20 |
| 1.5 Herramientas para el desarrollo .....                                      | 20 |

|                                                        |           |
|--------------------------------------------------------|-----------|
| 1.5.1 Sistema gestor de base de datos. ....            | 20        |
| 1.5.2 Herramienta manejadora de base de datos .....    | 22        |
| 1.5.3 Herramienta para el modelado .....               | 22        |
| 1.5.4 Herramienta para la integración de datos.....    | 23        |
| 1.5.5 Herramientas para Inteligencia de Negocios ..... | 24        |
| Conclusiones del capítulo .....                        | 27        |
| <b>CAPÍTULO 2: ANÁLISIS Y DISEÑO.....</b>              | <b>28</b> |
| Introducción.....                                      | 28        |
| 2.1 Análisis de la Solución .....                      | 28        |
| 2.1.1 Definición del Negocio .....                     | 28        |
| 2.1.2 Tema de Análisis Identificado.....               | 29        |
| 2.1.3 Roles y permisos .....                           | 29        |
| 2.1.4 Reglas del Negocio.....                          | 30        |
| 2.1.5 Requisitos de información .....                  | 31        |
| 2.1.6 Requisitos multidimensionales .....              | 31        |
| 2.1.7 Requisitos funcionales.....                      | 32        |
| 2.1.8 Requisitos no funcionales.....                   | 33        |
| 2.1.9 Casos de uso del sistema.....                    | 34        |

|                                                         |    |
|---------------------------------------------------------|----|
| 2.2 Diseño de la Solución.....                          | 36 |
| 2.2.1 Modelo de Datos.....                              | 36 |
| 2.2.2 Matriz BUS o Matriz Dimensional .....             | 38 |
| Conclusiones del capítulo .....                         | 38 |
| CAPÍTULO 3: IMPLEMENTACIÓN.....                         | 40 |
| Introducción.....                                       | 40 |
| 3.1 Implementación del modelo de datos.....             | 40 |
| 3.2 Implementación del subsistema de integración .....  | 44 |
| 3.2.1 Arquitectura del subsistema de integración.....   | 44 |
| 3.3 Implementación de subsistema de visualización ..... | 48 |
| 3.3.1 Cubos OLAP .....                                  | 48 |
| 3.3.1 Navegación de la capa de visualización.....       | 49 |
| 3.3.3 Configurar la seguridad de los usuarios .....     | 51 |
| Conclusiones del capítulo .....                         | 52 |
| CAPÍTULO 4: VALIDACIÓN .....                            | 53 |
| Introducción.....                                       | 53 |
| 4.1 Diseño de casos de pruebas.....                     | 53 |
| 4.2 Lista de chequeo .....                              | 54 |

|                                                 |    |
|-------------------------------------------------|----|
| 4.3 Pruebas de aceptación.....                  | 56 |
| 4.4 Resultados y discusión de las pruebas ..... | 57 |
| Conclusiones.....                               | 57 |
| CONCLUSIONES GENERALES .....                    | 58 |
| REFERENCIA BIBLIOGRÁFICAS .....                 | 60 |
| BIBLIOGRAFÍA.....                               | 62 |
| GLOSARIO DE TÉRMINOS .....                      | 65 |

|                                                                                     |    |
|-------------------------------------------------------------------------------------|----|
| TABLA 1: ESQUEMA DE MATRIZ BUS.....                                                 | 38 |
| TABLA 2: ESQUEMA DEL MERCADO DE DATOS.....                                          | 43 |
| TABLA 3: DISEÑO DE CASO DE PRUEBA "ANALIZAR INFORMACIÓN DE LA CAPA DE OZONO". ..... | 54 |
| TABLA 4: LISTA DE CHEQUEO.....                                                      | 56 |

|                                                                             |    |
|-----------------------------------------------------------------------------|----|
| FIGURA 1: EJEMPLO DE ESQUEMA DE ESTRELLAS.....                              | 9  |
| FIGURA 2: EJEMPLO DE ESQUEMA DE COPO DE NIEVE .....                         | 10 |
| FIGURA 3: MODELO DE ALMACENAMIENTO ROLAP.[5] .....                          | 13 |
| FIGURA 4: MODELO DE ALMACENAMIENTO MOLAP.[5].....                           | 14 |
| FIGURA 5: MODELO DE ALMACENAMIENTO HOLAP.[5] .....                          | 15 |
| FIGURA 6: ARQUITECTURA DE LA INTEGRACIÓN. ....                              | 44 |
| FIGURA 7: ARQUITECTURA DE LA INTEGRACIÓN. ....                              | 45 |
| FIGURA 8: EJEMPLO DE TRANSFORMACIÓN DE LA DIMENSIÓN GRUPOS TAXONÓMICOS..... | 46 |
| FIGURA 9: EJEMPLOS DE TRANSFORMACIÓN DE HECHO LLUVIA MEDIA.....             | 46 |
| FIGURA 10: VALIDACIÓN DE LAS REGLAS DE TRANSFORMACIÓN. ....                 | 47 |
| FIGURA 11: DISEÑO DEL TRABAJO. ....                                         | 48 |
| FIGURA 12: DISEÑO DE LOS CUBOS. ....                                        | 49 |
| FIGURA 13: DISEÑO DE LOS LIBROS DE TRABAJO.....                             | 50 |
| FIGURA 14:ROL ANALISTA.....                                                 | 52 |
| FIGURA 15: ROL ADMINISTRADOR .....                                          | 52 |
| FIGURA 16: COMPORTAMIENTO DE LOS INDICADORES POR SECCIONES.....             | 57 |

## INTRODUCCIÓN

En el mundo actual, las tecnologías de la información forman parte indispensable en la vida del ser humano. Debido a la gran competencia existente en el mercado internacional, entre las compañías dedicadas al desarrollo de dichas tecnologías, surge la necesidad de adaptarse constantemente a los cambios que ocurren en este sentido. Las soluciones informáticas traen mejoras considerables para el entorno en el cual se aplican, ahorrando tiempo, recursos y generando ganancias.

Cuba, como país en vías de desarrollo, se esfuerza en garantizar el bienestar de la sociedad, siendo la educación y la salud dos de sus grandes prioridades. El auge alcanzado a nivel mundial por la Tecnologías de la Informática y las Comunicaciones (TIC), traza nuevas pautas en función de la informatización del país. En la actualidad se trabaja para alcanzar un desarrollo de punta, guiado por la telefonía, internet y la computación. Las TIC están produciendo importantes transformaciones a nivel nacional, hasta el punto de marcar una característica distintiva entre este momento histórico en relación con el pasado. Como parte de estos cambios, el país se dio la tarea de actualizar, tanto la tecnología existente, como la manera de recopilar y gestionar la información de muchas de las entidades, proveyendo a diferentes empresas de información oportuna y veraz, lo cual puede tanto facilitar la toma de decisiones como permitir el control de procesos críticos. Pero no solo se trabaja en la informatización de la industria cubana sino también de la sociedad.

La Universidad de las Ciencias Informáticas (UCI) ha dotado al país de un gran número de ingenieros capacitados en la producción de Software, los cuales desde sus años de estudio han aportado, no solo a la economía del país, sino también, al cumplimiento de uno de los principales objetivos para los cuales el compañero Fidel la creó, que es contribuir a la informatización de Cuba. La UCI devenida en motor impulsor para el desarrollo de la industria del software cubano y la economía del país, tiene diseñado un sistema productivo conformado por varios centros de desarrollo. Actualmente el Centro de Tecnologías de Gestión de Datos (DATEC) de la UCI trabaja en conjunto con la Oficina Nacional de Estadísticas (ONE) para hacer más eficiente el uso que hasta el momento le dan a las tecnologías informáticas.

La ONE fue creada en el 1994. Su misión es garantizar la producción de estadísticas de calidad a través del Sistema Estadístico Nacional ejerciendo una adecuada dirección, ejecución y control de la captación

de las cifras económicas y sociales, así como, una adecuada difusión de la información estadística de acuerdo con las necesidades del país.

La ONE de Cuba posee diversas deficiencias al almacenar y gestionar los datos que se generan para el área de Medio ambiente, estas dificultades hacen lentos y engorrosos los procesos de análisis de la información. Dentro de los principales problemas que presenta la ONE es que la misma posee información digital en formato .dbf que no puede ser consultada a no ser por un especialista de la informática y de la información, necesitando conocimiento del negocio para entender los ficheros. También existen múltiples versiones de los datos, que provoca que los mismos no estén debidamente integrados y la carencia de reportes flexibles con la información actualizada, lo que dificulta el trabajo manual por el gran número de documentos, aumentando las probabilidades de errores estadísticos. Se generan ficheros anuales de difícil interpretación y esta información se va acumulando año tras año, requiriendo un mayor esfuerzo para realizar su análisis.

La situación anteriormente expuesta conlleva al siguiente problema de la investigación: ¿Cómo contribuir a la toma de decisiones en el área de Medio ambiente en la Oficina Nacional de Estadísticas?

Objetivo general: Desarrollar el mercado de datos Medio ambiente del Sistema de Información de Gobierno que contribuya a la toma de decisiones.

El objeto de estudio es el almacén de datos.

El campo de acción estará centralizado en el mercado de datos para los indicadores de Medio ambiente de la Oficina Nacional de Estadísticas.

Objetivos específicos:

- Refinar el análisis y diseño del mercado de datos.
- Implementar el mercado de datos.
- Validar el mercado de datos.

Tareas de la Investigación:

- Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.
- Refinamiento del negocio.
- Refinamiento de requisitos.
- Refinamiento del modelo de datos.
- Refinamiento del subsistema de integración.
- Refinamiento del subsistema de visualización.
- Refinamiento de los casos de pruebas.
- Implementación del subsistema de integración.
- Implementación del subsistema de visualización.
- Aplicación de las listas de chequeo.
- Aplicación de las pruebas de aceptación.
- Aplicación de los casos de pruebas.

El presente trabajo de diploma está estructurado por cuatro capítulos que documentan cada tarea realizada en función del cumplimiento de los objetivos trazados para lograr el desarrollo del mercado de datos.

## Capítulo 1: Fundamentación teórica.

En este capítulo se describen los principales conceptos relacionados con el tema. Se realiza un estudio del estado del arte acerca del mercado de datos del área de información Medio ambiente, para comprobar cuál es la solución factible y similar al sistema a desarrollar. Se establecen la metodología y se especifican las herramientas a utilizar.

## Capítulo 2: Análisis y diseño

El capítulo se centra en la definición del negocio en el área de Medio ambiente. Se examinan los excel correspondientes a esta área para identificar los requisitos funcionales y no funcionales, además de los requisitos de información y los multidimensionales. Luego se definen las tablas de hechos, las medidas y las dimensiones para estructurar el modelo dimensional de la solución. Se realiza la matriz BUS para obtener una mejor visión de las relaciones de los hechos y las dimensiones.

## Capítulo 3: Implementación

En este capítulo se explica cómo se lleva a cabo el proceso para la implementación del mercado de datos. Para ello son definidos los esquemas y tablas correspondientes. Se muestra el tratamiento realizado a las reglas del negocio definidas con anterioridad, así como el diseño de las transformaciones y los cubos OLAP con sus respectivos reportes asociados. Se define los roles y usuarios para el sistema y se configura la seguridad para estos últimos.

## Capítulo 4: Validación

Se realiza el diseño de los casos de pruebas. Se aplica la lista de chequeo y se evalúan los resultados para validar la solución en aras de que cumpla con los requerimientos y las necesidades de los usuarios para los cuales está destinada la misma y de que contribuya a la toma de decisiones que es el principal objetivo de la misma.

## CAPÍTULO 1. FUNDAMENTACIÓN TEÓRICA

### Introducción

En este capítulo se describen los principales conceptos relacionados con el tema. Se realiza un estudio del estado del arte acerca del Mercado de Datos del área Medio ambiente, para determinar cuál es la solución más factible al sistema a desarrollar, además se establece la metodología y herramientas a utilizar.

### 1.1 Medio ambiente

El medio ambiente se refiere a todo lo que rodea a los seres vivos, está conformado por elementos biofísicos (suelo, agua, clima, atmósfera, plantas, animales y microorganismos), y componentes sociales que se refieren a los derivados de las relaciones que se manifiestan a través de la cultura, la ideología y la economía. La relación que se establece entre estos elementos es lo que, desde una visión integral, conceptualiza el medio ambiente como un sistema.

La conservación y empleo racional de los recursos naturales ofrece un reto para la humanidad. En el mundo han ocurrido cambios drásticos y cada día son mayores los daños que ocasionan las catástrofes ambientales, esto ha llevado a las naciones y a la comunidad internacional a la adopción de medidas globales, regionales y nacionales para prevenir, atenuar y controlar estos impactos. Por ello hoy en día el concepto de medio ambiente está ligado al de desarrollo sustentable. Esta relación permite entender los problemas ambientales y garantizar una adecuada calidad de vida para las generaciones presentes y futuras.

Cuba, a través de la ONE que es el órgano rector de estadísticas en el país, lleva históricamente datos que orientan a la nación y a todos sus habitantes sobre que área trabajar en la reducción de acciones negativas sobre el mundo en que viven. Para ello se realizan periódicamente una serie de resúmenes estadísticos con los cuales se mide el impacto de la sociedad sobre el Medio ambiente y viceversa. Las TIC facilitan ese trabajo permitiendo almacenar grandes volúmenes de información utilizando un espacio reducido a un dispositivo de almacenamiento. El empleo de estas en la ONE ha sido crucial pues agiliza y facilita el trabajo a realizar por parte de los especialistas de sus áreas. El mercado de datos propuesto

como solución al problema que presenta la ONE para el área de Medio ambiente, permitirá una vez culminado, mostrar de una forma más amena, organizada y representada mediante gráficas, toda la información que los usuarios quieran solicitar.

Para los especialistas de la ONE sería de gran ayuda contar con un almacén de datos que integre los mercados de datos construidos para cada área que posee la misma.

### 1.2 Almacén de datos

Un almacén de datos (AD) es una colección de información recogida de múltiples sistemas fuentes u operacionales dispersos y cuya actividad se centra en la “toma de decisiones”. 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 la toma de decisiones. [1]

Un AD se caracteriza por contener los datos orientados a temas, integrados, variante en el tiempo y no volátiles, que dan soporte al proceso de toma de decisiones.

**Organizado en torno a temas:** La información se clasifica en base a los aspectos que son de interés para la empresa.

**Integrados:** La integración de datos consiste en convenciones de nombres, codificaciones consistentes, medida uniforme de variables y no contiene datos redundantes.

**Dependiente del tiempo. Esta dependencia aparece de tres formas:**

- La información representa los datos sobre un horizonte largo de tiempo.
- Cada estructura clave contiene (implícita o explícitamente) un elemento de tiempo (día, semana, mes, etc.).
- La información, una vez registrada correctamente, no puede ser actualizada.

**No volátil:** El Almacén de datos sólo permite cargar nuevos datos y acceder a los ya almacenados, pero no permite borrar ni modificar los datos.

**El concepto del tiempo varía respecto a los sistemas tradicionales:** En los sistemas tradicionales, la caducidad o validez de los datos no suele exceder de dos o tres meses. En muchos casos los datos varían todos los días, mientras que en un AD los datos tienen una caducidad en cuanto a su uso mucho mayor, incluso de varios años. En los sistemas de gestión se trabaja con datos actuales, mientras que en un AD pueden verse como fotografías tomadas en distintos momentos, que no sufren actualizaciones. En general, los datos de un AD suelen contener algún tipo de referencia temporal. [2]

**La arquitectura de un almacén de datos está formada por tres componentes básicos que son:**

- Proceso de extracción, transformación y carga: Permite mover los datos desde múltiples fuentes tomando aquellos que son de utilidad para el almacén. Una vez que estos son extraídos, se transforman a una forma presentable y de valor para los usuarios, al terminar se cargan en el almacén.
- Procesos de gestión de datos: Se encarga de almacenar y procesar los datos en el almacén.
- Los servicios de consulta y acceso a la información almacenada.

**Ventajas del uso de los almacenes de datos:**

- La inversión que realiza una organización para una correcta implantación de un Almacén de datos conlleva un costo muy elevado, sin embargo, el retorno de la inversión es garantizado en gran medida.
- Mejoran la productividad de los responsables en la toma de decisiones de la organización debido a que:
  1. Los AD hacen más fácil el acceso a una gran variedad de datos.
  2. Se obtiene una base de datos clasificada por temas.
  3. Integración de información procedente de múltiples sistemas externos.

**Desventajas del uso de los almacenes de datos:**

1. Menospreciar el tiempo requerido para extraer, limpiar y cargar los datos en el almacén.
2. Problemas con los sistemas de origen de los datos.

3. Pueden suponer altos gastos, además de los gastos de mantenimiento que son muy elevados.
4. Pueden quedarse obsoletos relativamente pronto si los usuarios incrementan sus necesidades.
5. La construcción de un almacén de datos puede requerir de mucho tiempo.

Resulta fácil comprender que un AD no se refiere a una base de datos común, por lo que modelarlo también tiene una representación única.

### 1.2.1 Modelo multidimensional

La tecnología de los almacenes de datos debido a su orientación analítica, impone un procesamiento y pensamiento distinto, el cual se sustenta en un modelamiento de bases de datos propio, conocido como modelamiento multidimensional, el cual busca ofrecer al usuario su visión respecto a la operación del negocio. Modelamiento Dimensional es una técnica para modelar bases de datos simples y entendibles al usuario final. La idea fundamental es que el usuario visualice fácilmente la relación que existe entre los distintos componentes del modelo. [3]

#### Características del modelo multidimensional

La estructura básica de un AD está definida por dos elementos: esquemas y tablas.

**Tablas:** Como cualquier base de datos relacional, un AD se compone de tablas. Hay dos tipos básicos de tablas en el modelo multidimensional:

**Tablas de hechos:** contienen los valores de las medidas del negocio, por ejemplo: ventas promedio en dólares, número de unidades vendidas.

**Tablas dimensionales:** contienen el detalle de los valores que se encuentran asociados a la tabla hecho.

**Esquemas AD:** la colección de tablas en el almacén se conoce como esquema. Los esquemas caen dentro de dos categorías básicas: esquema de estrellas y esquema copo de nieve.

#### Esquema de estrellas

El modelo multidimensional es conocido también como esquema de estrellas, ya que su estructura es similar: una tabla central y un conjunto de tablas que la atienden radialmente. El nombre proviene debido a que el diagrama forma una estrella, con puntos radiales desde el centro. El centro de la estrella consta de una o más tablas hechos y las puntas son las tablas dimensionales. Las tablas dimensionales sólo tienen conexión con la tabla hecho y ninguna más. [3]

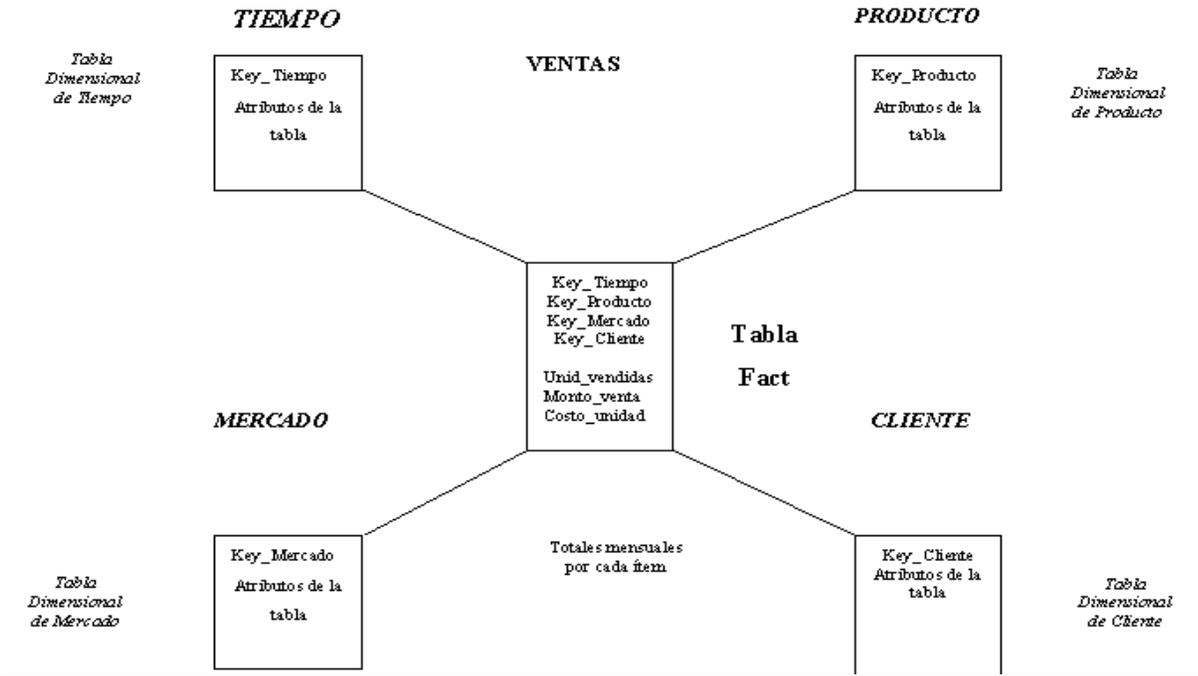


Figura 1: Ejemplo de esquema de estrellas.

Esquemas de copo de nieve

La diferencia del esquema de copo de nieve comparado con el esquema de estrellas, está dada en la estructura de las tablas dimensionales: las tablas dimensionales en el esquema de copo de nieve están normalizadas. Cada tabla dimensional contiene sólo el nivel que es clave primaria en la tabla y la llave foránea de su parentesco del nivel más cercano del diagrama. [3]

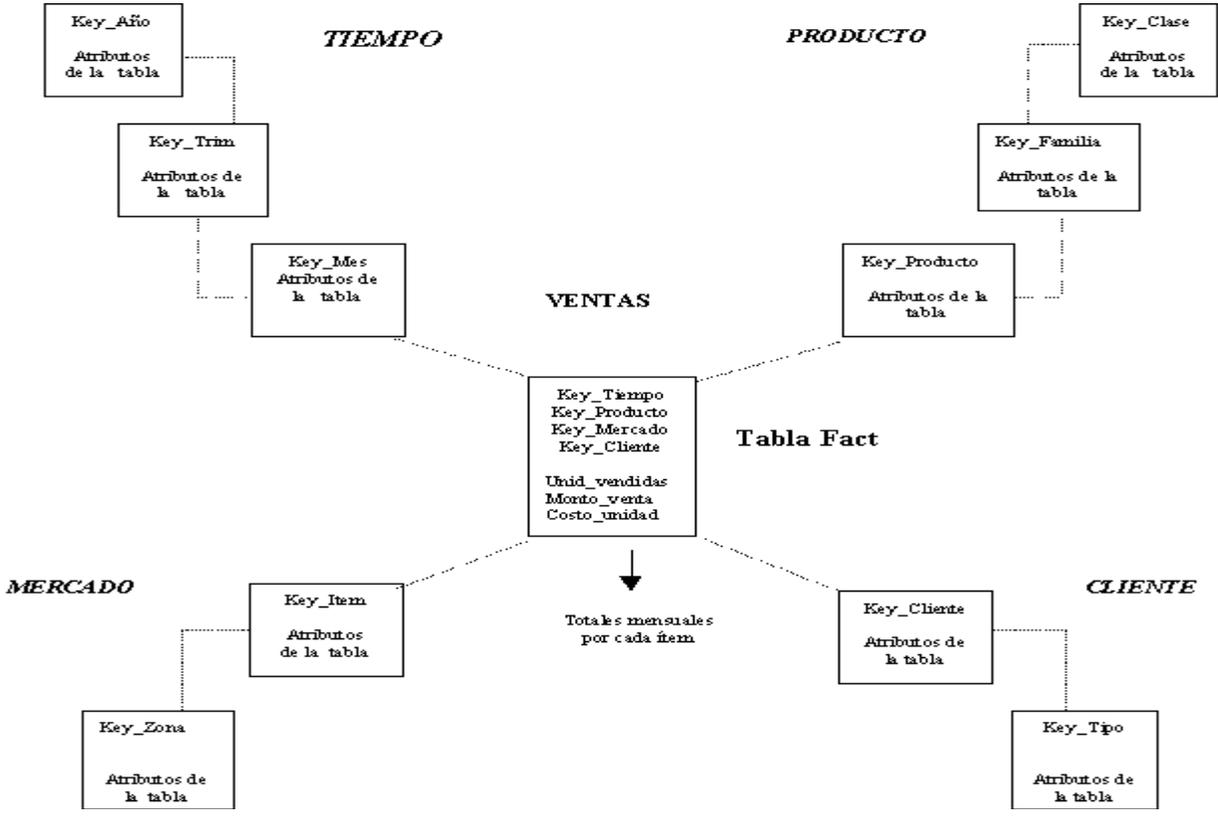


Figura 2: Ejemplo de esquema de copo de nieve

De estos esquemas el empleado ha sido el de estrella pues no se definió ninguna relación entre una dimensión y otra, por lo que todas las dimensiones están asociadas directamente a los hechos.

Este tipo de modelado, no solo se utiliza en la confección propia de un AD, debido a que uno de los términos más abordados cuando se habla de AD es el de Mercado de Datos (MD). Cuando se trata de almacenar información, los MD son posibles soluciones que están muy vinculadas a los AD.

### 1.2.2 Mercados de datos

El MD es un subconjunto de un almacén relativo a los requisitos de un departamento o área de negocio concreto. Este subconjunto de datos puede funcionar de forma autónoma, o bien enlazado al AD. El

motivo por el cual se crean MD es el crecimiento que tiene el almacén y así facilitar su construcción y utilización. [4]

### **Características de los MD:**

1. Los MD se enfocan en los requisitos de los usuarios que están asociados a un departamento específico de la empresa.
2. Su utilización y comprensión es sencilla debido a que contienen menor número de información que los almacenes de datos.

### **Ventajas del uso de los MD**

1. Poseen menos volumen de información, lo que posibilita mayor rapidez en las respuestas a las consultas y que su elaboración sea más fácil.
2. Se realizan sobre ellos consultas MDX y/o SQL sencillas que facilitan el acceso a los datos que son utilizados con frecuencia.
3. Facilita el resguardo de datos históricos.
4. Validación directa de la información.

Conocidos los conceptos abordados en los anteriores epígrafes y teniendo en cuenta que tanto los AD como los MD surgen para almacenar información que se empleará por largo tiempo, es necesario abordar temas relacionados con los datos como el que se explica a continuación.

### **1.2.3 Metadatos**

Existen varias definiciones que expresan el significado del término metadato. Algunas de las más difundidas son: datos sobre datos. También existen otros conceptos como informaciones sobre datos, datos sobre informaciones e informaciones sobre informaciones.

Los metadatos pueden describir colecciones de objetos y también los procesos en los que están involucrados, describiendo cada uno de los eventos, sus componentes y cada una de las restricciones que

se les aplican. Los metadatos definen las relaciones entre los objetos, como las tuplas en una base de datos o clases en orientación a objetos, generando estructuras.

### 1.2.4 Modos de Almacenamiento de Datos.

Procesamiento Analítico Relacional en Línea (Relational Online Analytical Process, "ROLAP"), Procesamiento Analítico Multidimensional en Línea (Multidimensional Online Analytical Process, "MOLAP") y Procesamiento Analítico Híbrido en Línea (Hybrid Online Analytical Process, "HOLAP") son tres modelos para el proceso analítico en línea (OLAP). En ellos el proceso de análisis se desarrolla de igual forma, variando en uno y otro caso, la metodología de almacenamiento. [1]

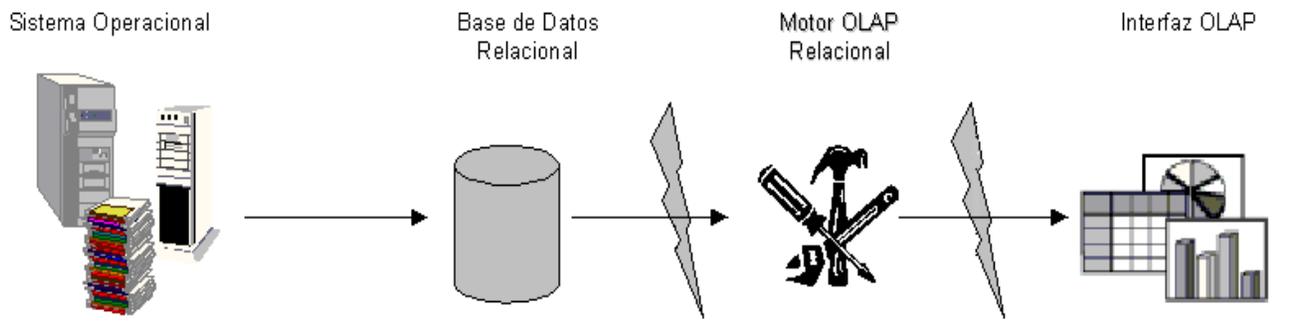
#### ROLAP

En el Procesamiento Analítico Relacional en Línea, los datos son almacenados en filas y columnas de forma relacional.

Los datos de los usuarios se presentan en forma de dimensiones de negocio. La semántica de las etiquetas de los metadatos es creada para encubrir las estructuras de almacenamiento y mostrar los datos dimensionales. Estas soportan el mapeo de las dimensiones a las tablas relacionales. Los metadatos son almacenados en tablas relacionales. Este modelo es utilizado fundamentalmente cuando la información no se consulta frecuentemente.

#### Características:

- Posee tablas relacionales que resumen los datos disponibles.
- Tiene un volumen alto de datos.
- Todos los datos de acceso están en la bodega de almacenamiento.
- Utiliza SQL complejo para los datos del depósito. Los cubos de datos son creados sobre la marcha por el servidor de análisis.
- Muestra vistas multidimensionales en la capa de presentación.
- Tiene un ambiente conocido.
- Disponibilidad de las herramientas.



**Figura 3: Modelo de almacenamiento ROLAP.[5]**

## MOLAP

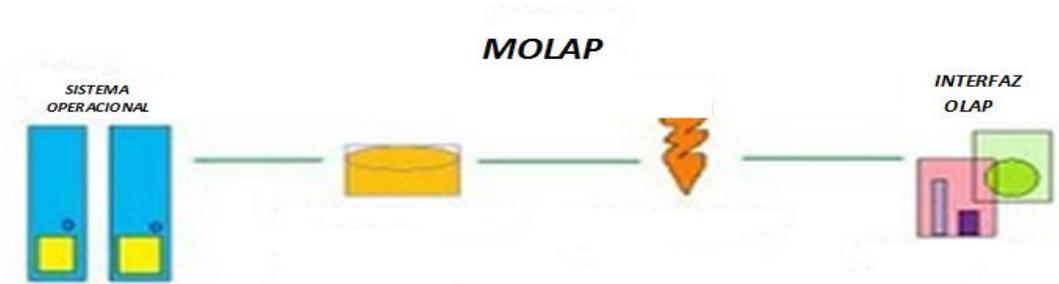
El Procesamiento Analítico Multidimensional en Línea almacena los datos de forma dimensional a diferencia del ROLAP. MOLAP utiliza una arquitectura de dos niveles: las bases de datos multidimensionales y el motor analítico. La BD multidimensional es la encargada del manejo, acceso y obtención del dato.

Las estructuras de almacenamiento son una copia de la fuente de datos y físicamente permanecen en la misma estación de trabajo donde está instalada la herramienta AD. Esto provoca que el acceso a la información sea de manera rápida y efectiva, donde el tiempo, en velocidad de respuesta, es crítico.

### Características:

- Diversos resúmenes de datos en bases de datos propietarias.
- Volumen de datos moderados.
- Resúmenes de acceso a datos detallados en bases de datos Multidimensionales.
- Crea cubos de datos prefabricados por el motor OLAP.
- Usa tecnología propietaria para almacenar las vistas multidimensionales en arreglos.
- Utiliza una matriz de alta velocidad para la recuperación de datos.
- Maneja poca tecnología de matriz de datos para gestionar los resúmenes.
- Tiene una fuente de librería de funciones para el cálculo complejo.

Facilita el análisis independientemente de la cantidad de dimensiones.



**Figura 4: Modelo de almacenamiento MOLAP.[5]**

## HOLAP

El Procesamiento Analítico Híbrido en Línea es una combinación entre los métodos de ROLAP y MOLAP que permite almacenar una parte de los datos como un sistema ROLAP y el resto como uno MOLAP. El control que ejerce el operador de la aplicación sobre este aprisionamiento varía de un producto a otro.

El aprisionamiento vertical almacena las agregaciones como un MOLAP para mejorar la velocidad de las consultas, detallando los datos en ROLAP para optimizar el tiempo en que se procesa en cubo. El aprisionamiento horizontal en el modo HOLAP almacena una sección de datos, generalmente los más recientes en modo MOLAP para mejorar la velocidad de consulta y los datos más antiguos en ROLAP. Se pueden almacenar algunos cubos en MOLAP y otros en ROLAP.

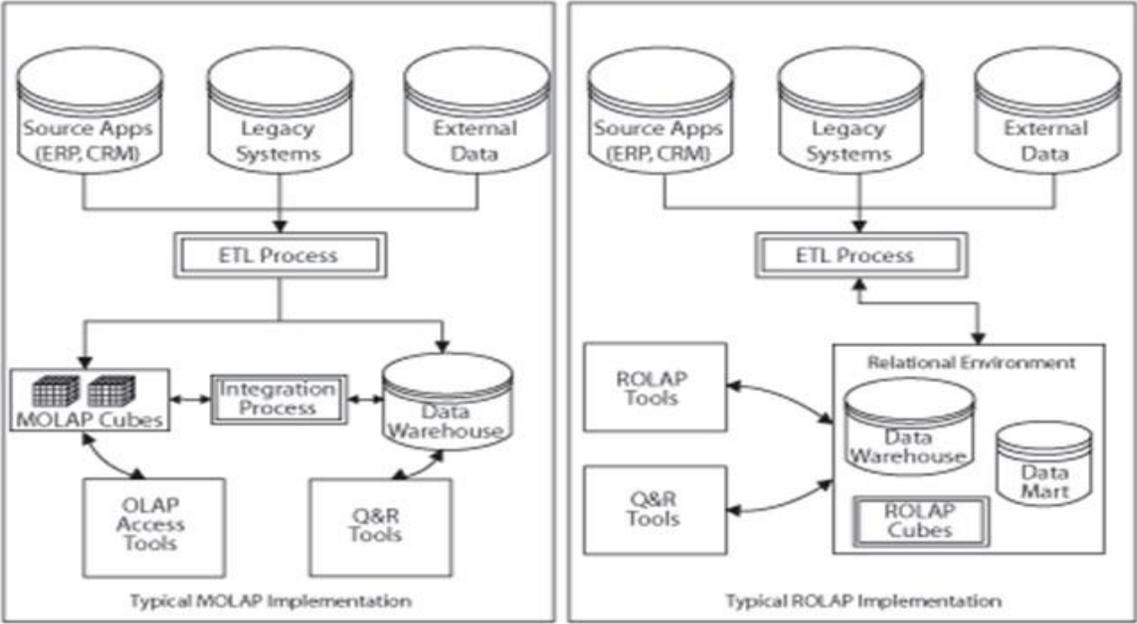


Figura 5: Modelo de almacenamiento HOLAP.[5]

Se utiliza ROLAP porque es el modo de almacenamiento establecido para PostgreSQL, siendo este el Gestor de Base de datos definido para la solución del Mercado de Datos del área de Medio ambiente de la ONE. Posee organización física que se implementa sobre tecnología relacional. ROLAP guarda la información en bases de datos relacionales, aprovechando así la tecnología relacional, permitiendo usar la integridad y seguridad de los sistemas gestores de bases de datos relacionales y gratuitos, y es capaz de manejar grandes volúmenes de datos.

1.2.5 Metodologías para el desarrollo

El vocablo metodología, es la ciencia que estudia los métodos del conocimiento. Se refiere a los métodos o procedimientos de investigación que se siguen para alcanzar una gama de objetivos en una ciencia. Además, puede categorizarse como el conjunto de métodos que se rigen en una investigación científica o en una exposición doctrinal. En la tecnología de Almacenes de Datos se han destacado un conjunto de metodologías que caracterizan y orientan todo el proceso de desarrollo. Existen criterios que han marcado fuertemente su tendencia, manejando a la comunidad mundial en este tema. Dichas metodologías son

conocidas como la Metodología Kimball por su creador Ralph Kimball y la de Inmon nombrada también por su creador Bill H. Inmon. [6]

Como propuesta de desarrollo a utilizar en la Línea de Almacenes de Datos e Inteligencia de Negocios (BI) del Centro de Tecnologías de Gestión de Datos (DATEC), se tomó la Metodología para el diseño de Kimball y se tuvo en cuenta por lo siguiente:

Crea los conceptos de hechos y dimensiones, que garantiza que sea eficaz el proceso de toma de decisiones y proporciona mayor agilidad en el proceso de desarrollo.

Propone incluir el Almacén de datos a través de la construcción de los Mercados de Datos departamentales, lo que forma una buena estrategia y coincide con la división lógica de las empresas, entidades y organismos.

La misma cubre todas las fases por las que pasa la construcción de un Almacén de datos, desde el levantamiento de información inicial hasta la Capa de Visualización.

Es importante mencionar que es una propuesta de metodología resistente y adaptable ante los cambios.

### **1.3 Integración de datos**

Según Ralph Kimball, es el proceso de unificación de los datos provenientes de múltiples fuentes. En el proceso de integración, el principal problema lo constituye la diversidad de los datos, los cuales se encuentran dispersos y generalmente no estandarizados. Esto provoca que la información proveniente de los sistemas externos sea inconsistente y de baja calidad, convirtiendo este proceso en un trabajo costoso y complejo. Existen tres maneras de realizar este proceso:

- Replicación de datos.
- Integración de Información Empresarial (EII).
- Extracción, Transformación y Carga de datos (ETL).[7]

### 1.3.1 Replicación de datos

“La replicación de datos consiste en el transporte de datos entre dos o más servidores, permitiendo que ciertos datos de la base de datos estén almacenados en más de un sitio, y así aumentar la disponibilidad de los datos y mejorar el rendimiento de las consultas globales. El modelo de replicación está formado por: publicador, distribuidor, suscriptor, publicación, artículo y suscripción; y varios agentes responsabilizados de copiar los datos entre el publicador y el suscriptor. A los tipos básicos de replicación (de instantáneas, transaccional y de mezcla), se le incorporan opciones para ajustarse aún más a los requerimientos del usuario”. [1]

### 1.3.2 Integración de Información Empresarial:

En este proceso, los datos se mantienen en la fuente sin modificar su origen, o sea, no se mueve la información permanentemente a un almacén. Para ello se establece una interfaz sobre las fuentes, de modo que en tiempo real se integren datos de las mismas y que sean recuperados como un sistema de información heterogéneo virtualmente integrado.

Por lo general este tipo de solución consiste en crear un mediador o intermediario de tal forma que contengan directorios de bases de datos y que a su vez sirvan de canal de consulta y representación de la información recuperada. Es el mejor método para la Integración de datos, ya que la información es capturada en tiempo real de las fuentes de datos, lo cual implica que las fuentes deben tener una estructura tecnológica sólida y bien establecida. [8]

### 1.3.3 Extracción, Transformación y Carga de datos (ETL)

“Es la tecnología enfocada a la Integración de datos, tanto por lote como a tiempo real hacia Almacenes de Datos”. [7] Se emplea además para migrar y sincronizar información de diversas aplicaciones y también para convertir bases de datos de un tipo o formato a otro. Por estas características este es el modo de integración propuesto para el mercado de datos Medio ambiente. Este proceso está conformado por tres subprocesos que se unen para como un todo lograr la integración.

### **Extracción**

El proceso de extracción consiste en obtener los datos de la fuente destino sin tener en cuenta sistemas ni hardware. Los datos se llevan a un formato establecido necesario para dar paso al proceso de limpieza y transformación. Pero antes se validan los mismos, se aceptan o se rechazan de acuerdo con lo deseado.

### **Limpieza y Transformación**

En esta fase se aplican una serie de Reglas de Negocios sobre los datos extraídos, con el objetivo de convertirlos en datos aptos para ser cargados. Aquí es necesario lograr una buena calidad de los datos y para ello es necesario el control de los valores válidos, garantizar la coherencia entre los valores, la eliminación de duplicaciones y comprobar que las reglas del negocio no han sido forzadas. Lo primero es garantizar la limpieza de los datos ya que estos pueden estar incompletos o ser incoherentes y para dar paso a la transformación de los mismos se debe lograr que sean consistentes y haber resuelto el problema de identidad que puedan presentar. Una vez realizada la limpieza, se aplican las reglas de transformación que pueden ser: combinar los datos de distintas fuentes, tratar valores nulos, buscar valores en distintas tablas, entre otras.

### **Carga**

El principal objetivo de esta fase es lograr que los datos estén listos para ser consultados. La información resultante de la fase anterior es incluida en el sistema de destino.

La integración de datos sirve como punto de partida para realizar el proceso de inteligencia de negocios, el cual una vez culminado permitirá mostrar los datos almacenados a través de una interfaz visual amigable que permita el análisis de los mismos.

### **1.4 Inteligencia de Negocios.**

El término Inteligencia de Negocios no escapa a la variedad de interpretaciones. Sin embargo, queda esencialmente claro que: “no es una metodología, software, sistema o herramienta específica, es más bien una colección de tecnologías que van desde arquitecturas para almacenar datos, metodologías,

técnicas para analizar información y software, entre otros, con un fin común para el apoyo a la toma de decisiones”. [9]

La Inteligencia de Negocios es conformada por todas las actividades relacionadas con la organización y entrega de información así como el análisis del negocio.

Entre estas actividades se encuentran:

- Sistemas de Almacenamiento de Datos (Almacén de datos y Mercados de Datos).
- Aplicaciones Analíticas.
- Minería de Datos.
- Herramientas de Procesamiento Analítico de Datos (OLAP).
- Herramientas de Consulta y Reporte de Datos.
- Herramientas de Producción de Reportes Personalizados.
- Herramientas de Extracción, Transformación y Carga de Datos (ETL).
- Herramientas de Administración de Sistemas.
- Portales de Información Empresarial.
- Sistemas de Bases de Datos.
- Sistemas de Administración del Conocimiento.

Estas soluciones proporcionan amplias ventajas al permitir que los datos se conviertan en un centro de beneficios, facilitando el análisis de información útil a las empresas, apoyando el proceso de toma de decisiones.

Las principales ventajas que ofrece una solución de Inteligencia de Negocio son:

- Control de costes, al tener un solo sistema que permite manejar fácilmente los distintos programas que se encuentran en los diferentes departamentos de su compañía.
- Mejora de la colaboración y la calidad de las decisiones, facilitando el acceso a la información en todos los niveles de la organización.
- Orienta las soluciones tecnológicas hacia el usuario, porque reduce los tiempos de aprendizaje mediante las herramientas de uso cotidiano.

- Proporciona una profunda visión del negocio a través de un sistema integrado: Cuadros de Mando Integrales, Cuadros de Mando Digitales, Reportes, OLAP y Minería de Datos.
- Asiste a los ejecutivos para planear y pronosticar el trabajo, presentando una descripción común de los procesos del negocio de una compañía.

### 1.4.1 Ejemplo del mundo real de aplicaciones de Inteligencia de Negocios

Twentieth Century Fox utiliza Inteligencia de Negocios para predecir qué actores, argumentos y películas serán populares en cada vecindario. Evitando ciertos argumentos en cines específicos, la compañía tiene ahorros de aproximadamente 100 millones de dólares alrededor del mundo cada año. Esa misma tecnología la utilizan para seleccionar los "trailers" (cortos previos a la presentación de una película) alternativos para cada película en cada cine y así maximizar las ventas. Una película puede tener varios trailers diferentes, cada cual puede percibirse de diferente forma por cada tipo de audiencias. [10]

Existen una gran cantidad de casos de aplicación exitosa de sistemas de Inteligencia de Negocios, proporcionándole no solo una notoria eficiencia y ventaja competitiva, sino que a la larga puede ser la diferencia entre la supervivencia o desaparición de la empresa.

## 1.5 Herramientas para el desarrollo

### 1.5.1 Sistema gestor de base de datos

PostgreSQL es un Sistema de Gestión de Bases de Datos Objeto-Relacionales (ORDBMS) que ha sido desarrollado de varias formas desde 1977. Comenzó como un proyecto denominado Ingres en la Universidad Berkeley de California. En 1986 otro equipo dirigido por *Michael Stonebraker*<sup>1</sup> continuó el desarrollo del código de Ingres, para crear un sistema de bases de datos objeto-relacionales llamado Postgres. En 1996, debido a un nuevo esfuerzo de código abierto y a la incrementada funcionalidad del software, Postgres fue renombrado a PostgreSQL. El proyecto PostgreSQL sigue actualmente un activo proceso de desarrollo a nivel mundial, gracias a un equipo de desarrolladores y contribuidores de código

---

<sup>1</sup> Científico especializado en bases de datos de investigación y desarrollo. Fue profesor de Ciencias de la Computación en la Universidad Berkeley de California y ayudó a crear la mayoría de las bases de datos relacionales del mercado existentes. Es fundador de Ingres, ilustraciones, Cohera, entre otros. También es un editor para el libro de lecturas en sistemas de bases de datos.

abierto. Está ampliamente considerado como el sistema de bases de datos de código abierto más avanzado del mundo.

El PostgreSQL corre en casi todos los sistemas operativos: Linux, Unix, BSDs, Mac OS, Beos y Windows. Cuenta con una documentación muy bien organizada, pública y libre, con comentarios de los propios usuarios, es altamente adaptable a las necesidades del cliente. Trabaja con varios lenguajes de programación como son: PHP, C, C++, Perl, Python, entre otros. Soporta todas las características de una base de datos profesional (triggers, funciones, secuencias, relaciones, reglas, tipos de datos definidos por usuarios, vistas materializadas). PostgreSQL funciona con grandes cantidades de datos y una alta concurrencia, con muchos usuarios accediendo a la vez el sistema. Sus características técnicas la hacen uno de los servidores de bases de datos más potentes y robustos del mercado. Su desarrollo comenzó hace muchos años y durante ese tiempo, estabilidad, potencia, robustez, facilidad de administración e implementación de estándares, han sido las características que más se han tenido en cuenta durante su desarrollo.

Ventajas de PostgreSQL:

- Se ajusta al número de CPU y la cantidad de memoria que ofrece el sistema de forma óptima, soportando mayor cantidad de peticiones simultáneas sin provocar errores.
- Comprueba la integridad referencial así como también la de almacenar procedimientos en la propia base de datos.
- Permite la creación de funciones personalizadas, el manejo y configuración de disparadores.

Algunas desventajas son:

- La velocidad de respuestas que ofrece este gestor es un poco deficiente.
- Tienen un límite de 8k por filas que se pueden aumentar hasta 32 lo que trae consigo que disminuya su rendimiento.

### 1.5.2 Herramienta manejadora de base de datos

Para administrar las bases de datos en PostgreSQL 8.4, se escogió la herramienta PgAdmin3 versión 1.10 porque es un software libre que accede a todos los objetos del PostgreSQL y responde las necesidades de todos los usuarios.

Algunas características de este manejador de datos son:

- Cuenta con una sintaxis SQL editor, servidor en el editor de código.
- La conexión es utilizando TCP/IP o Unix Domain Sockets y no requiere de drives para comunicarse con el servidor.
- Muy rápido para la visualización y entrada de datos.
- Procedimientos almacenados.
- Transacciones.

### 1.5.3 Herramienta para el modelado

Cuando se hace la planificación de la base de datos, la primera etapa del ciclo de vida de las aplicaciones de bases de datos, también se puede escoger una herramienta de ingeniería de software orientada por computadora (CASE de sus siglas en ingles Computer Aided Software Engineering) que permita llevar a cabo el resto de tareas del modo más eficiente y efectivo posible. El uso de las herramientas CASE puede mejorar la productividad en el desarrollo de una aplicación de bases de datos. Una herramienta CASE suele incluir:

- Un diccionario de datos para almacenar información sobre los datos de la aplicación de bases de datos.
- Herramientas de diseño para dar apoyo al análisis de datos.
- Herramientas que permitan desarrollar el modelo de datos corporativo, así como los esquemas conceptual y lógico.
- Herramientas para desarrollar los prototipos de las aplicaciones. [11]

La utilización de herramientas de modelado permite gestionar y mantener aplicaciones que trabajan con un gran volumen de datos. Entre estas herramientas se encuentra el Visual Paradigm que permite transformar, migrar e integrar amplias cantidades de datos procedentes de diversas fuentes de la empresa.

### **Visual Paradigm**

Es una herramienta CASE de modelado que utiliza UML (Lenguaje Unificado de Modelado). Soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue. Los sistemas de modelado UML ayudan a una rápida construcción de aplicaciones de calidad y a un menor coste. Permite dibujar todos los tipos de diagramas de clases, ingeniería inversa, generar código a partir de diagramas, generación de objetos a partir de bases de datos y generación de bases de datos a partir de diagramas de entidad relación. Posee una plataforma capaz de integrarse con Eclipse, NetBeans, Oracle JDeveloper, JBuilder, IntelliJ IDEA, WebLogic Workshop, Microsoft Visual Studio, entre otras.

La herramienta es colaborativa, es decir, soporta múltiples usuarios trabajando sobre el mismo proyecto; genera la documentación del proyecto automáticamente en varios formatos como Web o Pdf, y permite el control de versiones. Cabe destacar igualmente su robustez, usabilidad y portabilidad. El Visual Paradigm es un producto de calidad, soporta aplicaciones Web, genera código para Java y es fácil de instalar y actualizar.

### **1.5.4 Herramienta para la integración de datos**

Para la integración de datos se escogió el Pentaho Data Integration porque es una herramienta de código abierto adoptado por Pentaho BI. Proporciona la extracción de gran alcance, Transformación y Carga (ETL) utilizando un enfoque innovador, orientado a los metadatos. Con una interfaz intuitiva, gráfica de arrastre, una probada arquitectura escalable y basada en estándares, Pentaho Data Integration es cada vez más la elección de las organizaciones tradicionales. [12]

Presenta un enfoque orientado a los metadatos que simplemente es indicar qué quiere hacer, pero no cómo desea hacerlo. Ahora los desarrolladores ETL, BI y los administradores pueden crear complejas

transformaciones y el empleo en un entorno gráfico, arrastrar y soltar sin tener que generar ningún código personalizado.

El Pentaho Data Integration se utiliza para:

- Poblar el Almacén de datos.
- Exportación de bases de datos a texto, archivos u otras bases de datos.
- Importación de datos en bases de datos, que van desde archivos de texto a hojas de excel.
- Migración de datos entre aplicaciones de base de datos.
- Exploración de datos en bases de datos existentes. (tablas, vistas, sinónimos).
- Enriquecimiento de información por la búsqueda de datos en varios almacenes de información (bases de datos, archivos de texto, hojas de excel).
- Limpieza de datos mediante la aplicación de condiciones complejas en transformaciones de datos.
- Integración de aplicaciones.

Entre las ventajas de la utilización del Pentaho Data Integration se encuentran:

- Capacidad para redireccionar filas erróneas sin afectar al rendimiento.
- Capacidad de lectura y grabación de archivos directamente desde la URL usando Apache VFS.
- Mejora la organización y navegación de pasos en la realización del trabajo.[13]

### 1.5.5 Herramientas para Inteligencia de Negocios

#### **Pentaho Schema Workbench**

Pentaho Schema Workbench es un entorno visual para el desarrollo y prueba de cubos OLAP Mondrian. Si bien la definición del XML para esquemas Mondrian no es extremadamente compleja, en la práctica resulta engorroso recordar cada uno de los elementos junto a sus atributos y sub-elementos. Con esta aplicación, se puede configurar una conexión JDBC como el modelo físico, para luego elaborar el esquema lógico de manera simple y efectiva. Permite crear y probar los cubos OLAP visualmente para que luego el motor de Mondrian procese las solicitudes MDX con los esquemas creados. Los esquemas

son modelos metadatos XML, que se crean en una estructura específica utilizada por el motor de Mondrian. [14]

Ofrece las siguientes funcionalidades:

- Editor de esquemas integrados con un origen de datos subyacente para su validación.
- Prueba de consultas MDX contra el esquema y la base de datos.
- Examinar la estructura subyacente de bases de datos.

### **Mondrian OLAP Server**

Para obtener la funcionalidad de OLAP se utilizan otras dos aplicaciones: el servidor OLAP Mondrian, que combinado con Jpivot, permiten realizar consultas al AD y permite que los resultados sean presentados mediante un navegador, de modo que el usuario pueda realizar las actividades típicas de navegación. Mondrian utiliza MDX como lenguaje de consulta, que fue un lenguaje propuesto por Microsoft. Funciona sobre las bases de datos estándares del mercado: Oracle, DB2, SQL-Server, MySQL, etc., lo cual habilita y facilita el desarrollo de negocios basados en la plataforma Pentaho.

Mondrian es una de las aplicaciones más importantes de la plataforma Pentaho BI. Es un servidor OLAP open source, que gestiona comunicación entre una aplicación OLAP (escrita en Java) y la base de datos con los datos fuentes.

Entre sus ventajas están:

- Agilizar la consulta de grandes cantidades de datos.
- Alta velocidad de respuesta.
- Permite realizar consultas a MD.
- Es un motor ROLAP con caché. (Portada sobre la plataforma Pentaho Open)

### **Pentaho BI Server**

La plataforma Pentaho BI Server, provee el soporte y la infraestructura necesaria para crear soluciones de inteligencia empresarial a problemas de negocios. El marco proporciona los servicios básicos, incluidos

autenticación, registro, auditoría, servicios web y motor de reglas. La plataforma también incluye un motor de soluciones que integra reportes, análisis, tableros de comandos y componentes de minería de datos. El diseño modular y arquitectura basada en plug-in permite a todos o parte de la plataforma estar inmersa en aplicaciones de terceros por los usuarios finales, así como fabricantes de equipos originales.

La aplicación Pentaho BI Server funciona como un sistema basado en administración web de informes, el servidor de integración de aplicaciones y un motor de flujo de trabajo ligero (secuencias de acción). Está diseñado para integrarse fácilmente en cualquier proceso de negocio.

Algunas de sus ventajas son:

- Integración con procesos de negocios.
- Administra y programa reportes.
- Administra seguridad de usuarios.

### **El servidor Apache Tomcat**

Tomcat es el servidor web más utilizado a la hora de trabajar con Java en entornos web; es una implementación completamente funcional de los estándares de JSP y Servlets. También puede especificarse como el manejador de las peticiones JSP y Servlets recibidas por servidores webs populares, como el servidor Apache HTTP de la fundación de software de Apache o el servidor Microsoft Internet Information Server (IIS). Está integrado en la implementación de referencia Java 2 Enterprise Edition (J2EE) de Sun Microsystems. [15]

Las ventajas de Tomcat

Es gratis, es fácil de instalar, se ejecuta en máquinas más pequeñas y es compatible con las API más recientes de Java. Puede descargarse, instalarse y probarse en menos de una hora.

Pone a disposición de todo el mundo las últimas actualizaciones de Java. Por ejemplo, Tomcat 4.0 es compatible con JSP 1.2 y Servlet 2.3 (como WAS 5.0), Tomcat 3.x son compatibles con JSP 1.1 y Servlet 2.2. Además, como puede ejecutarse utilizando la JVM que se quiera, puede utilizarse JDK 1.4 (por cierto

que Java 1.4 ofrece aserciones, expresiones regulares, mejor rendimiento que Java 1.3 y una infraestructura de registro cronológico estándar). Tomcat 4.0 ahora tiene seguridad de nivel de aplicación, una aplicación de administración intuitiva basada en web, expresiones regulares compatibles con JDK 1.2 y 1.3 y, como se mencionaba antes, mejor escalabilidad y rendimiento.

### Conclusiones del capítulo

En el presente capítulo se realizó un bosquejo sobre el medio ambiente y la significación que tiene para el país manejar estadísticas del estado del mismo. Se abordaron los conceptos relacionados con almacenes de datos dejando constancia de las facilidades que proporciona la utilización de un mercado de datos para la solución del problema planteado. Los modos de almacenamiento también fueron objeto de estudio y se concluyó con la elección de ROLAP. Como metodología fue elegida Kimball con la adaptación establecida por la línea de Almacenes de Datos. Fueron abordados además los conceptos de integración de datos e inteligencia del negocio, dos fases muy importantes dentro de la construcción de un mercado de datos. El proceso de extracción, transformación y carga fue el escogido para la integración.

Se estudiaron todas las posibles herramientas a utilizar seleccionando:

- Visual Paradigm para el modelado del mercado de datos.
- PostgreSQL 8.4 como gestor de bases de datos.
- PgAdmin III versión 1.10 como herramienta manejadora de base de datos pues es el cliente para PostgreSQL 8.4.
- Pentaho Data Integration (PDI) versión 4.1.0 para la integración de los datos.

Para la inteligencia de negocio se emplearon:

- Pentaho Schema Workbench
- Mondrian OLAP Server
- Pentaho BI Server
- El servidor Apache Tomcat

### CAPÍTULO 2: ANÁLISIS Y DISEÑO

#### Introducción

Cuando se realiza cualquier producto informático la principal meta que se sigue es lograr que tenga la calidad requerida. A la hora de realizar un mercado de datos y teniéndose en cuenta que la calidad anteriormente mencionada se logra cuando se satisfacen las necesidades del cliente y el mismo queda complacido con el resultado obtenido, resulta de vital importancia la realización de una investigación previa que deviene en lo que se conoce como análisis del negocio. Una vez que se haya analizado se procede a realizar el diseño con el cual queda representado y establecido el mercado de datos.

En el presente capítulo se muestran los aspectos más relevantes del análisis y diseño realizado para el mercado de datos Medio ambiente de la ONE.

#### 2.1 Análisis de la Solución

##### 2.1.1 Definición del Negocio

La estructura institucional de la ONE está distribuida por todo el país. Existen 169 oficinas municipales que se subordinan a las 16 oficinas provinciales las cuales se encargan de interactuar directamente con los centros informantes.

El mecanismo de captación de la información está compuesto por varios procesos orientados a los tipos de modelos existentes. Los modelos de Estadísticas Continuas y Encuestas Periódicas son unos de estos procesos definidos. Las diferencias entre ambos se basan en los períodos de captura que puede ser mensual, trimestral, semestral, anual, entre otros, en dependencia de la labor que cada cual realice.

Los datos estadísticos de los indicadores de Medio ambiente se capturan mediante las Encuestas Periódicas del modelo 1003 de cada centro informante en los períodos mensual y anual. Estas oficinas están obligadas a entregar en la fecha establecida la información solicitada por la ONE; algunas de las fuentes son: Instituto Nacional de Recursos Hidráulicos; Cuerpo Nacional de Guardabosques del Ministerio del Interior y el Centro Nacional de Investigaciones Sismológicas.

La sustancial tarea de la ONE es absorber los datos estadísticos de todos los sectores para luego almacenarla y permitir la disposición de la misma, facilitando consultarla de una manera alígera y eficaz por lo que se puede proceder a identificar el tema de análisis.

### 2.1.2 Tema de Análisis Identificado

Para una mejor organización de la información recogida en el área Medio ambiente se divide la información de la misma en varias sub-temas de análisis que permite relacionarla. Estos son:

- Atmósfera
- Sustancias agotadoras de la capa de ozono
- Energía renovable
- Agua
- Carga contaminante
- Cobertura de agua potable y saneamiento
- Acueducto y alcantarillado
- Suelos
- Superficie cubierta de bosques
- Biodiversidad
- Áreas protegidas
- Desastres naturales
- Desechos sólidos
- Materias primas

### 2.1.3 Roles y permisos

Dentro de un sistema informático, la seguridad con que se maneje la información que en él se gestiona es de suma importancia. Para ello se gestionan los roles y permisos.

El rol puede ser entendido como el papel que pone en práctica la persona en el drama social o en un sentido más preciso, como el sistema de expectativas sociales que acompañan a la presentación pública de los sujetos de un determinado estado social o estatus. Un rol de un sistema informático, sería el papel que juega cada uno de ellos al interactuar con el sistema en función de los permisos que tiene al utilizar el producto. Básicamente existen tres roles predefinidos, los cuales son:

1. Analista: es la persona que utilizará el sistema para consultar la información referente a los indicadores ambientales.
2. Administrador: es la persona que podrá almacenar, mostrar y consultar los datos sobre el Indicador estadístico Medio Ambiente.

3. Especialista de ETL: es la persona que realizará el proceso de extraer, transformar y cargar los indicadores ambientales.

### 2.1.4 Reglas del Negocio

Las reglas del negocio describen las políticas, normas y restricciones presentes en un software, los cuales son de vital importancia para alcanzar los objetivos propuestos. Para un buen funcionamiento del producto propuesto es necesario seguir las siguientes reglas desglosadas:

#### Reglas de Negocio de Análisis

- En el excel 2.3 la lluvia total media por provincia se obtiene de la lluvia total media de cada mes del año correspondiente a dicha provincia.
- En el excel 2.3 la lluvia total media de Cuba se obtiene de la suma de las lluvias totales medias de cada mes del año en todo el país.
- En el excel 2.4 el total de ciclones de diferentes intensidades que han azotado Cuba en un período se obtiene de la suma del total de los ciclones por categoría.
- En el excel 2.4 el total de ciclones por mes de la temporada ciclónica en un período se obtiene de la suma de los ciclones de diferentes categorías que azotaron a Cuba en dicho mes.
- En el excel 2.4 el total de ciclones por categoría de un período de obtiene de la suma de los ciclones de dicha categoría de cada mes de la temporada ciclónica.

#### Reglas de Negocio de ETL

- Se pueden encontrar fechas que no están completas por lo que se cubrirá con el año solamente o acompañada con el día.
- Se pueden hallar horas que no están registradas por lo que se representará con tres puntos.
- En caso de que algún excel sea representado de la siguiente forma:
  - ✓ . No se efectuó la operación indicada por falta de algún dato.
  - ✓ .. Indicador no aplicable.
  - ✓ ... Cifras no disponibles al terminarse la redacción.

✓ - Resultado igual a cero.

Se accederá a cargar los datos como vacíos pues no se conocen sus valores y la asignación de algún valor puede provocar errores en las medidas calculables que se establecen en las reglas del negocio de análisis.

- En los excel se presentan casos en que los nombres de las provincias, las estaciones de monitoreo y los meses no están escritos de la misma manera que se definieron en las dimensiones.

### 2.1.5 Requisitos de información

Describen qué información debe almacenar el sistema para satisfacer las necesidades de clientes y usuarios. Identifican los conceptos relevantes sobre los que se debe almacenar información y los datos específicos que son de interés. Son los reportes que el cliente necesita visualizar. Fueron definidos sesenta y uno requisitos de información, a continuación se presentan:(Ver anexos requisitos de información)

RI1: Obtener las temperaturas máximas por provincias y estaciones de monitoreo, anualmente.

RI2: Obtener las temperaturas mínimas por provincias y estaciones de monitoreo, anualmente.

RI3: Obtener la dirección de los vientos predominantes, por provincias y estaciones de monitoreo, anualmente.

RI4: Obtener la cantidad de días con lluvia por provincias y estaciones de monitoreo, anualmente.

RI5: Obtener la nubosidad media por provincias y estaciones de monitoreo, anualmente.

### 2.1.6 Requisitos multidimensionales

Se realizará un análisis de cada uno de los indicadores ambientales para extraer todos los valores de entrada (VE) y salida (VS). Los VE son las informaciones como provincia, categorías, años, entre otras, que están estrechamente relacionadas con los indicadores ambientales porque dan a conocer su

comportamiento y los VS no son más que aquellos datos estadísticos que serán visibles para el usuario como pueden ser la cantidad de huracanes, frentes de frío, lluvias, entre otras.(Ver anexos requisitos multidimensionales)

### Atmósfera

- Principales indicadores del clima.

VE: provincias con sus estaciones y años; VS: temperatura máxima y mínima, rapidez del viento, cantidad de días con lluvias; por ciento de humedad relativa y nubosidad media.

- Temperaturas máximas y mínimas absolutas registradas.

VE: provincias con sus estaciones y período; VS: temperatura máxima y mínima.

### 2.1.7 Requisitos funcionales

Los requisitos funcionales son capacidades o condiciones que el sistema debe cumplir, lo que es muy importante para satisfacer las expectativas del cliente. De ahí a que se hayan definido los siguientes:

RF1-Extraer información de archivos excel del área Medio ambiente.

RF2-Transformar y cargar información de archivos excel del área Medio ambiente.

RF3-Insertar usuario.

RF4-Eliminar usuario.

RF5-Insertar rol.

RF6-Eliminar rol.

RF7-Insertar reporte.

RF8-Eliminar reporte.

RF9-Modificar reporte

RF10- Configurar elementos del cubo.

RF11- Ordenar resultados

RF12- Mostrar padres.

RF13- Mostrar propiedades

RF14- Suprimir filas

RF21- Editar consulta MDX

RF15- Invertir eje

RF22- Exportar reporte a .pdf

RF16- Detallar miembros

RF23- Exportar reporte a .xls

RF17- Entrar en detalles

RF24- Imprimir reporte

RF18- Mostrar datos de origen

RF25- Ver un reporte

RF19- Mostrar gráfico

RF26- Autenticar usuario.

RF20- Configurar gráfico

### 2.1.8 Requisitos no funcionales

Los requisitos no funcionales son propiedades o cualidades que el producto debe tener, y que lo hacen atractivo, usable, rápido o confiable. A continuación se listan los que debe cumplir el sistema en proposición.

RNF1: Software

- Sistema Operativo Windows 98 o superior y Linux.

RNF2: Hardware

- Procesador Pentium III o superior.
- 1Gb de RAM.
- 80Gb de capacidad en disco duro.

RNF3: Usabilidad

- El sistema podrá ser usado por la persona que gestione la información referente a los indicadores estadísticos por lo que es necesario que los usuarios del producto tengan un conocimiento básico del manejo de las computadoras y los sistemas operativos.
- La aplicación debe ser asequible, amena y funcional tanto para usuarios expertos como para novatos.

### RNF6: Eficiencia

- Los reportes y estadísticas que se obtendrán deben ser precisos y reales.

### RNF7: Portabilidad

- El producto deberá funcionar en los sistemas operativos Windows y Linux.

### RNF8: Restricciones en el diseño y la implementación

- El producto se desarrollará con PostgreSQL 8.4.
- La herramienta manejadora de datos será PgAdmin3 1.10.
- Para el modelado de los datos se utilizará el Visual Paradigm.

### RNF9: Sesión de usuarios.

- Los permisos correspondientes a los usuarios se activarán una vez que estos se autenticuen teniendo acceso sólo a la información que le compete de acuerdo con sus privilegios.

### 2.1.9 Casos de uso del sistema

Describir el sistema es significativo para el desarrollo del producto pues se muestra a través de una representación gráfica y descriptiva el funcionamiento de este. En el mismo desempeña un papel primordial el actor que interactúa con los casos de usos bajo la forma de acciones y reacciones. Por lo tanto, se establece una relación entre los casos de usos y actores sobre las condiciones y requisitos que debe cumplir el sistema.

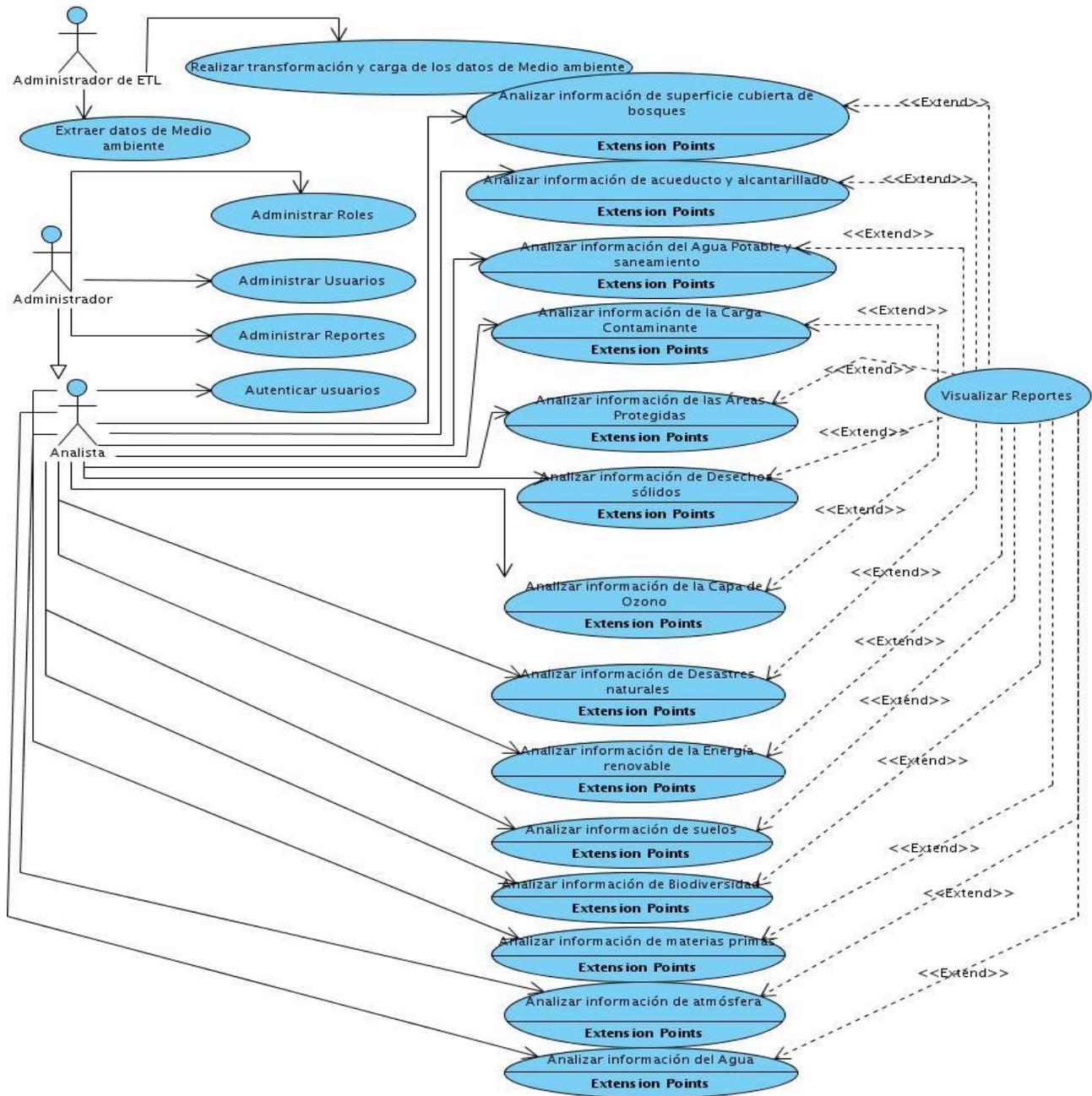


Figura 6. Diagrama de casos de uso.

### 2.2 Diseño de la Solución

#### 2.2.1 Modelo de Datos

El modelo de datos se diseña teniendo en cuenta los requisitos de información y los requisitos multidimensionales obtenidos anteriormente. Para ello se identifican las tablas de dimensiones y de hechos con las medidas correspondientes a estas últimas en aras de lograr un buen diseño. (Ver anexos modelo de datos).

#### Tablas de hechos:

Luego de haber estudiado los epígrafes anteriores, se extrajeron las temáticas principales con que cuenta el Medio Ambiente. A continuación se muestran las tablas de hechos que servirán de repositorio central a diferentes dimensiones.

#### Tabla de hecho clima:

- El clima recoge varios valores estadísticos relacionados con las temperaturas, cantidad de días con lluvia, rapidez del viento entre otras, tiene asociación con las dimensiones: Provincias, estaciones y años.

#### Tabla de hecho temperatura:

- La temperatura es el valor más alto o más bajo respectivamente, entre los valores máximos o mínimos de temperatura diaria, esta tabla se asocia a varias dimensiones como son: temporal, Provincias, estaciones y el período.

#### Dimensiones

Luego de la obtención de las tablas de hechos, se precisaron múltiples dimensiones que posteriormente se utilizará para la realización del diseño dimensional; entre ellas se encuentran:

**Años (dim\_temporal\_annos):** Se almacenan los años de la información.

Atributos asociados:

- dim\_temporal\_anno\_id: Identificador de la dimensión.
- anno\_numero: Almacena el número de cada año.
- anno\_codigo: Se almacena el código de cada año.
- anno\_numero: Almacena el número de los años de la información.

**DPA (dim\_dpa):** Se almacenan todas las provincias incluyendo el país.

Atributos asociados:

- dim\_provincia\_id: Es la llave primaria de la dimensión.
- provincia\_codigo: Almacena el código oficial establecidos por provincias
- provincia\_nombre: Se almacena el nombre de cada provincia.
- provincia\_descripcion: Se almacena la descripción de las provincias.

### Medidas

Las medidas definidas se almacenarán en las tablas de hechos, los extraídos de cada temática aparecen a continuación.

#### Tabla de hecho clima:

- Temperatura mínima media
- Temperatura máxima media
- Cantidad de días con lluvia
- Dirección del viento
- Rapidez del viento
- Nubosidad media
- Humedad Relativa
- Lluvia total

#### Tabla de hecho temperatura:

- Temperatura mínima absoluta registrada

- Temperatura máxima absoluta registrada
- Fecha máxima
- Fecha mínima

**2.2.2 Matriz BUS o Matriz Dimensional**

Al instante de realizar el modelo lógico del trabajo correspondiente al Indicador ambiental, se perpetrará la matriz BUS que demostrará a través de una tabla las relaciones entre los hechos y las dimensiones. Para ejemplificarlo se muestra una parte de la matriz resultante, la cual brinda una mejor visión de las dimensiones que quedarán relacionadas con cada hecho.

| <b>Hechos</b>          | Años | mes | Provincia | Regiones | Periodo | Estaciones |
|------------------------|------|-----|-----------|----------|---------|------------|
| Clima                  | x    |     | x         |          |         | x          |
| Temperatura            |      |     | x         |          | x       | x          |
| Lluvias                | x    | x   | x         |          |         |            |
| huracanes meses        |      | x   |           |          | x       |            |
| Huracanes regiones     |      |     |           | x        | x       |            |
| frentes fríos meses    |      | x   |           |          | x       |            |
| Frentes fríos regiones |      |     |           | x        | x       |            |
| Concentración de gases | x    |     |           |          |         | x          |
| Acidez de la lluvia    | x    |     |           |          |         | x          |

**Tabla 1: Esquema de matriz bus.**

**Conclusiones del capítulo**

Se definió el proceso del negocio para conocer el comportamiento del flujo de información del Indicador Medio Ambiente y el tema de análisis se centralizó en este Indicador, de la información capturada de esta

temática se desarrollaron todos los epígrafes restantes. Se realizó un largo estudio de cada uno de los excel ambientales para extraer todas las informaciones que ayudarán a desarrollar el mercado de datos. Se extrajeron las reglas del negocio, se establecieron 61 requisitos de información y las variables de entrada y salida de cada tabla de excel. Se describieron los requerimientos del sistema. Se detallaron los requisitos funcionales y no funcionales extraídos. Para entender el funcionamiento del sistema se realizaron los diagramas y las descripciones de casos de usos, además se determinaron los roles y las acciones que debe hacer al interactuar con el sistema. Culminado el análisis se prosiguió a realizar el diseño del modelo dimensional; para ello se identificaron las medidas, tablas centrales, dimensiones y se realizó la matriz BUS para relacionarlos. Simultáneamente con estas fases se fueron confeccionando los artefactos definidos para documentar las mismas. Una vez concluido este capítulo quedan establecidas las bases para proceder a la realización de la implementación del mercado de datos.

## CAPÍTULO 3: IMPLEMENTACIÓN

### Introducción

En este capítulo se implementarán los subsistemas de integración donde se realizarán las transformaciones necesarias para la carga de los datos, los cuales se llevan a cabo en el proceso de ETL. Además, se le dará solución a los requisitos de información a través de la implementación de los subsistemas de visualización.

### 3.1 Implementación del modelo de datos.

#### Estructuras de datos

Una estructura de datos es una forma de organizar un conjunto de datos. Se caracteriza por las funciones que se usan para almacenar y acceder a los elementos individuales de los datos. Para un mejor entendimiento de la solución se estructuraron los datos por esquemas y tablas.

#### Esquemas y tablas

Para el desarrollo de la solución se cuenta con dos esquemas: el esquema dimensiones, que contiene todas las dimensiones de las tablas que se encuentran definidas por la ONE, y el esquema mart\_medio\_ambiente, que agrupa todos los hechos y las dimensiones que son únicas para cada mercado de datos.

El modelo físico está estructurado por 75 tablas en total: 35 tablas de dimensiones, 40 tablas de hechos. Todas ellas están distribuidas en los esquemas propuestos anteriormente quedando de la siguiente manera:

En la siguiente tabla se describe dicha estructura:

| Esquemas    | Tablas            |
|-------------|-------------------|
| dimensiones | dim_temporal_anno |
| dimensiones | dim_provincia     |

|                     |                                         |
|---------------------|-----------------------------------------|
| dimensiones         | dim_temporal_mes                        |
| mart_medio_ambiente | dim_actividades_consumidoras            |
| mart_medio_ambiente | dim_actividades_del_uso_gases           |
| mart_medio_ambiente | dim_acueducto_alcantarillado            |
| mart_medio_ambiente | dim_agotadores_capa_ozono               |
| mart_medio_ambiente | dim_agua_potable_saneamiento            |
| mart_medio_ambiente | dim_areas_protegidas                    |
| mart_medio_ambiente | dim_categoria                           |
| mart_medio_ambiente | dim_categoria_huracanes                 |
| mart_medio_ambiente | dim_categoria_manejo_areas_protegidas   |
| mart_medio_ambiente | dim_clasificacion_agroproductica_suelos |
| mart_medio_ambiente | dim_clasificacion_genetica_suelos       |
| mart_medio_ambiente | dim_cuencas_hidrograficas               |
| mart_medio_ambiente | dim_distritos_geograficos               |
| mart_medio_ambiente | dim_especies_biota_cubana               |
| mart_medio_ambiente | dim_estaciones_meteorologicas           |
| mart_medio_ambiente | dim_extraccion_agua_dulce               |
| mart_medio_ambiente | dim_factores_limitantes_edaficos        |
| mart_medio_ambiente | dim_fauna_animal                        |
| mart_medio_ambiente | dim_gases_efecto_invernadero            |
| mart_medio_ambiente | dim_generadores_energia                 |
| mart_medio_ambiente | dim_grupos_taxonomicos                  |
| mart_medio_ambiente | dim_habitat_especies                    |
| mart_medio_ambiente | dim_huracanes                           |
| mart_medio_ambiente | dim_instituciones                       |
| mart_medio_ambiente | dim_intensidades_frente_frio            |
| mart_medio_ambiente | dim_localidades_cubanas                 |
| mart_medio_ambiente | dim_periodo                             |

|                     |                                          |
|---------------------|------------------------------------------|
| mart_medio_ambiente | dim_productos_produccion_materias_primas |
| mart_medio_ambiente | dim_regiones                             |
| mart_medio_ambiente | dim_significacion_areas_protegidas       |
| mart_medio_ambiente | dim_superficie                           |
| mart_medio_ambiente | dim_zona_residencial                     |
| mart_medio_ambiente | hech_acidez_lluvia                       |
| mart_medio_ambiente | hech_afectaciones_huracanes              |
| mart_medio_ambiente | hech_afectaciones_incendios_forestales   |
| mart_medio_ambiente | hech_agroproductiva_suelos               |
| mart_medio_ambiente | hech_agua_dulce_renovable                |
| mart_medio_ambiente | hech_agua_potable_saneamiento            |
| mart_medio_ambiente | hech_areas_protegidas                    |
| mart_medio_ambiente | hech_areas_protegidas_categoria_manejo   |
| mart_medio_ambiente | hech_areas_protegidas_internacional      |
| mart_medio_ambiente | hech_biota_cubana                        |
| mart_medio_ambiente | hech_capa_ozono                          |
| mart_medio_ambiente | hech_capa_ozono_actividades              |
| mart_medio_ambiente | hech_capa_ozono_sustancias               |
| mart_medio_ambiente | hech_carga_contaminante_cuencas          |
| mart_medio_ambiente | hech_carga_contaminante_provincias       |
| mart_medio_ambiente | hech_clima                               |
| mart_medio_ambiente | hech_cobertura                           |
| mart_medio_ambiente | hech_concentracion_gases                 |
| mart_medio_ambiente | hech_endemismo_biota                     |
| mart_medio_ambiente | hech_endemismo_vegetal                   |
| mart_medio_ambiente | hech_energia                             |
| mart_medio_ambiente | hech_extraccion_agua                     |
| mart_medio_ambiente | hech_fauna                               |

|                     |                                               |
|---------------------|-----------------------------------------------|
| mart_medio_ambiente | hech_flora                                    |
| mart_medio_ambiente | hech_flora_fauna_proteccion_exsitu            |
| mart_medio_ambiente | hech_frentes_frios_meses                      |
| mart_medio_ambiente | hech_frentes_frios_regiones                   |
| mart_medio_ambiente | hech_gases                                    |
| mart_medio_ambiente | hech_genetica_suelos                          |
| mart_medio_ambiente | hech_huracanes_meses                          |
| mart_medio_ambiente | hech_huracanes_regiones                       |
| mart_medio_ambiente | hech_lluvia                                   |
| mart_medio_ambiente | hech_principales_factores_limitantes_edaficos |
| mart_medio_ambiente | hech_produccion_materias_primas               |
| mart_medio_ambiente | hech_sistema_acueducto_alcantarillado         |
| mart_medio_ambiente | hech_superficie_cubierta_bosques              |
| mart_medio_ambiente | hech_temperatura                              |
| mart_medio_ambiente | hech_terremotos                               |
| mart_medio_ambiente | hech_tratamiento_recoleccion_desechos         |
| mart_medio_ambiente | hech_volumen_desecho                          |

**Tabla 2: Esquema del Mercado de datos.**

**Usuarios y privilegios.**

PostgreSQL ofrece la posibilidad de agrupar a los usuarios, según las necesidades de permisos y accesos que necesitará cada rol para realizar su función como trabajador del sistema. Los usuarios y roles definidos en la base de datos contribuyen a garantizar la seguridad de la misma.

A continuación se describen los roles establecidos:

- **Administrador de ETL:** Su función es realizar los procesos ETL en la interacción con la BD.

**Privilegios:**

Los privilegios que se les asigna a los usuarios del sistema son basados en el rol que desempeñan, en caso del Administrador de ETL, por el rol que este usuario desempeña, le son asignados los privilegios de *Select*, *Insert*, *Update*, *Delete*, *Refresh* y *Trigger* de los datos almacenados en el almacén.

### 3.2 Implementación del subsistema de integración

#### 3.2.1 Arquitectura del subsistema de integración

La arquitectura del proceso de desarrollo de un software, es la representación de la estructura del sistema al nivel más alto y consiste en un grupo de patrones que sirven de guía para la elaboración de un software. No se recomienda comenzar la construcción sin haber definido la arquitectura. El diseño de arquitectura para el mercado de datos Medio ambiente es el siguiente:

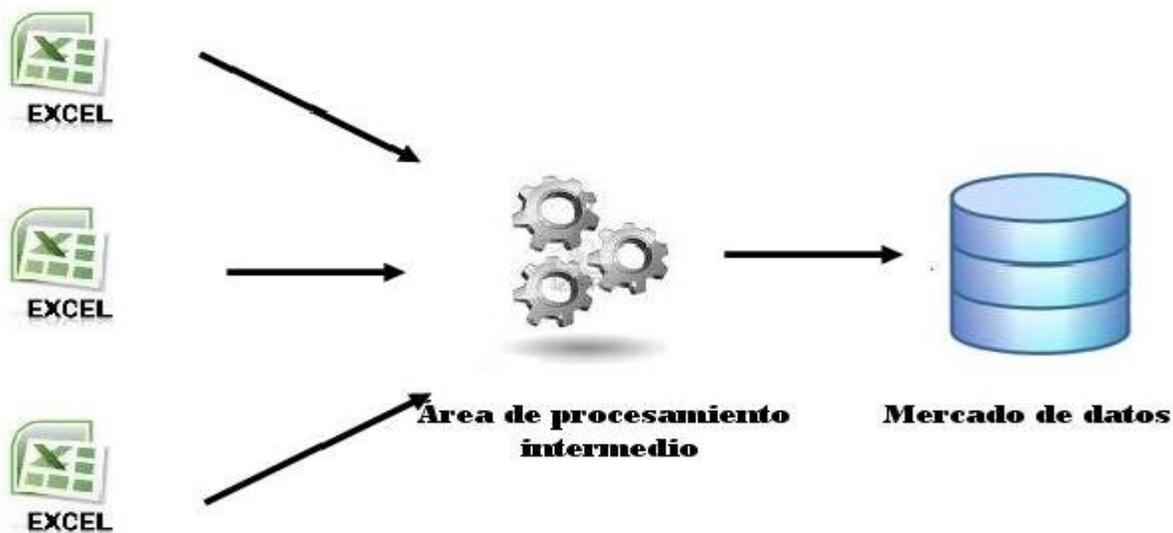


Figura 6: Arquitectura de la Integración.

**Fuente de Datos:** Son archivos de extensión .xls que contienen la información.

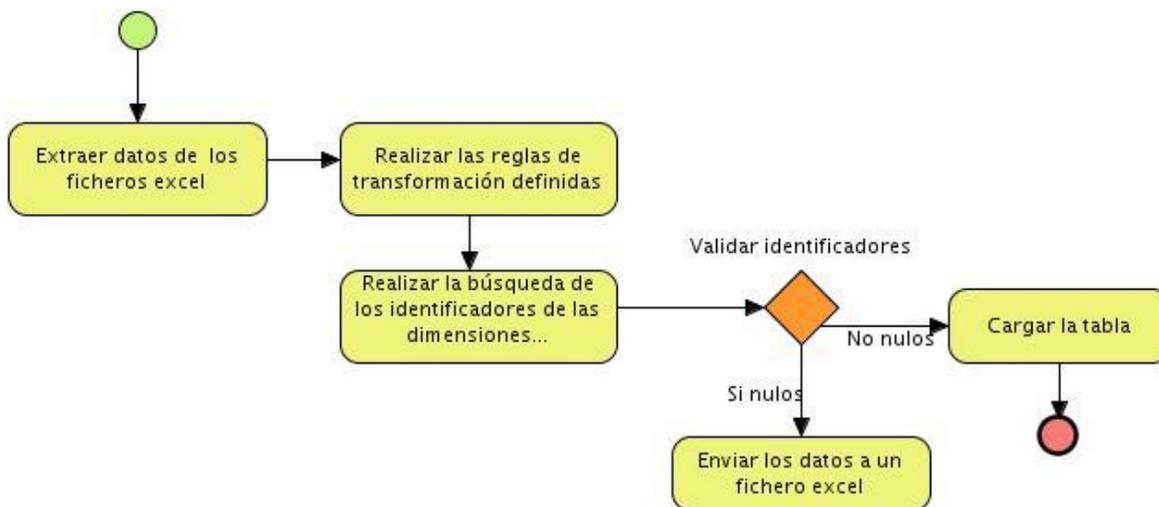
**Área de procesamiento intermedio:** Se almacenan temporalmente los datos después de haber realizado la extracción, transformación y carga de la información recibida de la fuente. Una vez extraídos los datos

de la fuente, se procede a transformarlos y limpiarlos. Para ello se aplican las reglas del negocio anteriormente seleccionadas y se validan otras nuevas que se pueden detectar. Entre los errores más comunes se encuentran: valores no significativos, nombres de provincias y estaciones de monitoreo escritos de diferentes formas.

**Mercado de Datos:** Destino donde son cargados los datos para su futuro análisis.

### Diseño de transformaciones:

Un diseño de transformaciones describe a grandes rasgos los pasos esenciales del proceso de ETL que son comunes para la carga de las tablas de hecho de un mercado de datos. Así se ejemplifica en la Figura 7.



**Figura 7: Arquitectura de la Integración.**

La figura 8 muestra el proceso efectuado para cargar la tabla de dimensiones **dim\_grupos\_taxonomicos**. En esta como en las demás transformaciones para la carga de dimensiones se realizan las validaciones necesarias para aquellos campos que pueden aparecer nulos y que cargarlos indebidamente provocaría errores en la información que se maneja. Una vez ejecutada esta transformación quedarán guardados en esta tabla los datos relacionados a cada grupo taxonómico.

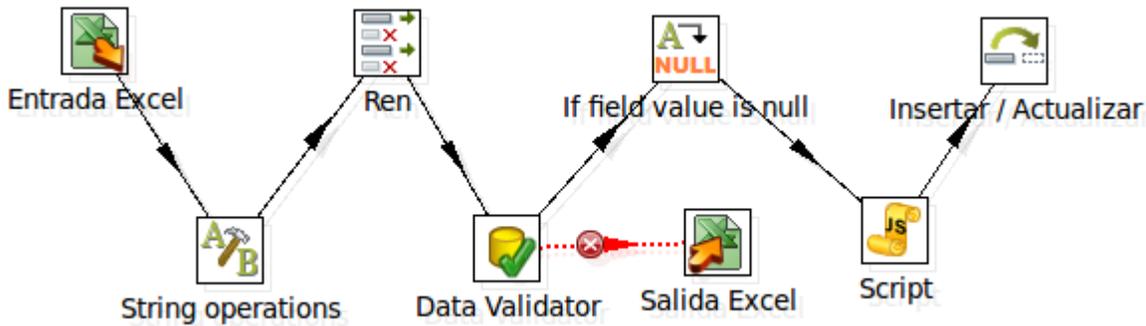


Figura 8: Ejemplo de transformación de la dimensión grupos taxonómicos.

En la figura 9 se muestra como fue realizado el proceso para cargar la tabla de hecho **hech\_lluvia**.

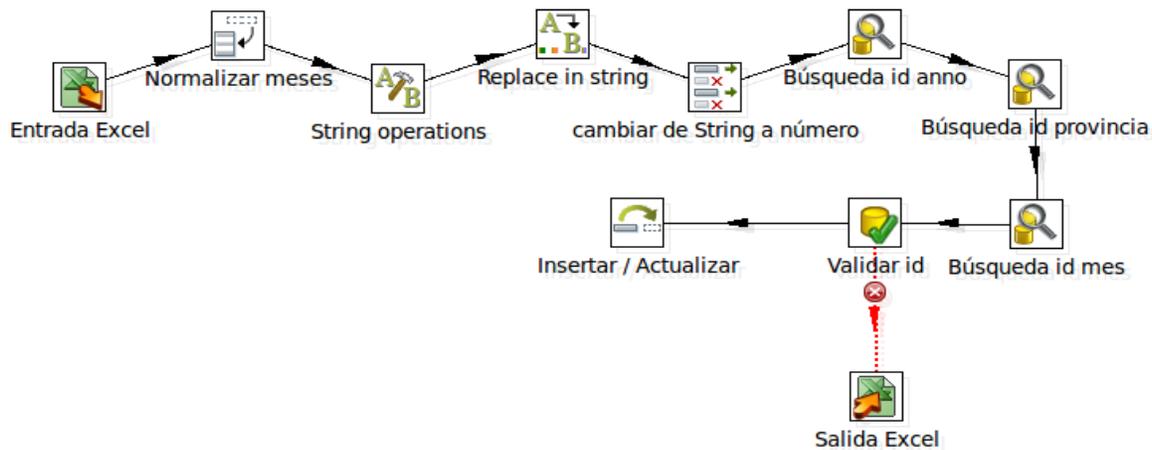
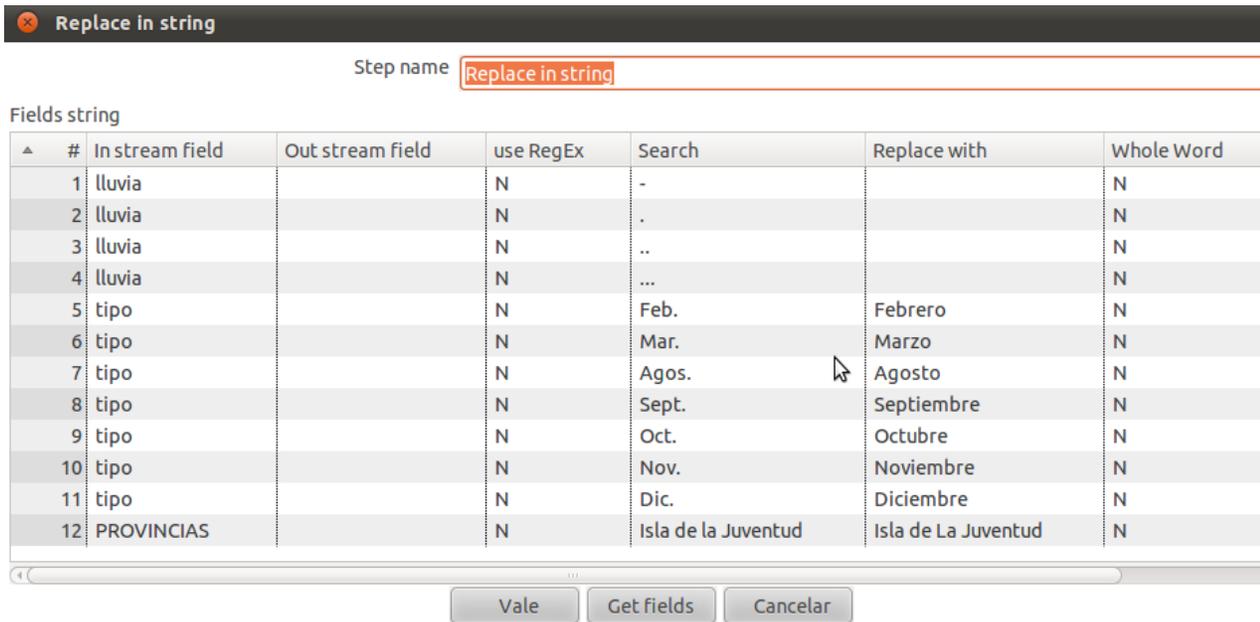


Figura 9: Ejemplos de transformación de hecho lluvia media.

Para validar las reglas del negocio definidas se utilizó el componente Replace in String. En la figura 10 se puede observar el uso que se le da al mismo para las transformaciones propias de la tabla de hecho representada en la figura 9.



**Figura 10: Validación de las reglas de transformación.**

**Trabajo**

El trabajo, también conocido como Job es una transformación resultante de unir las transformaciones realizadas para las tablas de dimensiones y hechos para ser ejecutados consecutivamente de acuerdo con el orden de prioridad que se establece, con el fin de realizar la carga de los datos en el mercado. El paso inicial es asegurarse de que la conexión al mercado este en perfecto estado. La ejecución de los restantes pasos da como resultado la población del mismo.

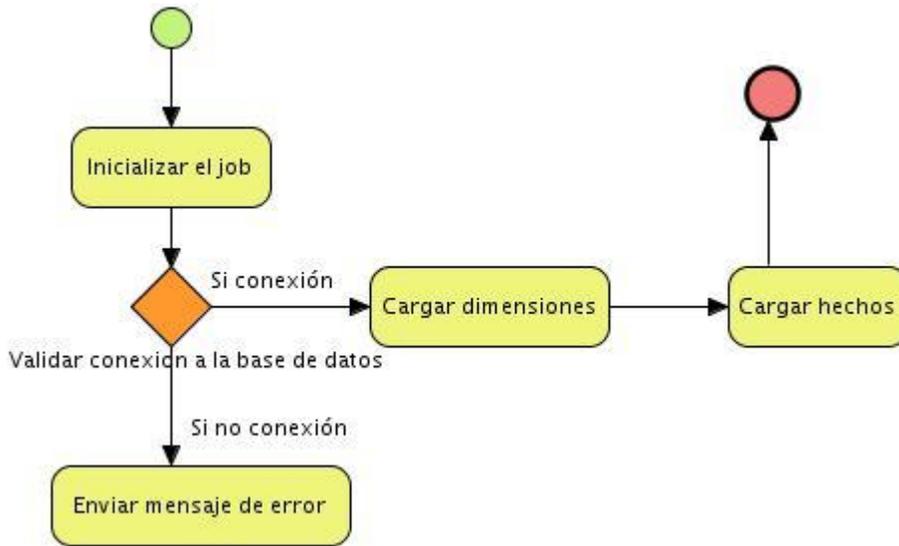


Figura 11: Diseño del trabajo.

### 3.3 Implementación de subsistema de visualización

#### 3.3.1 Cubos OLAP

Para la implementación del subsistema de visualización, es necesaria la creación de los cubos multidimensionales, en los cuales se definen las dimensiones, los niveles de jerarquía de las dimensiones y las medidas. Esto se realiza a través de la herramienta Pentaho Schema Workbench, donde la misma genera un fichero XML, el cual contendrá la estructura de todos los cubos OLAP.

En la realización del diseño de los cubos, se modelaron en el esquema Medio\_ambiente 40 cubos OLAP, donde cada uno está formado por sus dimensiones y medidas que los caracterizan. La siguiente imagen muestra el diseño de los cubos.

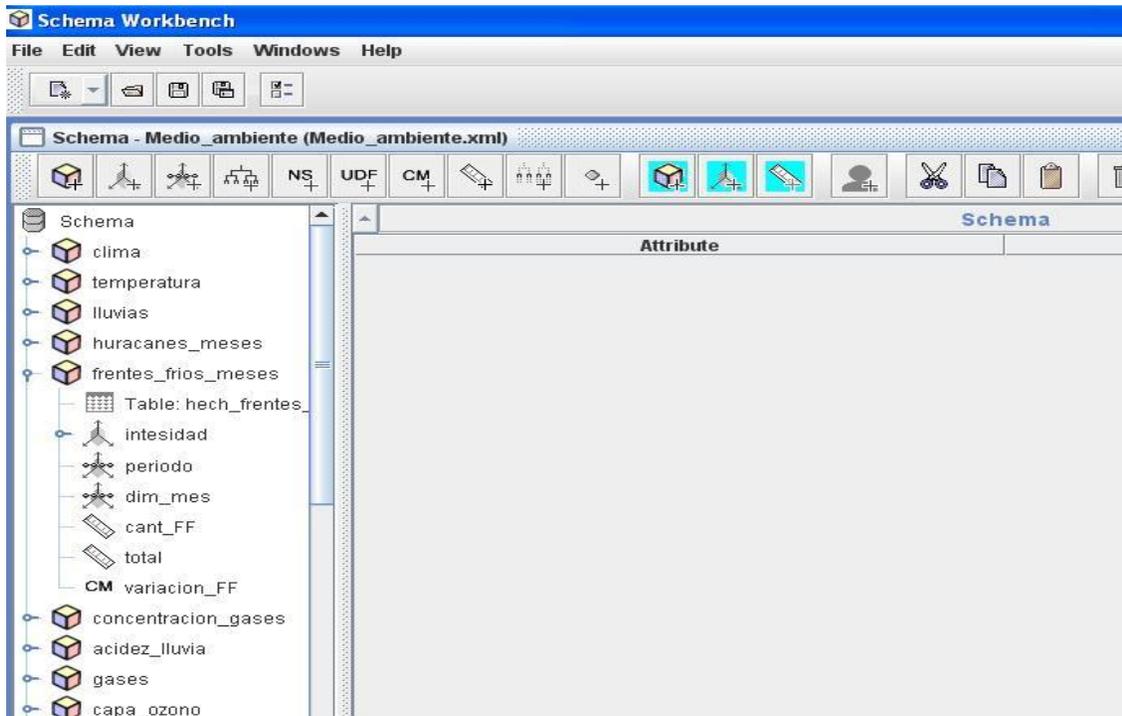
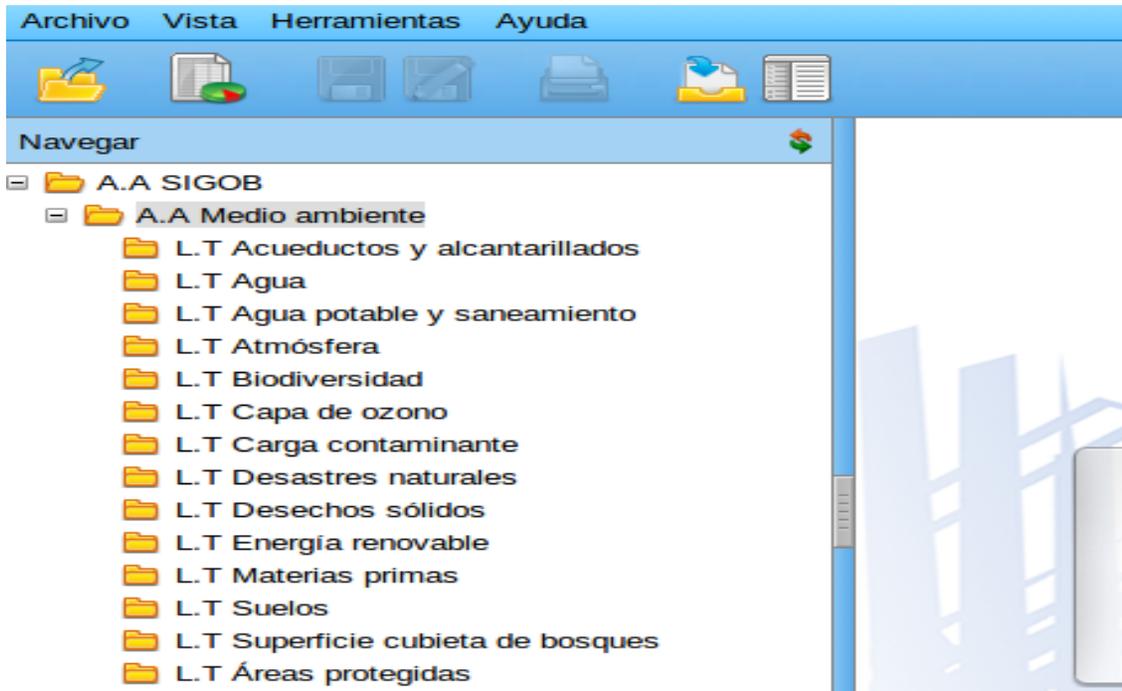


Figura 12: Diseño de los cubos.

### 3.3.1 Navegación de la capa de visualización

El mapa de navegación es la representación gráfica en que está organizada la información. El mercado de datos Medio ambiente está compuesto por 1 Área Análisis General (A.A.G), 1 Área de Análisis (A.A) y 14 Libros de Trabajo (L.T). A continuación se detalla la estructura en la cual será presentado en la capa de visualización.



**Figura 13: Diseño de los libros de trabajo.**

Descripción del Área de Análisis General.

**A.A.G SIGOB:** Agrupa toda la información de los mercados de datos de la ONE, formando el almacén SIGOB.

Descripción del Área de Análisis.

**A.A Medio ambiente:** Agrupa toda la información referente a los indicadores ambientales, atmósfera, temperatura, clima, áreas protegidas entre otros. Además, contiene reportes estadísticos que muestran el comportamiento de los diferentes indicadores.

Descripción de uno de los libros de Trabajo.

**L.T Energía renovable:** Libro de trabajo contenido dentro del área de análisis Medio ambiente, contiene 2 reportes que permiten realizar un análisis de datos correspondiente a la energía renovable en Cuba.

Descripción de los Reportes del L.T Energía renovable

**TS1**–Obtener la cantidad de dispositivos generadores de energía y biomasa empleada como combustible, por dispositivos, biomasa y anualmente.

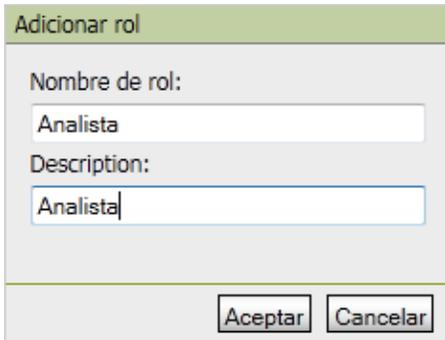
**TS2**–Obtener la cantidad de oferta renovable por dispositivos y biomasa, anualmente.

El mercado de datos Medio ambiente está compuesto por 51 reportes en total, los cuales fueron distribuidos dentro de los 14 libros de trabajo. Para la realización de estos se utilizaron como herramientas: el Mondrian para la visualización de los reportes, utilizando el Tomcat como servidor web y el Pentaho BI Server, donde queda estructurado finalmente el mapa de navegación junto con todos los reportes.

### 3.3.3 Configurar la seguridad de los usuarios

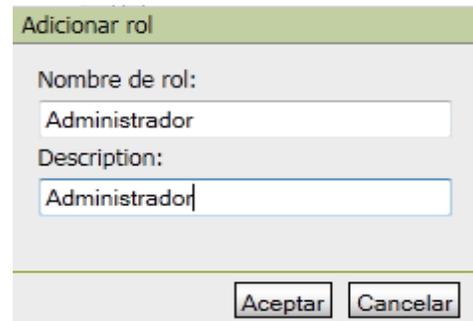
Durante la implementación del subsistema de visualización del mercado de datos Medio ambiente, se crearon 2 usuarios y roles los cuales tienen diferentes permisos de acceso a la información, proporcionando una mayor seguridad al sistema.

- ✓ El rol administrador tiene todos los permisos de la aplicación y posee el usuario administrador del sistema.
- ✓ El rol analista tiene permiso de solo lectura y posee el usuario analista del sistema.



The screenshot shows a dialog box titled "Adicionar rol" with a light green header. It contains two text input fields: "Nombre de rol:" with the value "Analista" and "Description:" with the value "Analista". At the bottom right, there are two buttons: "Aceptar" and "Cancelar".

**Figura 14: Rol Analista**



The screenshot shows a dialog box titled "Adicionar rol" with a light green header. It contains two text input fields: "Nombre de rol:" with the value "Administrador" and "Description:" with the value "Administrador". At the bottom right, there are two buttons: "Aceptar" and "Cancelar".

**Figura 15: Rol Administrador**

### Conclusiones del capítulo

Al concluir este capítulo se culminó la implementación del mercado de datos. Se tuvo en cuenta para la realización de la misma las reglas del negocio definidas en el capítulo anterior. Se estableció la arquitectura de la información y se realizó el diseño de las transformaciones. Se realizaron los cubos OLAP para cada una de las tablas de hechos del mercado de datos, así como los reportes asociados a los mismos satisfaciendo las necesidades de los usuarios. Se obtuvo como resultado final del mismo un mercado de datos poblado y listo para pasar al proceso de validación del producto.

**CAPÍTULO 4: VALIDACIÓN**

**Introducción**

Este capítulo tiene como objetivo, fundamentar el conjunto de validaciones realizadas al MD para demostrar la calidad de su implementación. Un producto tiene calidad si satisface las necesidades de quienes lo emplearán. Para ello se aplica la lista de chequeo y se realizaron los casos de pruebas y las pruebas de aceptación del cliente.

**4.1 Diseño de casos de pruebas**

Las pruebas del software constituyen un elemento fundamental para garantizar la calidad de cualquier producto, de esta manera, se puede verificar si nuestro sistema cumple con lo esperado por el cliente. Existen diversas pruebas aplicables al mercado de datos como lo son, listas de chequeo y las pruebas de interfaz. En el mercado de datos del área medio ambiente se diseñaron los casos de pruebas de interfaz, a continuación se muestra un ejemplo para el caso de uso de información “Analizar información de la capa de ozono”

| Escenario                                                                                   | Descripción                                                                                                | Perfiles de análisis                                                   | Indicadores a medir                                | Respuesta del sistema                                                                                                                                     | Flujo central                                                                                                                                                     |
|---------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------|----------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| EC 1.1:<br>Consumo de sustancias agotadoras de la capa de ozono                             | Se obtiene la cantidad de consumo por DPA, agotadores de la capa de ozono, actividades consumidoras y años | DPA<br>Agotadores de la capa de ozono<br>actividad consumidora<br>Años | -cantidad de consumo                               | El sistema muestra todas las variables disponibles para los análisis, ubicados en las filas y las columnas que pueden ser visualizadas para cada reporte. | Se abre la aplicación.<br>Se autentifica.<br>Se entra al sistema.<br>Se selecciona la opción Nueva Vista de Análisis.<br>Se selecciona el esquema Medio_ambiente. |
| EC 1.2:<br>Consumo de sustancias agotadoras de la capa de ozono de acuerdo con el potencial | Se obtiene la cantidad y variación del consumo por DPA, agotadores de la capa de ozono y años              | DPA<br>Agotadores de la capa de ozono<br>Años                          | -cantidad de consumo<br><br>-variación del consumo |                                                                                                                                                           | Se selecciona el cubo <b>Capa de ozono, Capa de ozono por sustancias o Capa de ozono por</b>                                                                      |

|                                                                                        |                                                                                         |                                |                                                    |  |                                                                                                                                                                                                                                       |
|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|--------------------------------|----------------------------------------------------|--|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| de agotamiento                                                                         |                                                                                         |                                |                                                    |  | <b>actividades.</b><br>Se selecciona la opción Abrir navegador OLAP en la parte superior izquierda.<br><br>Se verifica que estén disponibles en las filas y las columnas de todos los perfiles de análisis y los indicadores a medir. |
| EC 1.3: Consumo de sustancias agotadoras de la capa de ozono por actividad consumidora | Se obtiene la cantidad y variación del consumo por DPA, actividades consumidoras y años | DPA actividad consumidora Años | -cantidad de consumo<br><br>-variación del consumo |  |                                                                                                                                                                                                                                       |

**Tabla 3: Diseño de caso de prueba "Analizar información de la capa de ozono".**

**4.2 Lista de chequeo**

La lista de chequeo contiene diferentes indicadores a evaluar los cuales se encuentran distribuidos en tres secciones fundamentales: Estructura del documento, Indicadores definidos en el desarrollo y Semántica del documento

| <b>Estructura del documento</b>               |                                                                                                      |             |             |                                        |
|-----------------------------------------------|------------------------------------------------------------------------------------------------------|-------------|-------------|----------------------------------------|
| <b>Peso</b>                                   | <b>Indicadores a evaluar</b>                                                                         | <b>Eval</b> | <b>(NP)</b> | <b>Cantidad de elementos afectados</b> |
| crítico                                       | 1. ¿Los Artefactos presentan las secciones obligatorias establecidas para el expediente de proyecto? | 0           |             | 0                                      |
| <b>Indicadores definidos en el desarrollo</b> |                                                                                                      |             |             |                                        |

| <b>Peso</b> | <b>Indicadores a evaluar</b>                                                                                                             | <b>Eval</b> | <b>(NP)</b> | <b>Cantidad de elementos afectados</b> |
|-------------|------------------------------------------------------------------------------------------------------------------------------------------|-------------|-------------|----------------------------------------|
|             | 1. ¿El lenguaje utilizado contiene sentencias que son expresables mediante una sintaxis bien definida?                                   | 0           |             | 0                                      |
| crítico     | 2. ¿Los reportes se configuran a través de la interfaz del sistema?                                                                      | 0           |             | 0                                      |
|             | 3. ¿La interfaz está orientada a facilitar el uso de las funciones del sistema por parte de los usuarios?                                | 0           |             | 0                                      |
| crítico     | 4. ¿No existen restricciones para construir cubos OLAP con dimensiones y niveles de agregación ilimitados?                               | 0           |             | 0                                      |
| crítico     | 5. ¿Los reportes pueden ser modificados dependiendo de las necesidades de los usuarios?                                                  | 0           |             | 0                                      |
|             | 6. ¿Una vez que el usuario solicita la información el sistema responde con rapidez?                                                      | 0           |             | 0                                      |
|             | 7. ¿El sistema refleja cualquier lógica del negocio para poder responder a preguntas específicas?                                        | 0           |             | 0                                      |
| crítico     | 8. ¿En dependencia del rol asignado al usuario, el sistema es capaz de garantizar la confidencialidad y seguridad de acceso a los datos? | 0           |             | 0                                      |

|                                |                                                                                                                      |             |                                             |
|--------------------------------|----------------------------------------------------------------------------------------------------------------------|-------------|---------------------------------------------|
|                                | 9. ¿La toma de decisiones en la Institución se apoya en los datos e información derivados del proceso de análisis?   | 0           | 0                                           |
| crítico                        | 10. ¿Los cambios en los datos se reflejan automáticamente en los reportes de forma instantánea?                      | 0           | 0                                           |
| <b>Semántica del documento</b> |                                                                                                                      |             |                                             |
| <b>Peso</b>                    | <b>Indicadores a evaluar</b>                                                                                         | <b>Eval</b> | <b>(NP) Cantidad de elementos afectados</b> |
| crítico                        | 1. ¿Se han identificado errores ortográficos en los entregables?                                                     | 0           | 0                                           |
| crítico                        | 2. ¿Se entiende claramente lo que se ha especificado en el documento?                                                | 0           | 0                                           |
|                                | 3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página? | 0           | 0                                           |

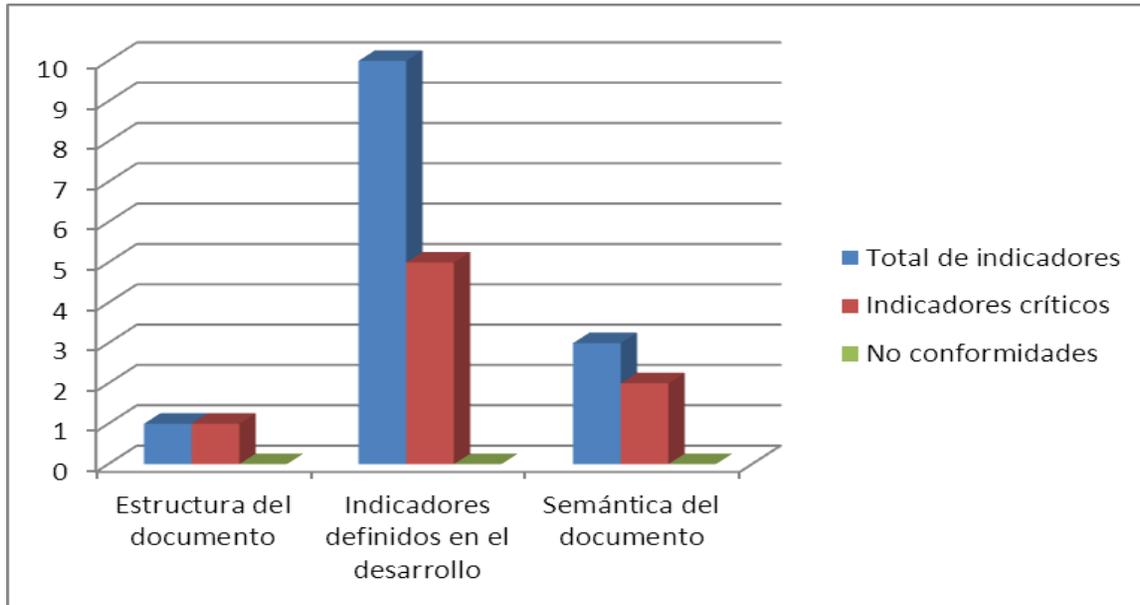
**Tabla 4: Lista de chequeo.**

**4.3 Pruebas de aceptación.**

Para aprobar la propuesta de solución se realizó un encuentro con el cliente: Elena Fernández García. En este se detectaron algunos errores en datos calculables y en la carga de valores de algunas medidas. Se obtuvo un total de 12 no conformidades las cuales fueron solucionadas y posteriormente aprobadas por la persona anteriormente mencionada.

#### 4.4 Resultados y discusión de las pruebas

En el siguiente gráfico de barras aparece el comportamiento de los indicadores definidos para la lista de chequeo elaborada. De forma general se identificaron 14 indicadores, de ellos 8 críticos y luego de aplicada la lista de chequeo, se generaron 0 no conformidades.



**Figura 16: Comportamiento de los indicadores por secciones.**

#### Conclusiones

Se aplicó la lista de chequeo elaborada sin que se obtuvieran no conformidades. Se realizaron 14 Casos de pruebas correspondiendo cada uno de ellos a un requisito de información. Se validó la propuesta con el cliente detectándose y corrigiéndose 12 no conformidades.

### CONCLUSIONES GENERALES

Después de haber desarrollado el mercado de datos Medio ambiente se llegan a las siguientes conclusiones:

- ✓ El estudio del estado del arte permitió un entendimiento de la implementación del mercado de datos.
- ✓ Las herramientas seleccionadas garantizaron de una manera correcta el buen funcionamiento del MD.
- ✓ Los requisitos definidos dan solución a las necesidades del cliente.
- ✓ Las estructuras dimensionales implementadas cumplen con las condiciones necesarias para integrar la solución.
- ✓ El proceso de integración de datos y la inteligencia de negocios fueron desarrollados satisfactoriamente.
- ✓ Se realizaron las pruebas pertinentes para la validación del producto, demostrando la calidad requerida.
- ✓ La validación de la solución demuestra que el sistema cumple con las expectativas del cliente.
- ✓ Se desarrolló un MD que contribuye a la toma de decisiones en el área Medio ambiente del Sistema de Información de Gobierno, dándole cumplimiento a las tareas de la investigación, objetivos específicos y generales.

### Recomendaciones

- Mantener actualizado el Mercado de Datos ante los posibles cambios que ocurran para/con los datos almacenados.
- Integrar el mercado de datos con el almacén de datos correspondiente.

### REFERENCIA BIBLIOGRÁFICA

1. **GM, David. 2009** . Eureka Digital. [En línea] 9 de Febrero de 2009 . [Citado el: 5 de Diciembre de 2010.] **SINNEXUS. 2008.** SINNEXUS. *Business Intelligence+Informatica estrinta*. [En línea] 2008. [Citado el: 29 de Noviembre de 2010.] [http://www.sinnexus.com/business\\_intelligence/index.aspx](http://www.sinnexus.com/business_intelligence/index.aspx).
2. **Caramazana, Alberto. 2002.** using System.Makondo. *Doctorado en ingeniería del software*. [En línea] Septiembre de 2002. [Citado el: 1 de Diciembre de 2010.] <http://www.willydev.net/descargas/Articulos/General/bi.pdf>.
3. **Wolff, Carmen Gloria. 2002.** Departamento Ingenieria y ciencia de la computación. [En línea] 28 de Agosto de 2002. [Citado el: 29 de Noviembre de 2011.] <http://www.inf.udec.cl/~revista/ediciones/edicion4/modmulti.PDF>.
4. **Cabrera, María Evelia Casales. 2009.** Universidad autónoma de México. [En línea] 2009. [Citado el: 28 de noviembre de 2010.] <http://hp.fciencias.unam.mx/~alg/bd/dwh.pdf>.
5. **Dataprix. 2010.** Dataprix. *OLAP, MOLAP y ROLAP*. [En línea] 2010. [Citado el: 30 de Noviembre de 2010.] <http://www.dataprix.com/olap-rolap-molap>.
6. **Espinosa, Roberto. 2010.** El Rincón del BI. *Descubriendo el Business Intelligence*. [En línea] Kimball vs Inmon. Ampliación de conceptos del Modelado Dimensional., 19 de Abril de 2010. [Citado el: 25 de Noviembre de 2010.] <http://churriwifi.wordpress.com/2010/04/19/15-2-ampliacion-conceptos-del-modelado-dimencional/>.
7. **Ralph Kimball, Joe Caserta. 2004.** *The Data Warehouse ETL toolkit*. s.l. : Wiley Publishing, Inc., 2004.
8. **Información, Sistemas avanzados de recuperación de Infomación. 2010.** SISTEMAS AVANZADOS DE RECUPERACIÓN DE INFORMACIÓN. *INTRODUCCIÓN A LOS METADATOS*. [En línea] 2010. [Citado el: 30 de Noviembre de 2010.] <http://sistemasavanzadosderecuperaciondeinformacion.iespana.es/>.
9. **Inteligencia de negocio. 2010.** slideshare. *Inteligencia de negocio*. [En línea] 2010. [Citado el: 27 de Noviembre de 2010.] <http://www.slideshare.net/rosmelys/trabajo-business-intelligence>.
10. **Caceres, Ana Mercedes. 2006.** Galéon.com. *HispaVista*. [En línea] 21 de JULIO de 2006. [Citado el: 1 de Diciembre de 2010.] <http://www.galeon.com/rcruz0423/docs/case.pdf>.

11. **Summan. 2010.** Summan. *Manejo documental e infraestructura informatica*. [En línea] 2010. [Citado el: 30 de Noviembre de 2010.] <http://www.summan.com/index.php/productos/software/pentaho-.html>
12. **Sánchez, Leopoldo Zenaido Zepeda. 2008.** *Metodología para el Diseño*. Valencia : s.n., 2008.
13. **Torres, Ing.Lilliam Vega. 2008.** *LA INTELIGENCIA DE NEGOCIO.SU IMPLEMENTACIÓN MEDIANTE LA PLATAFORMA PENTAHO*. Ciudad de la Habana : s.n., 2008.
14. **Sherman Wood, JasperSoft. 2007.** pentaho. *Mondrian Documentation*. [En línea] Abril de 2007. [Citado el: 2 de Diciembre de 2011.] <http://mondrian.pentaho.com/documentation/workbench.php>.
15. **Esteban, Eloy A. 2010.** Programación en castellano. *Artículo*. [En línea] 2010. [Citado el: 3 de Diciembre de 2011.] [http://www.programacion.com/articulo/tomcat - introduccion\\_134#tomcat1](http://www.programacion.com/articulo/tomcat - introduccion_134#tomcat1).

## BIBLIOGRAFÍA

1. **Albuerne, Juan Luis Martínez. 2008.** Archivo Digital UPM. [En línea] 2008. [Citado el: 27 de noviembre de 2010.] [http://oa.upm.es/1024/1/PFC\\_JUAN\\_LUIS\\_MARTINEZ\\_ALBURNE.pdf](http://oa.upm.es/1024/1/PFC_JUAN_LUIS_MARTINEZ_ALBURNE.pdf).
2. **Cabrera, María Evelia Casales. 2009.** Universidad autónoma de México. [En línea] 2009. [Citado el: 28 de noviembre de 2010.] <http://hp.fciencias.unam.mx/~alg/bd/dwh.pdf>.
3. **Caceres, Ana Mercedes. 2006.** Galéon.com. *HispaVista*. [En línea] 21 de JULIO de 2006. [Citado el: 1 de Diciembre de 2010.] <http://www.galeon.com/rcruz0423/docs/case.pdf>.
4. **Caramazana, Alberto. 2002.** using System.Makondo. *Doctorado en ingeniería del software*. [En línea] Septiembre de 2002. [Citado el: 1 de Diciembre de 2010.] <http://www.willydev.net/descargas/Articulos/General/bi.pdf>.
5. **Casters, Matt. 2010.** Softpedia. [En línea] 08 de Abril de 2010. [Citado el: 04 de Diciembre de 2010.] <http://www.softpedia.com/es/programa-Kettle-80094.html>.
6. **Catrin, Franco. 2009.** Dosis diarias de tecnología en español. *Fayer Wayer*. [En línea] 2009. [Citado el: 27 de Noviembre de 2010.] <http://www.fayerwayer.com/2009/07/postgresql-84-mantiene-distancia-con-sus-rivales/>.
7. **Dataprix. 2010.** Dataprix. *OLAP, MOLAP y ROLAP*. [En línea] 2010. [Citado el: 30 de Noviembre de 2010.] <http://www.dataprix.com/olap-rolap-molap>.
8. **Denzer, Patricio. 2002.** Proyecto Personal - Programación de Sistemas. *PostgreSQL*. [En línea] 23 de Octubre de 2002. [Citado el: 4 de Diciembre de 2010.] <http://profesores.elo.utfsm.cl/~aqv/elo330/2s02/projects/denzer/informe.pdf>.
9. **Descarga de software. 2007.** Descarga de software. [En línea] 2007 de Marzo de 2007. [Citado el: 3 de Diciembre de 2010.] [http://www.freedownloadmanager.org/es/downloads/Paradigma\\_Visual\\_para\\_UML\\_%28M%C3%8D%29\\_14720\\_p/](http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML_%28M%C3%8D%29_14720_p/).
10. **Díaz, Josep Curto. 2009.** Beyenetwork. *Actualización menor de DataCleaner*. [En línea] 2 de Abril de 2009. [Citado el: 3 de Diciembre de 2010.] <http://www.beyenetwork.es/blogs/curtodiaz/archives/data-cleansing/>.
11. **Díaz, Josep Curto. 2007.** Information Management. *Reflexiones sobre las tecnologías de la información*. [En línea] 7 de Octubre de 2007. [Citado el: 29 de Noviembre de 2010.]

- <http://informationmanagement.wordpress.com/2007/10/07/data-warehousing-data-warehouse-y-datamart/>.
12. **Equipo Danysoft. 2006.** Danysoft. *Modelado de bases de datos*. [En línea] 25 de Agosto de 2006. [Citado el: 02 de Diciembre de 2010.] <http://www.danysoft.com/free/model2.pdf>.
  13. **Espinosa, Roberto. 2010.** El Rincón del BI. *Descubriendo el Business Intelligence*. [En línea] Kimball vs Inmon. Ampliación de conceptos del Modelado Dimensional., 19 de Abril de 2010. [Citado el: 25 de Noviembre de 2010.] <http://churriwifi.wordpress.com/2010/04/19/15-2-ampliacion-conceptos-del-modelado-dimencional/>.
  14. **Esteban, Eloy A. 2010.** Programación en castellano. *Artículo*. [En línea] 2010. [Citado el: 3 de Diciembre de 2011.] <http://www.programacion.com/articulo/tomcat - introduccion 134#tomcat1>.
  15. **GM, David. 2009 .** Eureka Digital. [En línea] 9 de Febrero de 2009 . [Citado el: 5 de Diciembre de 2010.] <http://eurecadigital.wordpress.com/2009/02/09/replicar-base-de-datos-sql-server-2000-en-2-servidores-distintos/>.
  16. **Gómez, Isabel Criado y López, Paloma Sánchez. 2010.** NorSistemas. [En línea] Empresa Consultora Externa, 2010. [Citado el: 29 de Noviembre de 2010.] <http://www.csae.map.es/csi/silice>.
  17. **González, Carlos D. 2010.** Curso PostgreSQL, SQL avanzado y PHP. *Diseño natural, usable, seguro, accesible*. [En línea] Noviembre de 2010. [Citado el: 28 de Noviembre de 2010.] <http://www.usabilidadweb.com.ar/postgre.php>.
  18. **Ibarra, Antonio Aliaga. 2008.** I.E.S. San Vicente. *I.E.S. San Vicente*. [En línea] 21 de Enero de 2008. [Citado el: 3 de Diciembre de 2011.] <http://www.iessanvicente.com/colaboraciones/postgreSQL.pdf>.
  19. **Información, Sistemas avanzados de recuperación de Infomación. 2010.** SISTEMAS AVANZADOS DE RECUPERACIÓN DE INFORMACIÓN. *INTRODUCCIÓN A LOS METADATOS*. [En línea] 2010. [Citado el: 30 de Noviembre de 2010.] <http://sistemasavanzadosderecuperaciondeinformacion.iespana.es/>.
  20. **Information Builders. 2010.** Information Builders. *Integración empresarial*. [En línea] 2010. [Citado el: 29 de Diciembre de 2010.] <http://www.informationbuilders.es/pdfs/A4 iWay Brochure ES.pdf>.
  21. **Inteligencia de negocio. 2010.** slideshare. *Inteligencia de negocio*. [En línea] 2010. [Citado el: 27 de Noviembre de 2010.] <http://www.slideshare.net/rosmelys/trabajo-business-intelligence>.

22. **León, Eduardo. 2007.** Blog de Eduardo León. *Visual Paradigm, una herramienta de lo más útil.* [En línea] 2 de Abril de 2007. [Citado el: 3 de Diciembre de 2010.] <http://slion2000.blogspot.com/2007/04/visual-paradigm-una-herramienta-de-lo.html>.
23. **Portada sobre la plataforma Pentaho Open Source. 2010.** Portada sobre la plataforma Pentaho Open Source . *Business Intelligence.* [En línea] 2010. [Citado el: 2 de Diciembre de 2010.] <http://pentaho.almacen-datos.com/mondrian.html>.
24. **ralfm. 2007.** El rincón de Linux. *Introducción a PostgreSQL - Instalación e inicialización.* [En línea] Rafael Martinez, 08 de Junio de 2007. [Citado el: 29 de Noviembre de 2010.] <http://www.linux-es.org/node/536>.
25. **Ralph Kimball, Joe Caserta. 2004.** *The Data Warehouse ETL toolkit.* s.l. : Wiley Publishing, Inc., 2004.
26. **Sánchez, Leopoldo Zenaido Zepeda. 2008.** *Metodología para el Diseño.* Valencia : s.n., 2008.
27. **Sherman Wood, JasperSoft. 2007.** pentaho. *Mondrian Documentation.* [En línea] Abril de 2007. [Citado el: 2 de Diciembre de 2011.] <http://mondrian.pentaho.com/documentation/workbench.php>.
28. **SINNEXUS. 2008.** SINNEXUS. *Business Intelligence+Informatica estrinta.* [En línea] 2008. [Citado el: 29 de Noviembre de 2010.] [http://www.sinnexus.com/business\\_intelligence/index.aspx](http://www.sinnexus.com/business_intelligence/index.aspx).
29. **Soluciones AG. 2009.** Soluciones AG. *Soluciones para alta gerencia AG, SA.* [En línea] 2009. [Citado el: 25 de Noviembre de 2010.] [http://www.soluciones-ag.com/pdf\\_productos/Spanish\\_ER-Studio\\_Datasheet\\_2009.pdf](http://www.soluciones-ag.com/pdf_productos/Spanish_ER-Studio_Datasheet_2009.pdf).
30. **Torres, Ing.Lilliam Vega. 2008.** *LA INTELIGENCIA DE NEGOCIO.SU IMPLEMENTACIÓN MEDIANTE LA PLATAFORMA PENTAHO.* Ciudad de la Habana : s.n., 2008.
31. **Summan. 2010.** Summan. *Manejo documental e infraestructura informatica.* [En línea] 2010. [Citado el: 30 de Noviembre de 2010.] <http://www.summan.com/index.php/productos/software/pentaho-.html>
32. **Web, Elección del servidor de aplicaciones. 2003.** Información técnica y de gestión para IBM. [En línea] Febrero de 2003. [Citado el: 3 de Diciembre de 2011.] <http://www.help400.es/asp/scripts/nwart.asp?Num=131&Pag=10&Tip=T>.
33. **Wolff, Carmen Gloria. 2002.** Departamento Ingenieria y ciencia de la computación. [En línea] 28 de Agosto de 2002. [Citado el: 29 de Noviembre de 2011.] <http://www.inf.udec.cl/~revista/ediciones/edicion4/modmulti.PDF>.

### GLOSARIO DE TÉRMINOS

- **ONE:** Oficina Nacional de Estadísticas.
- **AD:** Almacenes de Datos.
- **MD:** Mercado de datos.
- **DATEC:** Centro de Tecnologías de datos.
- **BI:** Inteligencia de Negocio.
- **UCI:** Universidad de Ciencias Informáticas.
- **SGBD:** Sistema Gestor de base datos.
- **ROLAP:** Procesamiento Analítico Relacional en Línea.
- **MOLAP:** Procesamiento Analítico Multidimensional en Línea.
- **HOLAP:** Procesamiento Analítico Híbrido en Línea.
- **Matriz BUS:** Es la relación existente de cada uno de los hechos con sus respectivas dimensiones en una matriz.
- **RI:** Requisito de información.
- **RF:** Requisito funcional.
- **RNF:** Requisito no funcional.
- **DPA:** División política administrativa.
- **CASE:** Computer Aided Software Engineering(ingeniería de software orientada por computadoras)