

Universidad de las Ciencias Informáticas

Facultad 6



Título: Mercado de datos Series históricas de agricultura, ganadería, silvicultura y pesca para el Sistema de Información de Gobierno

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

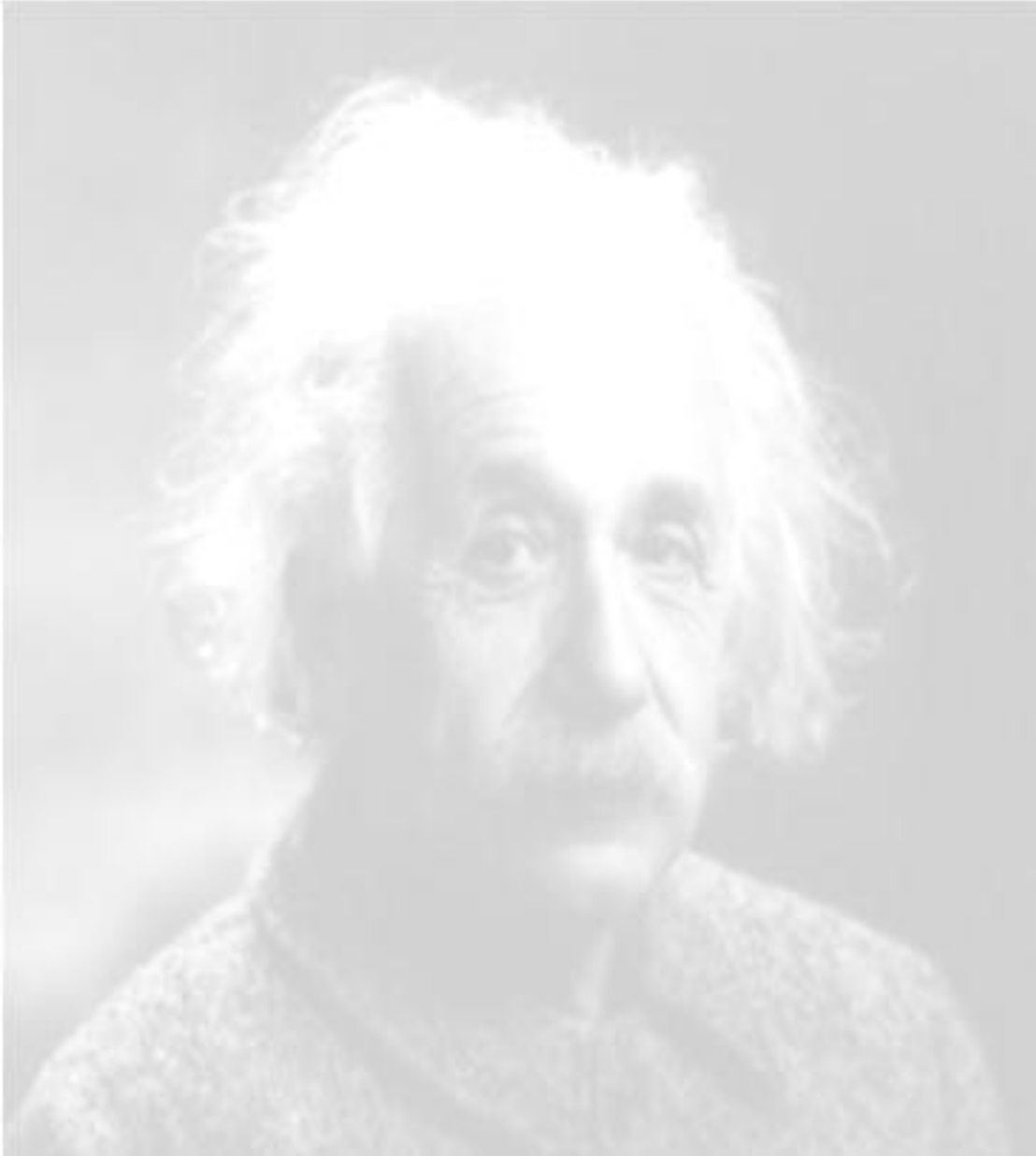
Autores: Liniuska Cardero Dieguez

Livan López González

Tutora: Ing. Themis Patricia Díaz Morales

La Habana, Junio de 2012

“Año 54 de la Revolución”



¿Por qué esta magnífica tecnología científica, que ahorra trabajo y nos hace la vida más fácil, nos aporta tan poca felicidad? La respuesta es ésta, simplemente: porque aún no hemos aprendido a usarla con tino.

Albert Einstein

DECLARACIÓN DE AUTORÍA

Declaramos ser autores del presente trabajo “Mercado de datos Series históricas de agricultura, ganadería, silvicultura y pesca para el Sistema de Información de Gobierno”, 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 _____.

Livan López González

Firma del Autor

Liniuska Cardero Dieguez

Firma de la Autora

Ing. Themis Patricia Díaz Morales

Firma de la Tutora

DATOS DE CONTACTO

Tutora: Ing. Themis Patricia Díaz Morales

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: Instructor

Categoría Científica: No

Años de experiencia en el tema: 3

Años de graduado: 2

Correo Electrónico: tpdiaz@uci.cu

Agradecimientos

A mis padres por apoyarme en todo lo que he necesitado a pesar de todas las cosas por las que se han pasado, a mi hermano que a pesar de haber dejado su carrera me dio ánimo y fuerza para que hiciera mi sueño realidad, a mis abuelos que siempre han estado al tanto de todas mis preocupaciones, mis tíos y primos, en fin a toda esta familia poco perfecta que tengo pero que en el fondo sé que me quieren y me han dado la fuerza que tengo.

A mi tutora Themis Patricia por el tiempo dedicado y sacrificado para que esta tesis salga de la mejor manera posible.

A los 30 compañeros que vinieron a pasar este último año en la facultad 6, lejos de donde pasaron posiblemente los mejores momentos de su vida y en especial a Leisy, Chema, Osmil, Gretel, Evelin, Yanez, en fin a todos.

A todo el claustro de profesores que de una forma u otra han tenido que ver en mi preparación y en especial a los del centro DATEC por la ayuda incondicional y dedicación brindada. A todos los amigos que hice en los cuatro años que estuve en la regional de Artemisa con los cuales viví momentos muy felices, en especial a Leodan Díaz, Henry Chirino, Jennifer Álvarez, Yasmany Arzuaga, Robiel Perdomo y Nagyara Fernández.

A todas mis amistades del barrio que siempre se han preocupado por mí a pesar de mi ausencia debido a la dedicación que le he tenido que presentar a los estudios.

Livan López González

Agradecimientos

A Fidel Castro y a la Revolución por haberme brindado la increíble oportunidad de estudiar en esta magnífica universidad. A mis padres, por su apoyo y dedicación durante todos estos años, por ser mi principal fuente de inspiración ante cada paso importante de mi vida. A mis hermanos, por ser también un eterno apoyo ante cada situación, por constituir un ejemplo para mí en todos los sentidos, por su comprensión y cariño. A todos los profesores que han contribuido en mi formación durante toda mi vida estudiantil. A esos que he considerado mis amigos imprescindibles durante estos seis años de universidad y que hicieron posible también que hoy pueda haber logrado mi sueño, a Tay, Martha, Yoendy, Fabian, Manresa, Orelmis, Ransel, Susel y Yuniel; a mis amigos de la FEU, que tanto me apoyaron y enseñaron, en especial a Freddy y Jorge. A Leonel, por todo su apoyo, dedicación, amor y ayuda durante estos años de universidad. A todas las personas maravillosas que conocí en el CDI Las flores durante mi estancia en Venezuela y que desde entonces no han dejado de preocuparse por mis resultados, a Gina, Rubén, Bety, Migdalis, Arianna, Betsy, Kenia, Marcia, Vero, Vicente, Emilci, Madelsy, Geysi y Yalian. A mi tutora Themis Patricia por todo su esfuerzo, preocupación constante e interés por lograr un buen resultado. A mi dúo de tesis, por haber aceptado que formara parte de este trabajo a pesar de las circunstancias, por ser un chico emprendedor y muy capaz, pero sobre todas las cosas, por ser un excelente compañero.

Liniuska Cardero Dieguez

Dedicatoria

En primer lugar a mí por el sacrificio realizado todos estos años para lograr el sueño más grande que he tenido.

A mi papá por estar siempre atento a pesar de su ausencia física debido a su trabajo, gracias por todo lo que has hecho por mí y sé que este también es tu sueño.

A mi madre que a pesar de todas sus malcriadeces y las mías siempre me apoyó y me dio ánimo cuando mas lo necesité.

A mi hermano que todos los problemas que hemos tenido no nos han podido separar.

A todas las personas de las cuales estoy muy agradecido por la influencia positiva que tuvieron sobre mí para que este gran sueño se lograra.

Livan López González

Dedicatoria

A mi familia, en especial a mis padres, mi hermano, hermana y abuelos, por ser siempre un motivo de impulso en mi vida, por su amor y apoyo incondicional en cada momento y ante cada situación.

A mis amigos de la quinta graduación de la UCI, estén donde estén.

Liniuska Cardero Dieguez

RESUMEN

La presente investigación surge como parte de la colaboración que existe entre la Universidad de las Ciencias Informáticas (UCI) y la Oficina Nacional de Estadísticas e Información (ONEI). Dicha entidad tiene como principal objetivo obtener, analizar y difundir las cifras económicas y sociales del país. La UCI con el Centro de Tecnologías de Gestión de Datos (DATEC) de la facultad 6 asumió la tarea de crear un Almacén de Datos (AD) con el objetivo de integrar toda la información gestionada por la ONEI, este AD tiene como nombre Sistema de Información de Gobierno (SIGOB). En el presente Trabajo de Diploma se realiza el Mercado de Datos (MD) Series históricas de agricultura, ganadería, silvicultura y pesca, el cuál se integró al AD SIGOB. Este trabajo tiene como principal objetivo apoyar la toma de decisiones en esta área y permitir además almacenar gran cantidad de información, viabilizando así la integración de los datos de la institución que posteriormente serán analizados. Para la construcción del MD se realiza un estudio y caracterización de las metodologías, herramientas y tecnologías utilizadas en el desarrollo de este tipo de soluciones. Se realiza el análisis, diseño e implementación de los subsistemas de almacenamiento, integración y visualización. Finalmente, con la utilización de las listas de chequeo, los casos de prueba y la carta de aceptación del cliente se valida la solución, lo que permite verificar que el MD responde a una correcta visualización de las solicitudes de información hechas por los especialistas de la organización.

Palabras claves: Almacén de Datos, Centro de Tecnologías de Gestión de Datos, Mercado de Datos, Metodologías, Oficina Nacional de Estadísticas e Información, Sistema de Información de Gobierno, Universidad de las Ciencias Informáticas.

Índice	
INTRODUCCIÓN.....	1
CAPÍTULO 1: FUNDAMENTOS TEÓRICOS SOBRE EL DESARROLLO DE ALMACENES DE DATOS	
.....	4
Introducción	4
1.1 Gestión de la información de agricultura, ganadería, silvicultura y pesca en la Oficina Nacional de Estadística e Información.....	4
1.2 Almacenes de datos.....	4
1.2.1 Revisión conceptual	4
1.2.2 Características principales.....	6
1.2.3 Ventajas y desventajas de la utilización de un Almacén de Datos	9
1.3 Mercado de Datos.....	10
1.3.1 Revisión conceptual	10
1.3.2 Características principales.....	10
1.4 Experiencias del uso de los Almacenes de Datos	10
1.5 Etapas de desarrollo de un Almacén de Datos.....	11
1.5.1 Análisis y diseño.....	11
1.5.2 Extracción, transformación y carga.....	11
1.5.3 Inteligencia de negocios	12
1.6 Metodologías de desarrollo de un Almacén de Datos	13
1.6.1 Ciclo de vida Kimball	14
1.6.2 Modelo para el Desarrollo de soluciones de Almacenes de Datos e Inteligencia de Negocios	
.....	14
1.7 Herramientas para el desarrollo de Mercados de Datos.....	15
1.7.1 Herramientas de modelado	15
1.7.2 Herramientas de administración de Base de Datos	16
1.7.3 Herramientas para el proceso de extracción, transformación y carga.....	17
1.7.4 Herramientas de Inteligencia de Negocios.....	17
Conclusiones	18
CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS	19
Introducción	19
2.1 Definición del negocio	19
2.2 Temas de análisis	19

2.3 Reglas del Negocio	20
2.4 Descripción de los actores del sistema	20
2.5 Necesidades de los usuarios	20
2.5.1 Requisitos de Información	22
2.5.2 Requisitos Funcionales.....	23
2.5.3 Requisitos No Funcionales	24
2.6 Casos de Uso del Sistema	27
2.6.1 Descripción de los Casos de Uso críticos	28
2.7 Arquitectura	30
2.8 Diseño del subsistema de almacenamiento	31
2.8.1 Dimensiones.....	32
2.8.2 Hechos y medidas	32
2.8.3 Matriz bus.....	33
2.8.4 Modelo de datos	34
2.9 Diseño del subsistema de integración	35
2.10 Diseño del subsistema de visualización	35
2.11 Políticas de seguridad.....	36
2.11.1 Salva de la Base de Datos	37
Conclusiones	37
CAPÍTULO 3: IMPLEMENTACIÓN DEL MERCADO DE DATOS	38
Introducción	38
3.1 Implementación del subsistema de almacenamiento	38
3.2 Implementación del subsistema de integración	39
3.2.1 Implementación de los procesos de Extracción, Transformación y Carga	39
3.2.2 Implementación del trabajo.....	40
3.3 Implementación del subsistema de visualización	41
3.3.1 Implementación de los cubos OLAP	41
Arquitectura de información.....	41
3.3.2 Implementación de los reportes.....	42
Conclusiones	42
CAPÍTULO 4: VALIDACIÓN DEL MERCADO DE DATOS.....	44
Introducción	44
4.1 Pruebas	44

4.2 Herramientas para validar el Mercado de Datos	45
4.2.1 Listas de chequeo	45
4.2.2 Casos de prueba	48
4.3 Evaluación de los resultados.....	49
Conclusiones	51
CONCLUSIONES	52
RECOMENDACIONES	53
REFERENCIAS BIBLIOGRÁFICAS	54
BIBLIOGRAFÍA.....	55

INTRODUCCIÓN

La capacidad de razonamiento del hombre ha hecho posible que todos los aspectos relacionados con sus principales necesidades hayan alcanzado un importante auge en casi todos los sectores de la sociedad, dándole paso a la creación de nuevos métodos, herramientas y tecnologías. Los avances tecnológicos han sido un eslabón fundamental para el desarrollo de la humanidad y con este el de las tecnologías de la información. El gran esfuerzo que se ha realizado en Cuba por experimentar los logros relacionados con la informática y la computación es incuestionable, por lo que se ha alcanzado un vertiginoso progreso en estas ciencias y se ha adquirido experiencia durante el transcurso de la última década. En aras de lograr la informatización del país y en medio de un sin número de sucesos, surge la Universidad de Ciencias Informáticas (UCI), institución que se caracteriza por producir software. La UCI cuenta con diversos centros productivos, entre los que se encuentra el Centro de Tecnologías de Gestión de Datos (DATEC). Este centro tiene la misión de proveer soluciones integrales, productos y servicios que apoyan la gestión de información. Dentro de sus áreas temáticas podemos encontrarlos “Almacenes de Datos”, el cual se encuentra trabajando en conjunto con la Oficina Nacional de Estadística e Información (ONEI).

La ONEI es el órgano rector de la estadística en Cuba, que abarca las diferentes esferas socioeconómicas del país. Para un mejor control de dicha entidad la información gestionada se divide en áreas, entre las que se encuentra agricultura, ganadería, silvicultura y pesca. Dentro de las deficiencias identificadas en la ONEI en esta área, se encuentran que los datos no pueden ser consultados a no ser por especialistas de la informática y con un alto conocimiento del negocio. Los ficheros son generados anualmente, por lo que se hace muy difícil la obtención de información, la cual va acumulándose año tras año y es cada vez más complejo su análisis. Los datos no están integrados, lo que atenta contra la calidad de la información, la cual posee diferente codificación, provocando la existencia de múltiples versiones en los datos y afectando la efectividad del tratamiento de la información. El análisis estadístico se realiza a través de mecanismos no automatizados, poco confiables y en algunas ocasiones resulta tedioso, a causa de que la información es almacenada en herramientas basadas en aplicaciones de oficina como Excel. La recuperación y creación de los informes se torna engorroso y costoso en cuanto a tiempo y esfuerzo. El proyecto Sistema de Información de Gobierno (SIGOB) arrojó como parte de un trabajo de diploma precedente el MD correspondiente al área de agricultura, ganadería y silvicultura. Las series históricas de agricultura, ganadería, silvicultura y pesca no fueron incluidas en esta solución, imposibilitando un correcto proceso de análisis y aprovechamiento de la información manejada para la toma de decisiones.

Por toda la situación anteriormente descrita se plantea como **problema de la investigación**: ¿cómo contribuir a la toma de decisiones correspondiente a los datos que pertenecen a las series históricas en el área de agricultura, ganadería, silvicultura y pesca del Sistema de Información de Gobierno?

La investigación tiene como **objeto de estudio** los Almacenes de Datos, enmarcado en el **campo de acción** Mercado de Datos Series históricas de agricultura, ganadería, silvicultura y pesca.

El **objetivo general** de este trabajo es desarrollar el Mercado de Datos Series históricas de agricultura, ganadería, silvicultura y pesca para el Sistema de Información de Gobierno.

En correspondencia con el objetivo general se plantean los siguientes **objetivos específicos**:

1. Fundamentar la selección de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de los Almacenes de Datos.
2. Realizar el análisis y diseño del Mercado de Datos Series históricas de agricultura, ganadería, silvicultura y pesca para el Sistema de Información de Gobierno.
3. Realizar la implementación del Mercado de Datos Series históricas de agricultura, ganadería, silvicultura y pesca para el Sistema de Información de Gobierno.
4. Realizar la validación del Mercado de Datos Series históricas de agricultura, ganadería, silvicultura y pesca para el Sistema de Información de Gobierno.

Para darle cumplimiento a los objetivos específicos se definen las siguientes **tareas de investigación**:

1. Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de los Almacenes de Datos.
2. Levantamiento de requisitos.
3. Descripción de los casos de uso del Mercado de Datos.
4. Definición de los hechos, las medidas y las dimensiones del Mercado de Datos.
5. Diseño del modelo de datos.
6. Definición de la arquitectura del Mercado de Datos.
7. Diseño del subsistema de integración.
8. Diseño del subsistema de visualización.
9. Diseño de los casos de prueba.
10. Implementación del modelo de datos.
11. Implementación del subsistema de integración.
12. Implementación del subsistema de visualización.

13. Aplicación de las listas de chequeo.

14. Aplicación de los casos de prueba.

El Trabajo de Diploma está estructurado de la siguiente manera: introducción, cuatro capítulos, conclusiones, recomendaciones, referencias bibliográficas, bibliografía, anexos y glosario de términos.

Capítulo 1: Fundamentos teóricos sobre el desarrollo de Almacenes de Datos

En este capítulo se abordan definiciones y conceptos relacionados con AD y MD, así como características, ventajas y desventajas de su utilización. Asimismo, se documentan las metodologías, herramientas y tecnologías para el desarrollo de un AD.

Capítulo 2: Análisis y diseño del Mercado de Datos

En este capítulo se realiza un análisis profundo y detallado del negocio, con el propósito de comprender los principales aspectos de relevancia para la organización. Se especifican las necesidades de información, Reglas del Negocio (RN), Requisitos Funcionales (RF), Requisitos No Funcionales (RNF) y Requisitos de Información (RI), así como los Casos de Uso del Sistema (CUS). Además se realiza el diseño de los subsistemas de almacenamiento, integración y visualización.

Capítulo 3: Implementación del Mercado de Datos

En este capítulo se implementan los procesos definidos en el diseño, como el subsistema de almacenamiento, definiendo la estructura de los datos, esquemas y tablas de la base de datos, así como el subsistema de integración, a través del trabajo y los flujos de transformación. Se realiza la implementación del subsistema de visualización de los datos, mediante la estructura de navegación y los reportes candidatos.

Capítulo 4: Validación del Mercado de Datos

En este capítulo se valida la solución, mediante la utilización de las listas de chequeo, los casos de prueba y la carta de aceptación del cliente.

CAPÍTULO 1: FUNDAMENTOS TEÓRICOS SOBRE EL DESARROLLO DE ALMACENES DE DATOS

Introducción

En este capítulo se abordan definiciones y conceptos relacionados con AD y MD, así como características, ventajas y desventajas de su utilización. Asimismo, se documentan las metodologías, herramientas y tecnologías para el desarrollo de un AD.

1.1 Gestión de la información de agricultura, ganadería, silvicultura y pesca en la Oficina Nacional de Estadística e Información

En la actualidad, la gestión de la información constituye uno de los procesos claves a los que tienen que enfrentarse las organizaciones. Para Cuba es importante llevar el control estadístico sobre los datos relacionados con la agricultura, ganadería, silvicultura y la pesca, para determinar el comportamiento de sus principales indicadores, dando la voz de alerta sobre la existencia de un problema y permitiendo tomar medidas para solucionarlo. Para llevar todo este control, la ONEI recoge los datos relacionados con los principales indicadores de este sector en tablas, las cuales se encuentran en formato Excel, estos datos dependen de información censal de un período de tiempo. Para analizarlos es necesaria la participación de varios especialistas en la materia, quienes realizan el análisis de los datos de forma manual con la ayuda de herramientas informáticas. Los analistas deben realizar un trabajo minucioso, revisando tabla por tabla para detectar incongruencias en los datos calculados, en un intervalo de tiempo entre un censo y otro. El proceso descrito anteriormente trae consigo que se dificulte la manera de realizar los análisis estadísticos sobre las esferas de la agricultura, ganadería, silvicultura y pesca; corriéndose el riesgo de perder información útil al no contar con una herramienta informática que contribuya a mejorar la eficiencia del tratamiento de la información.

1.2 Almacenes de datos

1.2.1 Revisión conceptual

La tecnología de almacenamiento de datos se ha ido posicionando como la vía más acertada para la realización de análisis de información histórica, además de convertirse en una potente herramienta para la recuperación efectiva de las más complejas consultas y de servir como base para la toma de decisiones. En cualquier revisión que se realice sobre lo que se entiende por AD, es difícil encontrar una concepción acabada y compartida por los autores, por el contrario, existen diversas aproximaciones teóricas; lo que demuestra que se trata de una herramienta en evolución y de compleja concepción. A continuación se muestran (ver tabla 1) las concepciones de diferentes autores [1].

Tabla 1: Almacén de Datos. Conceptos

ALMACÉN DE DATOS. CONCEPTOS.
<i>(1) Depósito (o archivo) de la información reunida a partir de varias fuentes, guardada según un esquema unificado en un único lugar. Una vez reunidos, los datos se guardan durante un tiempo, lo que permite el acceso a datos históricos.</i>
<i>(2) Un almacén de datos es un repositorio de datos que almacena e integra información procedente de toda la organización, asegurando una gestión más eficiente de la misma y proyectando una visión única de la realidad de la compañía.</i>
<i>(3) Colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia.</i>
<i>(4) Un AD es la colección de datos, organizados, integrados, historizados y disponibles para facilitar la toma de decisiones de usuarios finales.</i>
<i>(5) Un AD es una copia de los datos transaccionales específicamente estructurada para la consulta y el análisis.</i>
<i>(6) Colección de datos orientados a temas, integrada, variante en el tiempo, no volátil, que añade la geografía del dato.</i>
<i>(7) Depósito donde se almacenan los datos que la organización utiliza para saber cómo está funcionando. El almacenamiento de datos concentra mucha información proveniente de los procesos, de los sistemas operativos y financieros de los ERP y CRM.</i>
<i>(8) Almacenamiento efectivo, filtrado y ordenado de los datos estratégicos, tácticos y operativos, que permita su extracción y análisis de manera ordenada, clara, funcional y efectiva.</i>
<i>(9) Bodega donde están almacenados todos los datos necesarios para realizar las funciones de gestión de la empresa, de manera que puedan utilizarse fácilmente según se necesiten.</i>
<i>(10) Base de datos que almacena una gran cantidad de datos transaccionales integrados para ser usados para análisis gestionables por usuarios especializados (tomadores de decisión de la empresa).</i>
<i>(11) Colección de datos en la cual se encuentra integrada la información de la Institución y que se usa como soporte para el proceso de toma de decisiones gerenciales.</i>
<i>(12) Almacén de información temática orientado a cubrir las necesidades de</i>

aplicaciones de los Sistemas de Soporte de Decisiones (DSS) y de la Información de Ejecutivos (EIS), que permite acceder a la información corporativa para la gestión, control y apoyo a la toma de decisiones.

(13) Almacenamiento de información homogénea y fiable, en una estructura basada en la consulta y el tratamiento jerarquizado de la misma, y en un entorno diferenciado de los sistemas operacionales.

(14) Herramienta que se nutre de las bases de datos de gestión y de otras externas que permite, con total flexibilidad y en tiempo real, obtener y combinar todo tipo de datos, indicadores, comparativas y simulaciones para la simple información, el conocimiento, el análisis y la toma de decisiones.

En la presente investigación se asumirá la definición (11), ya que se considera el concepto más completo partiendo de las características de un AD.

1.2.2 Características principales

Un AD posee diversas características que lo distingue del resto de las tecnologías de almacenamiento de datos. Estos elementos contribuyen a que sea considerado el centro de atención por muchas empresas y organizaciones en los últimos tiempos cuando de gestión de información y toma de decisiones se trate. Dentro de sus principales características se encuentran [2]:

- Organizado por temas
- Integrado
- Dependiente del tiempo
- No volátil

Organizado por temas

La información se clasifica en base a los aspectos que son de interés para la empresa, los datos están organizados por materias o temas (ver figura 1). Esta característica contrasta con el clásico método orientado al proceso y funcionamiento de las aplicaciones utilizadas en sistemas operacionales más antiguos.



Figura 1: Organizado por temas

Integrado

Es considerado el aspecto más importante dentro de las características de un AD. Los datos necesitan ser almacenados en el AD de una forma globalmente aceptable y singular, aunque el programa operacional los almacene de una forma distinta (ver figura 2). Consiste en convenciones de nombres, codificaciones consistentes y medida uniforme de variables.

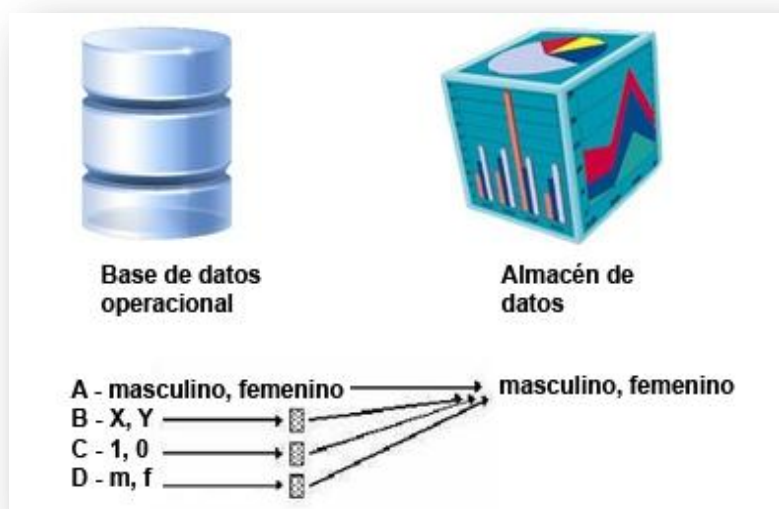


Figura 2: Integrado

Dependiente del tiempo

La dependencia del tiempo se observa en tres formas (ver figura 3):

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



Figura 3: Dependiente del tiempo

No volátil

Solo se permite la lectura y escritura de los datos, no incluyendo en esto la modificación o eliminación de datos existentes (ver figura 4). La ventaja principal de este tipo de Base de Datos (BD) radica en las estructuras en las que se almacena la información (modelos de tablas en estrella, en copo de nieve, cubos relacionales). Este tipo de persistencia de la información es homogénea y fiable, permitiendo la consulta y tratamiento jerarquizado de la misma (siempre en un entorno diferente a los sistemas operacionales) [3].



Figura 4: No volátil

1.2.3 Ventajas y desventajas de la utilización de un Almacén de Datos

El uso de AD presenta ventajas y desventajas para las empresas, su implementación puede beneficiar a una organización atendiendo a los siguientes aspectos:

- Integra y consolida diferentes fuentes de datos en una única plataforma sólida y centralizada.
- Provee la capacidad de analizar y explotar toda la información que posee.
- Permite reaccionar rápidamente a los cambios del mercado. Aumenta la competitividad en el mercado. Mejora la entrega de información, es decir, información completa, correcta, consistente, oportuna y accesible. Aprovecha el enorme valor potencial de los recursos de información y los transforma en valor verdadero.
- Permite al usuario adquirir mayor confianza acerca de sus propias decisiones y de las del resto, logrando así, un mayor entendimiento de los impactos ocasionados.
- Los usuarios pueden acceder directamente a la información en línea, lo que contribuye a su capacidad para operar con mayor efectividad en las tareas rutinarias o no.
- Los usuarios pueden tener a su disposición una gran cantidad de información multidimensional, presentada coherentemente como fuente única, confiable y disponible en sus estaciones de trabajo.

Utilizar AD también trae consigo algunos problemas como:

- La subestimación del tiempo requerido para extraer, limpiar y cargar los datos en el AD.
- Problemas con los sistemas de origen de los datos.
- Los datos obtenidos no son suficientes.

- Pueden suponer gastos muy elevados no solo en su elaboración sino también en su mantenimiento.

1.3 Mercado de Datos

1.3.1 Revisión conceptual

Los MD se pueden ver como pequeños AD, teniendo en cuenta que estos son creados de forma departamental. Un concepto más formal sería: *“es una BD departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento. Un MD puede ser alimentado desde los datos de un AD, o integrar por sí mismo un compendio de distintas fuentes de información”* [4]. De forma sencilla, se puede definir un MD como un AD que permite ser consultado rápidamente, pero a un nivel más pequeño (áreas), mientras que el AD es a nivel de toda la empresa.

1.3.2 Características principales

Los MD representan elementos básicos dentro de un AD, al constituir estos los componentes de un AD. Dentro de sus principales características se pueden encontrar:

- Poseen una estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos del departamento al cual está aplicado.
- Son más sencillos a la hora de utilizarlos y comprender sus datos, debido a que la cantidad de información que contienen es mucho menor que los AD.
- Se centran en los requisitos de los usuarios asociados a un departamento o área de negocio concreta.

1.4 Experiencias del uso de los Almacenes de Datos

En la actualidad los AD son muy usados en todo el mundo, surgieron como respuesta a la problemática de ¿qué hacer para lograr que la cantidad de información almacenada por grandes y medianas empresas en las distintas fuentes de datos fuera útil?, constituyendo uno de los soportes fundamentales para el proceso de toma de decisiones. Son numerosas las compañías que hacen uso de estas tecnologías entre las que se encuentran: Twentieth Century Fox para gestionar la información relacionada con las películas que se proyectan en distintos lugares de los Estados Unidos para predecir qué actores, argumentos y filmes serán más populares, con el objetivo de ganar audiencia en sus producciones. La compañía Wal-Mart, es una red mayorista de productos comestibles en los Estados Unidos, considerada la empresa más grande a nivel mundial. Esta cuenta con el uso de un AD para tomar decisiones acerca de todos los procesos que se realizan en el mercado internacional para de esta manera elevar su economía y mantenerse en competencia respecto a otras compañías.

En Cuba también existen centros que hacen uso de los AD como herramienta para mejorar la toma de decisiones, un ejemplo de ello es la empresa comercializadora CIMEX (Corporación de Exportaciones e Importaciones), sobresaliente por su estabilidad financiera y crecimiento constante, utilizando los AD para la gestión de inventarios. En la UCI también se ha venido experimentando en este campo, ejemplo de ello es el AD para la toma de decisiones en cuanto al consumo energético de la universidad. También el Centro de Inmunología Molecular (CIM) utiliza un AD, desarrollado por la UCI para analizar los ensayos clínicos que se gestionan en dicho centro. En colaboración con la ONEI y en apoyo al proceso de informatización del país, la universidad está desarrollando un AD para el apoyo a la toma de decisiones en esta institución, nombrado SIGOB.

1.5 Etapas de desarrollo de un Almacén de Datos

1.5.1 Análisis y diseño

Dentro del proceso de desarrollo de software, una de las etapas de mayor relevancia e impacto en el resultado final de cualquier producto informático lo constituye el análisis y diseño. El objetivo de esta etapa es tener un control acerca de las necesidades de los clientes y poder obtener finalmente un sistema que responda a los intereses del negocio. El análisis es el eslabón fundamental para el desarrollo del MD, pues a partir de él se sientan las bases para los posteriores procesos de diseño e implementación. Realizar el proceso de análisis es una labor compleja, partiendo de que se necesita un estudio del proceso del negocio que se pretende informatizar para entender de manera clara y transparente lo que el cliente necesita.

En la fase de análisis se generan un conjunto de artefactos que facilitan el desarrollo del sistema, permitiendo un mejor entendimiento del negocio y del funcionamiento de la organización en general. La realización de estos artefactos sitúa el avance del cumplimiento de las tareas planteadas, garantizando la utilidad y el éxito del diseño de las estructuras. Es necesario tener una descripción detallada acerca de la entidad que se desempeña como cliente. Durante el período de análisis se tiene en cuenta el levantamiento de los requisitos, estableciendo una meta para los desarrolladores en la fase de implementación, donde se definen las relaciones entre los hechos y dimensiones, se especifican los actores del negocio y del sistema, así como su relación con los diferentes Casos de Uso (CU), desarrollando el diagrama de CUS. En la etapa de diseño se construye la matriz bus, se obtiene y transforma el modelo lógico a modelo físico, quedando confeccionado de esta manera el diseño de la BD.

1.5.2 Extracción, transformación y carga

Es el proceso que organiza el flujo de los datos entre diferentes sistemas en una organización y aporta los métodos y herramientas necesarias para mover datos desde múltiples fuentes a un AD, reformatearlos, limpiarlos y cargarlos en otra BD [5].

Extracción

Es el proceso realizado sobre las fuentes de datos que disponen los clientes que deben ser introducidas al MD. Este proceso comienza con la extracción de los datos a partir de las fuentes dadas que provienen de diferentes sistemas de origen. Los datos se encuentran normalmente en BD relacionales, ficheros planos u otras estructuras diferentes. Esta parte del proceso convierte los datos a un formato preparado para iniciar el proceso de transformación.

Transformación

Una vez culminada la extracción de los datos, estos son analizados y transformados basándose precisamente en las RN o funciones sobre los datos extraídos. Luego de aplicadas las transformaciones sobre los datos, estos quedan listos para ser cargados en el AD.

Carga

Siendo esta la última fase del proceso de ETL se realiza una interacción directa con la BD destino, ya que los datos tienen que ser incluidos en el sistema. En los AD la información no puede sobrescribirse ya que estos guardan información histórica de los datos de manera que estos pueden ser auditados en cualquier momento.

1.5.3 Inteligencia de negocios

Luego de concluidos los procesos de ETL en el desarrollo de un MD el próximo paso para cumplir con el objetivo final es lograr la visualización de la información de manera correcta y comprensible para el cliente. Lo que se puede implementar a través de los procesos de BI, que en su conjunto ayudan al usuario a la exploración de los datos y generación de vistas de información. El análisis de datos es un proceso en el que, a través de las distintas técnicas de análisis, como: OLAP (por sus siglas en inglés Online Analytical Processing), la minería de datos, los reportes y consultas, se le da el valor real de los datos al extraer de ellos la información requerida para auxiliar la toma de decisiones y mostrarla de manera entendible.

La técnica OLAP está basada en BD orientadas al procesamiento analítico. Este análisis suele implicar, generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil. Este sistema es típico de los MD.

- El acceso a los datos suele ser de solo lectura. La acción más común es la consulta, con muy pocas inserciones, actualizaciones o eliminaciones.
- Los datos se estructuran según las áreas de negocio, y los formatos de los datos están integrados de manera uniforme en toda la organización.
- El historial de datos es a largo plazo, normalmente de dos a cinco años.
- Las BD OLAP se suelen alimentar de información procedente de los sistemas operacionales existentes, mediante un proceso de extracción, transformación y carga (ETL) [6].

Con el paso del tiempo y la utilización continua de estas tecnologías las diferentes empresas sintieron la necesidad de buscar nuevas vías para trabajar sus datos, de esta forma surgen los sistemas MOLAP (por sus siglas en inglés Multidimensional Online Analytical Process), ROLAP (por sus siglas en inglés Relational Online Analytical Process) y HOLAP (por sus siglas en inglés Hybrid Online Analytical Process), utilizando estos a OLAP como base.

MOLAP

La arquitectura MOLAP usa bases de datos multidimensionales para proporcionar el análisis. Un sistema MOLAP usa una BD propietaria multidimensional, en la que la información se almacena multidimensionalmente, para ser visualizada en varias dimensiones de análisis.

ROLAP

La arquitectura ROLAP, accede a los datos almacenados en un AD para proporcionar los análisis OLAP. La premisa de los sistemas ROLAP es que las capacidades OLAP se soportan mejor contra las bases de datos relacionales.

HOLAP

Un desarrollo un poco más reciente ha sido la solución OLAP híbrida (HOLAP), la cual combina las arquitecturas ROLAP y MOLAP para brindar una solución con las mejores características de ambas: desempeño superior y gran escalabilidad. Un tipo de HOLAP mantiene los registros de detalle (los volúmenes más grandes) en la BD.

En la solución propuesta se utiliza ROLAP debido a que el Sistema Gestor de Bases de Datos (SGBD) que se utiliza en la solución es PostgreSQL 9.1 donde se modela un diseño relacional que simula uno multidimensional a través de hechos y dimensiones. De esta forma el esquema obtenido es un esquema relacional, constelación de hechos, dada las características del gestor al no modelar directamente modelos multidimensionales.

1.6 Metodologías de desarrollo de un Almacén de Datos

Una metodología es el conjunto de métodos que rigen una investigación científica [7]. Dentro del mundo de la industria del software existen diversos tipos de metodologías, que a partir de sus características permiten a los ingenieros seleccionar la más adecuada para guiar todo el proceso de desarrollo, logrando un producto final de calidad y poco costoso en tiempo y esfuerzo. Lo mismo ocurre en el campo de desarrollo de los AD.

Las dos vertientes fundamentales para el desarrollo de almacenes de datos son las planteadas por Ralph Kimball y Bill Inmon. Kimball (principal promotor del enfoque dimensional para el diseño de almacenes de datos), considera que un AD es una copia de los datos transaccionales específicamente estructurada para la consulta y el análisis. Bill Inmon (conocido por muchos como el padre del AD), plantea que un AD es un conjunto de datos orientados por temas, integrados, variantes en el tiempo y

no volátiles, que tienen por objetivo dar soporte a la toma de decisiones [8]. Lo propuesto por Inmon, puede tener una implementación mucho más tardada, es recomendada cuando se hace demasiado difícil representar el modelo a través de dimensiones y la complejidad de la solución se hace demasiado grande. Por el contrario, la de Kimball es la más aceptada en todo el mundo como la más efectiva para desarrollar una solución de construcción de AD. Además es de fácil comprensión y rápida de implementar por etapas.

Existen en el mundo diferentes metodologías para el desarrollo de AD entre las que se encuentran Metodología Hefesto [9], Desarrollo de almacenes de datos dirigidos por modelos (Trujillo), Data Warehouse Engineering Process (DWEPE), Rapid Warehousing Methodology (RWM) y el Ciclo de vida Kimball.

1.6.1 Ciclo de vida Kimball

El ciclo de vida Kimball comienza con una planificación de proyecto, donde se define el alcance, se identifican y programan las tareas, se planifica el uso de los recursos, conformando con todo esto el plan de proyecto. En la segunda etapa de este ciclo se definen los requerimientos del negocio. Luego de definir los requerimientos del negocio, el proyecto se enfoca en tres líneas concurrentes: tecnología, datos y aplicaciones de la BI (ver figura 5).

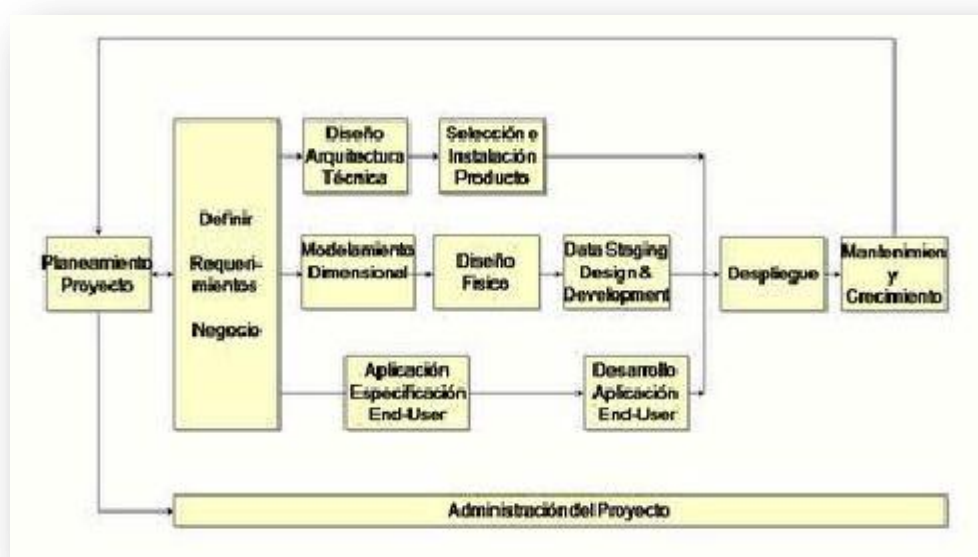


Figura 5: Ciclo de vida Kimball

1.6.2 Modelo para el Desarrollo de soluciones de Almacenes de Datos e Inteligencia de Negocios

La metodología utilizada por DATEC denominada Modelo para el Desarrollo de Soluciones de AD e BI en DATEC, abarca todas las fases para la construcción de un AD, comenzando con el levantamiento

de información inicial hasta la capa de visualización. Este modelo toma como base el ciclo de vida Kimball, además tiene en cuenta lo planteado por Leopoldo Zenaido Zepeda en su tesis de doctorado en cuanto al uso de Casos de Uso Informativos (CUI) [10]. Una primera fase inicia con el levantamiento de información a nivel de negocio para identificar los posibles indicadores y aspectos a medir en los análisis, quedando conformados los requisitos de información.

Se lleva a cabo un estudio de las fuentes de datos a cargar. Se garantiza que la información levantada sobre las necesidades de los clientes esté realmente almacenada en las fuentes correspondientes, para posteriormente, teniendo los requisitos de información correctamente definidos, proceder a diseñar la solución del AD. Una vez diseñada la estructura del AD, se realiza la carga de los datos desde las fuentes y posteriormente se implementan los procesos de BI. Las actividades y artefactos de la solución son realizados por cuatro grupos que conforman la línea, especializados en componentes específicos de la solución.

Ventajas del Modelo para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio en DATEC:

- La solución completa se puede implementar en poco tiempo.
- Cuenta con mayor velocidad de respuesta al cliente.
- Los productos son más comprensibles para los usuarios.
- Es resistente y tolerante ante los cambios.

Para el presente trabajo de diploma se decide utilizar el modelo propuesto por DATEC, ya que toma como base el ciclo de vida Kimball, adaptado a las políticas y necesidades tanto del centro como de la UCI, cubriendo todas las fases de construcción de un AD.

1.7 Herramientas para el desarrollo de Mercados de Datos

1.7.1 Herramientas de modelado

En el proceso de desarrollo de un producto informático es necesario realizar un correcto modelado de los artefactos, para esto se realiza un estudio de varias herramientas que utilizan técnicas de diseño automatizadas. Dentro de ellas se destacan las CASE (Computer Aided Software Engineering), que proveen a los ingenieros de software apoyo en el modelado de las soluciones.

A continuación se nombran algunas herramientas CASE orientadas a UML (Lenguaje Unificado de Modelado).

- Rational Rose
- ArgoUML
- Visual Paradigm

En el presente trabajo de diploma se decidió utilizar Visual Paradigm 8.0, ya que este 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 [11]. Permite realizar 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.

La herramienta es colaborativa, soporta múltiples usuarios trabajando sobre el mismo proyecto; genera la documentación automáticamente en varios formatos como Web o PDF y permite el control de versiones. Cabe destacar igualmente su usabilidad y portabilidad [12].

1.7.2 Herramientas de administración de Base de Datos

Los Sistemas Gestores de Base de Datos (SGBD) fueron diseñados para gestionar grandes volúmenes de información, tanto la definición de estructuras para el almacenamiento como los mecanismos para la gestión de los datos. Estos permiten a los usuarios definir, crear, mantener la BD y proporcionar un acceso controlado en todo momento.

Existen diversos SGBD que son usados por profesionales de la información y que se encuentran disponibles para el uso y desarrollo de aquellos interesados en el área de la gestión de la información. El 1 de Julio del 2009 el Grupo Global de Desarrollo de PostgreSQL liberó la versión 9.1, continuando con el rápido desarrollo de bases de datos de código abierto más avanzada del mundo. Esta versión contiene una gran cantidad de mejoras para hacer la administración, consulta y programación en PostgreSQL mucho más fácil que nunca.

Algunas de las mejoras realizadas sobre esta herramienta son [13]:

- Es estable y segura.
- Funciona en la mayoría de los sistemas operativos actuales.
- Posee puntos de recuperación en un momento dado.
- Posee replicación asincrónica.
- Permite la restauración de bases de datos en procesos paralelos, acelerando la recuperación de un respaldo hasta ocho veces.
- Privilegios por columna, que permiten un control más granular de datos confidenciales.
- Configuración de ordenamiento configurable por BD, lo cual hace a PostgreSQL más útil en entornos con múltiples idiomas.
- Nuevas herramientas de monitoreo de consultas que le otorgan a los administradores mayor información sobre la actividad del sistema.

En este trabajo queda definido como SGBD a utilizar PostgreSQL 9.1 para el desarrollo del MD Series históricas agricultura, ganadería, silvicultura y pesca para el SIGOB.

1.7.3 Herramientas para el proceso de extracción, transformación y carga

Una de las herramientas más potentes para este proceso es Pentaho Data Integration (PDI) 4.0.1 la cual abre, limpia e integra la información y la pone en manos del usuario proporcionando consistencia, una sola versión de todos los recursos de información y vías para la extracción, transformación y carga de datos. Algunas de las principales características de esta herramienta son:

- Posee entorno gráfico de desarrollo.
- Incluye el uso de tecnologías estándar: Java, XML, Java Script.
- Fácil de instalar y configurar.
- Es multiplataforma: Windows, Macintosh, Linux.
- Está basado en dos tipos de objetos: transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones).
- Es de código abierto.
- Sin costes de licencia.

Incluye cuatro herramientas fundamentales:

Spoon: para diseñar las transformaciones usando el entorno gráfico.

PAN: para ejecutar transformaciones diseñadas con Spoon.

CHEF: para crear trabajos.

Kitchen: para ejecutar trabajos.

También se hace necesario una herramienta para la integración, carga, limpieza y estandarización de los datos, para ellos se utiliza el DataCleaner 1.5.3.

El DataCleaner (DC) es una aplicación Open Source para el perfilado, la validación y comparación de datos, estas actividades ayudan a administrar y supervisar la calidad de estos, con el fin de garantizar que la información sea útil. Es la alternativa gratuita al software de gestión de datos maestros, metodologías de almacenamiento de datos, proyectos de investigación estadística y la preparación para el proceso de ETL.

1.7.4 Herramientas de Inteligencia de Negocios

Pentaho Workbench

Es una herramienta de análisis caracterizada por su potencia gráfica y capacidad multitarea. Posee un entorno visual para el desarrollo y prueba de los cubos OLAP, brinda la posibilidad de configurar una conexión JDBC (por sus siglas en inglés Java Database Connectivity) con el modelo físico. Provee un mecanismo para buscar datos con rapidez y tiempo de respuesta uniforme independientemente de la cantidad de datos en el cubo o la complejidad del procedimiento de búsqueda.

Mondrian OLAP Server

Como su nombre lo dice es un servidor OLAP, el cual se encarga de gestionar la comunicación entre una aplicación OLAP y la BD con los datos fuentes. Permite realizar consultas a un AD logrando que los resultados sean presentados mediante un navegador. Mondrian utiliza MDX (por sus siglas en inglés Multidimensional QueryExpression) como lenguaje de consulta, que fue un lenguaje propuesto por Microsoft. Funciona sobre las bases de datos estándares del mercado: Oracle, SQL-Server, MySQL, PostgreSQL, lo cual habilita y facilita el desarrollo del negocio basado en la plataforma Pentaho.

Servidor web Apache Tomcat

La suite de Pentaho no incluye esta herramienta, pero si la utiliza como servidor web. Tomcat es un servidor de código abierto, un contenedor de aplicaciones web basadas en Java que fue creado para ejecutar Servlets y Java Server Page (por sus siglas en inglés JSP), de aplicaciones web.

Se decide utilizar el servidor Mondrian OLAP 3.0.4, ya que tiene un alto desempeño, reflejado en el análisis interactivo de grandes o pequeños volúmenes de información. Debido a que debe ser ejecutado sobre un servidor web que soporte el lenguaje JSP se decidió utilizar el servidor web Apache Tomcat5.5. En el diseño de los cubos que se cargarán en el servidor Mondrian se propone utilizar la herramienta Pentaho Shema Workbench 3.2.0. Esta posee un entorno visual muy cómodo para el desarrollo y prueba de los cubos OLAP, que soporta además las consultas MDX, permitiendo la realización de pruebas y corrección de consultas sobre los cubos.

Conclusiones

Luego del desarrollo de este capítulo se pudieron arribar a las siguientes conclusiones:

- Se realizó un estudio sobre las metodologías, herramientas y técnicas de desarrollo para la construcción del MD.
- Se decidió utilizar como metodología de desarrollo el Modelo para el desarrollo de soluciones de AD e BI de DATEC (DW&BI), el cual toma como base la metodología Ciclo de vida Kimball.
- Se definieron como herramientas a utilizar el Visual Paradigm 8.0 como herramienta de modelado, como SGBD PostgreSQL9.1, para el proceso de ETL el Pentaho Data Integration 4.0.1, para el perfilado de los datos el DataCleaner 1.5.3 y para desarrollar los procesos de BI el Pentaho Shema Workbench 3.2.0 para el diseño de los cubos y el Mondrian OLAP 3.0.4 sobre el servidor web Apache Tomcat 5.5.

CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS

Introducción

En este capítulo se realiza un análisis profundo y detallado del negocio, con el propósito de comprender los principales aspectos de relevancia para la organización. Se especifican las necesidades de información, Reglas del Negocio (RN), Requisitos Funcionales (RF), Requisitos No Funcionales (RNF) y Requisitos de Información (RI), así como los Casos de Uso del Sistema (CUS). Además se realiza el diseño de los subsistemas de almacenamiento, integración y visualización.

2.1 Definición del negocio

La Oficina Nacional de Estadística e Información (ONEI), desde su creación en el año 1994 tiene como objetivo captar, analizar y difundir la información referente a las distintas esferas de la sociedad cubana [14]. Una de las áreas a ser analizadas por dicha entidad es la de agricultura, ganadería, silvicultura y pesca.

La información gestionada por la ONEI en este departamento se encuentra en series históricas, las cuales incluyen datos de distintos indicadores de manera anual. Estas series son almacenadas en ficheros de tipo Excel, provocando tardanza en la elaboración de informes y resultando costoso en tiempo y esfuerzo. El análisis de esta información por los especialistas de la ONEI constituye un proceso tedioso, pues son grandes volúmenes de datos y varios indicadores a analizar, además solo puede realizarse por un especialista con alto conocimiento del negocio. Todos estos aspectos afectan en gran medida el análisis de la información en el área de agricultura, ganadería, silvicultura y pesca y con ello el proceso de toma de decisiones.

2.2 Temas de análisis

Un tema de análisis es la división o clasificación de la información de una organización de acuerdo a las diferentes temáticas que esta incluye. Tiene como propósito lograr un mejor entendimiento del negocio y permite además agrupar las necesidades de información en correspondencia a los temas definidos. En el caso de la presente investigación se definieron los siguientes temas de análisis:

- Agricultura
- Ganadería
- Silvicultura
- Avicultura
- Apicultura
- Pesca

2.3 Reglas del Negocio

Las RN son condiciones que se establecen y deben tenerse en cuenta durante todo el proceso de desarrollo, la calidad del producto final depende en gran medida del cumplimiento de estas. Para la construcción del MD Agricultura, ganadería, silvicultura y pesca se definieron las siguientes reglas del negocio:

RN1:

Las dimensiones no deben contener campos nulos.

RN2:

Se define el cálculo del rendimiento agrícola como se muestra en la siguiente función:

Rendimiento agrícola=Superficie cosechada / Producción cosechada.

2.4 Descripción de los actores del sistema

Tabla 2: Actores del sistema

Actor	Descripción
Administrador	El administrador interactúa con el sistema para realizar las operaciones relacionadas con la administración de roles, usuarios y reportes.
Administrador ETL	El administrador de ETL gestiona los procesos de extracción, transformación y carga de los datos.
Especialista	El especialista interactúa con el sistema para analizar y consultar la información.

2.5 Necesidades de los usuarios

Luego de realizar un profundo estudio de la organización y del negocio sobre el área de agricultura, ganadería, silvicultura y pesca se definen las necesidades de los usuarios. Es de vital importancia que el resultado final esté en correspondencia con las necesidades identificadas inicialmente, pues ello va a definir que el sistema sea satisfactorio o no, respondiendo de igual manera a los intereses de los clientes. A continuación se describen las necesidades identificadas de acuerdo a los temas de análisis definidos:

- Distribución de la tierra del país y su utilización según formas de tenencia y tipos de empresas o entidades económicas en 31 de diciembre de 2007.
- Distribución de la tierra de acuerdo con su uso en las empresas y entidades estatales en diciembre 31.

- Distribución de la tierra de acuerdo con su uso en el sector no estatal en diciembre 31.
- Superficie cosechada, producción y rendimiento de la caña de azúcar por zafra, destino a industria.
- Superficie existente sembrada de cultivos permanentes seleccionados de la agricultura no cañera en diciembre 31.
- Superficie cosechada y en producción de cultivos seleccionados de la agricultura no cañera.
- Superficie cosechada y en producción de cultivos seleccionados de la agricultura no cañera, por sector estatal.
- Superficie cosechada y en producción de cultivos seleccionados de la agricultura no cañera, por sector no estatal.
- Producción agrícola por cultivos seleccionados de la agricultura no cañera.
- Producción agrícola por cultivos seleccionados de la agricultura no cañera, por sector estatal.
- Producción agrícola por cultivos seleccionados de la agricultura no cañera, por sector no estatal.
- Rendimiento agrícola por cultivos seleccionados de la agricultura no cañera.
- Rendimiento agrícola por cultivos seleccionados de la agricultura no cañera, por sector estatal.
- Rendimiento agrícola por cultivos seleccionados de la agricultura no cañera, por sector no estatal.
- Existencia de ganado vacuno, según sexo y categorías.
- Nacimientos y muertes del ganado vacuno.
- Indicadores seleccionados de la producción de leche de vaca.
- Entregas a sacrificio de ganado vacuno.
- Existencia, nacimientos y muertes del ganado porcino.
- Entregas a sacrificio de ganado porcino.
- Existencia total de aves en diciembre 31.
- Producción de huevos e indicadores seleccionados de gallinas ponedoras.
- Producción de carne de aves e indicadores seleccionados de pollos de ceba.
- Existencia de ganado équido, por sexos y formas de propiedad, en diciembre 31.
- Existencia de ganado ovino y caprino, en diciembre 31.
- Producción de leche y entrega a sacrificio de ganado ovino-caprino.
- Indicadores seleccionados de la apicultura.
- Plantaciones forestales realizadas.
- Otros indicadores seleccionados de la silvicultura.
- Captura por grupos de especies.

- Dinámica de la captura por grupos de especies.

2.5.1 Requisitos de Información

- RI1.** Obtener la distribución de la tierra por sector, año e indicadores de la agricultura.
- RI2.** Obtener el por ciento estructural de la tierra por sector, año e indicadores de la agricultura.
- RI3.** Obtener la cantidad de caña de azúcar por período de zafra, sector e indicadores de la zafra.
- RI4.** Obtener la superficie existente sembrada de cultivos permanentes por sector, año e indicadores de la agricultura no cañera.
- RI 5.** Obtener la superficie cosecha de cultivos seleccionados por sector, año e indicadores de la agricultura no cañera.
- RI 6.** Obtener la producción de cultivos seleccionados por sector, año e indicadores de la agricultura no cañera.
- RI 7.** Obtener el rendimiento agrícola por sector, año e indicadores de la agricultura no cañera.
- RI 8.** Obtener la existencia de ganado vacuno por año e indicadores del ganado vacuno.
- RI 9.** Obtener los nacimientos de ganado vacuno por sector, año.
- RI 10.** Obtener las muertes de ganado vacuno por sector, año.
- RI 11.** Obtener la producción de leche de vaca por sector, año.
- RI 12.** Obtener la existencia promedio de vacas en ordeño por sector, año.
- RI 13.** Obtener el rendimiento anual por vaca en ordeño por sector, año.
- RI 14.** Obtener las entregas a sacrificio de ganado vacuno por año e indicadores ganado vacuno.
- RI 15.** Obtener el ganado porcino por sector, año e indicadores ganado porcino.
- RI 16.** Obtener las entregas a sacrificio de ganado porcino por año e indicadores de sacrificio ganado porcino.
- RI 17.** Obtener la existencia total de aves por sector, año e indicadores avicultura.
- RI 18.** Obtener las gallinas ponedoras por sector, año e indicadores gallinas ponedoras.
- RI 19.** Obtener los pollos de ceba por sector, año e indicadores de los pollos de ceba.
- RI 20.** Obtener la existencia de ganado por sector, año e indicadores de ganadería.
- RI 21.** Obtener la existencia de ganado ovino por sector, año.
- RI 22.** Obtener la existencia de ganado caprino por sector, año.
- RI 23.** Obtener el ganado ovino-caprino por sector, año e indicadores ovino-caprino.
- RI 24.** Obtener valores de apicultura por sector, año e indicadores apicultura.
- RI 25.** Obtener las plantaciones forestales realizadas por año e indicadores silvicultura.
- RI 26.** Obtener las semillas procesadas por año.
- RI 27.** Obtener la producción de posturas por año.
- RI 28.** Obtener la reconstrucción de bosques por año.

RI 29. Obtener el mantenimiento de silvicultura por año.

RI 30. Obtener los tratamientos de silvicultura por año.

RI 31. Obtener las fajas verdes por año.

RI 32. Obtener las trochas corta fuegos por año.

RI 33. Obtener la captura por grupos de especies por año, indicadores de pesca.

RI34. Obtener la dinámica de la captura por grupos de especies por año, indicadores de pesca.

2.5.2 Requisitos Funcionales

Los RF son condiciones que debe cumplir el sistema, de acuerdo con las necesidades y especificaciones del cliente.

RF1. Autenticar usuario.

RF2. Adicionar roles.

RF3. Eliminar roles.

RF4. Adicionar usuarios.

RF5. Eliminar usuarios.

RF6. Insertar reportes.

RF7. Modificar reportes.

RF8. Eliminar reportes.

RF 9. Extraer información

RF 10. Realizar transformación y carga.

RF11. Abrir navegador OLAP.

RF12. Mostrar editor MDX.

RF13. Mostrar padres.

RF14. Ocultar repeticiones.

RF15. Intercambiar ejes.

RF16. Mostrar gráfico.

RF17 Configurar gráfico.

RF18. Configurar impresión.

RF19. Exportar a PDF.

RF20. Exportar a Excel.

RF21. Mostrar propiedades.

RF22. Suprimir filas.

RF23. Detallar miembros.

RF24. Entrar en detalles.

RF25. Mostrar datos de origen.

2.5.3 Requisitos No Funcionales

Los RNF son propiedades o cualidades que el producto debe tener. A continuación se exponen los requisitos no funcionales definidos:

Usabilidad

RNF 1. Cumplir con las pautas de diseño de las interfaces.

El sistema debe tener una interfaz gráfica uniforme que incluya pantallas, menús y opciones. Las pautas de diseño se realizarán siguiendo la arquitectura de información definida.

RNF 2. Mostrar los mensajes, títulos y demás textos que aparezcan en la interfaz del sistema en idioma español.

Los títulos de los componentes de la interfaz, los mensajes para interactuar con los usuarios y los mensajes de error, deben ser en idioma español y tener una apariencia uniforme en todo el sistema. Los mensajes de error deberán ser lo suficientemente informativos para dar a conocer la severidad del error.

RNF 3. Agilizar el acceso a los reportes del AD mediante la distribución de la información por áreas de análisis.

El usuario podrá acceder de manera rápida a la información que solicita en el área correspondiente de acuerdo al objetivo de su solicitud.

Confiabilidad

RNF 4. Asegurar la disponibilidad del sistema.

El sistema debe estar disponible durante el horario de trabajo. En caso de fallo, la recuperación del servicio no deberá exceder las 72 horas.

RNF 5. Asegurar la recuperación ante un fallo.

El sistema debe ser capaz de recuperarse ante un fallo, teniendo en cuenta la complejidad y naturaleza de este. El tiempo para su correcta recuperación fluctúa entre 10 minutos y 72 horas. Este tiempo comprende la solución al problema, así como su validación y prueba.

RNF 6. Garantizar la persistencia de la información.

Se debe realizar un respaldo total de los datos del AD con una frecuencia anual.

Eficiencia

RNF 7. Gestionar el tiempo promedio de respuesta para la obtención de un reporte.

El tiempo promedio para la obtención de un reporte es de aproximadamente 10 segundos.

Soporte

RNF 8. Lograr la homogeneidad de la estructura de los elementos definidos en el almacén.

Las estructuras del AD deben tener un nombre estándar teniendo en cuenta el tipo de estructura que sea. En la siguiente tabla se definen convenciones de nombrado con el objetivo de manejar un

vocabulario común en todo el AD, permitiendo un entendimiento claro y conciso por parte de los desarrolladores (ver tabla 3).

Tabla 3: Convenciones de nombrado

Estructura	Descripción	Ejemplo
Tablas de hechos	Todas las tablas de hechos tendrán una cadena que demuestra que son hechos y el concepto que describen.	hech_ <concepto>
Tablas de dimensiones	Todas las tablas de dimensiones tendrán una cadena que demuestra que son dimensiones y el concepto que describen.	dim_ <concepto>
Llaves primarias	Todas las llaves primarias tendrán una cadena que demuestra que son llaves primarias y el nombre de la tabla a la que pertenecen.	<tabla>_id
Atributos compuestos	En los atributos donde el nombre es compuesto se debe especificar el primer componente del atributo separado del segundo por un carácter de _.	<Primer nombre>_<Segundo nombre>

Restricciones de diseño

RNF 9. Utilizar el SGBD definido durante la investigación.

El SGBD que se utiliza es PostgreSQL 9.1 y como interfaz de administración de dicho gestor PgAdmin III.

RNF 10. Utilizar los lenguajes de programación definidos durante la investigación.

Como lenguaje dentro del SGBD para la programación en el AD se utiliza PL/pgSQL. En la implementación de los procesos de integración de datos se hace uso del lenguaje Java Script y MDX para realizar las consultas.

RNF 11. Utilizar la herramienta de integración de datos definida durante la investigación.

Para el proceso de integración de datos se usa la herramienta Pentaho Data Integration 4.0.1.

RNF12. Utilizar las herramientas para la implementación de la capa de BI definidas durante la investigación:

De la suite Pentaho, se usarán los siguientes componentes.

- Schema Workbench 3.2.0

- Pentaho BI Server
- Pentaho Administrator Console (en su traducción al inglés Consola de Administración del Pentaho)

Para el uso de las herramientas anteriores se requiere la instalación de la máquina virtual de java (Java Virtual Machine) como mínimo en su versión 6.0.

Requisitos para la documentación de usuarios en línea y ayuda del sistema

RNF 13. Confección de un manual de usuario.

El sistema debe estar acompañado de un documento que guiará la ejecución del usuario teniendo en cuenta cada funcionalidad.

Interfaz

RNF 14. Acceso al sistema.

El usuario deberá acceder a la aplicación mediante el protocolo HTTP (por sus siglas en inglés Hypertext Transfer Protocol), usando preferiblemente el navegador web Firefox 2.0 en adelante.

Interfaces de usuario

RNF 15. Garantizar una interfaz amigable al usuario.

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

Interfaces de hardware

RNF 16. Definir las interfaces de hardware que soportará el sistema.

El sistema podrá interactuar solamente con una interfaz de hardware: la impresora. Esta interacción ocurrirá cuando se necesite imprimir un reporte en formato físico. El acceso a la impresora será mediante el protocolo TCP/IP (por sus siglas en inglés Transmission Control Protocol /Internet Protocol), a través de la interfaz que ofrece el hardware.

RNF 17. Proporcionar características mínimas de hardware a los servidores.

Para lograr una explotación aceptable del sistema los servidores deben contar con los siguientes requerimientos de hardware:

- Windows server 2003.
- 1 GB RAM.
- 1 Microprocesador Core2Duo.

Interfaces de software

RNF 18. Instalar en las estaciones de trabajo los programas necesarios para el correcto funcionamiento del sistema.

Las configuraciones de software de las máquinas clientes deben contar al menos con

- Navegador web Firefox 2.0 o superior.
- Java Virtual Machine 6.0 y Schema Workbench 3.2.0 en caso de que un usuario capacitado requiera la construcción de esquemas multidimensionales para el diseño de nuevos reportes.

Requisitos legales, de derecho de autor y otros

RNF 19. Entregar el sistema a la ONEI.

El sistema debe ser transferido a la ONEI una vez que esté en explotación, incluyendo la documentación correspondiente.

2.6 Casos de Uso del Sistema

Para el diseño del diagrama de CUS se agruparon los 25 RF, los 34 de información en CU y se definieron las relaciones existentes entre ellos y los actores del sistema (ver figura 6).

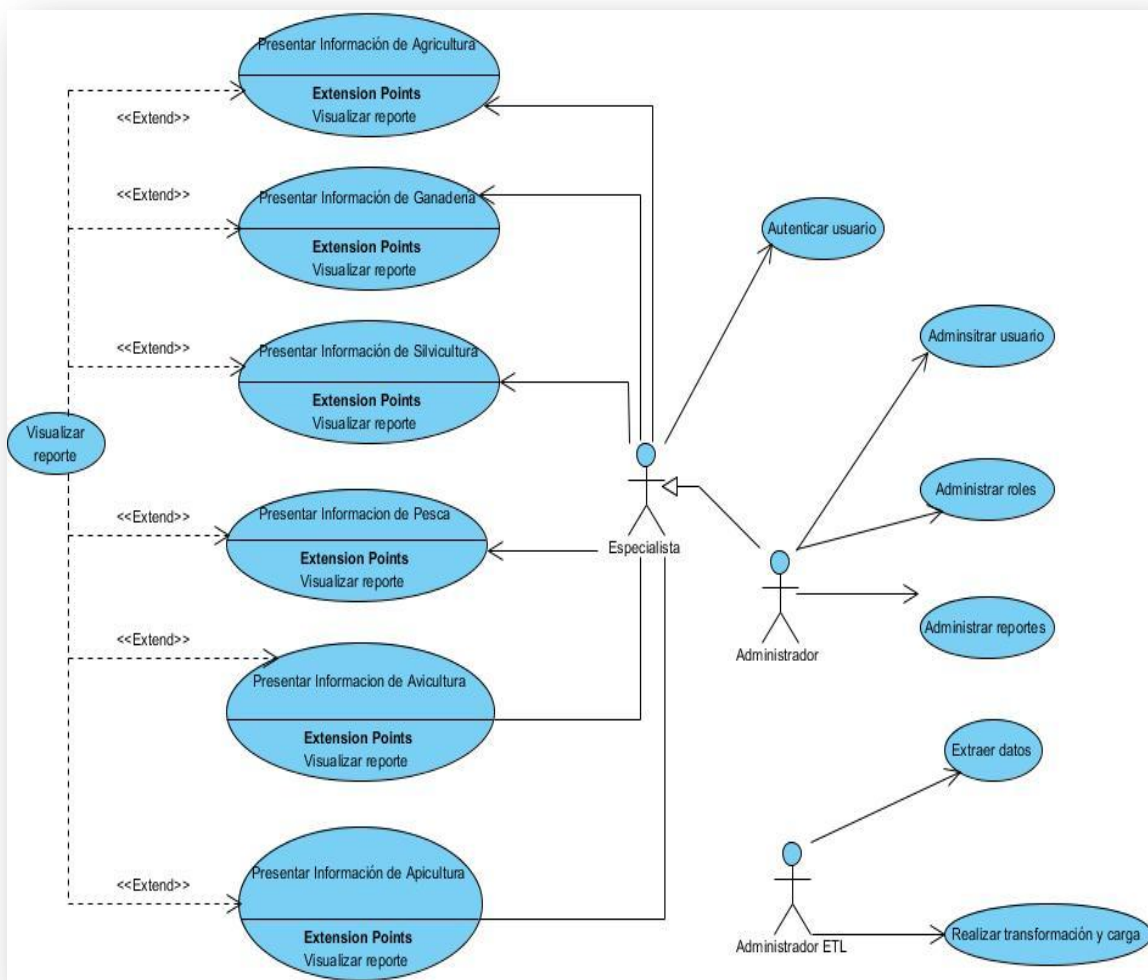


Figura 6: Diagrama de Casos de Uso del Sistema

2.6.1 Descripción de los Casos de Uso críticos

Para ver la totalidad de las descripciones de los CU referirse al artefacto “Especificación de Casos de Uso” del expediente de proyecto.

Extraer datos.

Tabla 4: Descripción Caso de Uso “Extraer datos”

Objetivo	Extraer los datos del sistema.	
Actores	Administrador ETL	
Resumen	El caso de uso inicia cuando el actor selecciona los datos a extraer. Se extraen los mismos de la fuente. Finaliza cuando los datos se encuentran en el área temporal.	
Complejidad	Media	
Prioridad	Crítico	
Precondiciones	Disponibilidad de las fuentes.	
Postcondiciones	Los datos del fichero ‘Excel’ correspondientes han sido extraídos de la fuente y almacenados en un área temporal.	
Flujo de eventos		
Flujo básico: Extraer datos.		
	Acción del Actor	Respuesta del Sistema
1.	El administrador de ETL realiza la Conexión al ‘Excel’ correspondiente.	
2.		Responde a la solicitud de conexión.
3.	El administrador de ETL selecciona la estructura o archivo a extraer.	
4.	El administrador de ETL realiza la extracción de los datos.	
5.		Ejecuta la extracción de los datos. Finaliza el caso de uso.
Flujos alternos		
2. No responde a la solicitud de conexión.		
	Actor	Sistema
2.1		No responde a la solicitud de conexión.

2.2		Notifica el error al administrador de ETL. Vuelve al paso 1 del flujo normal de eventos.
3.1.	Si hay control de cambios, el administrador de ETL verifica si hay modificaciones. <ul style="list-style-type: none"> • En caso afirmativo ir al paso 3 del flujo normal de eventos. • En caso negativo ir al paso 2 del flujo normal de eventos. 	
Relaciones	CU Incluidos	No aplica
	CU Extendidos	No aplica
Requisitos no funcionales		
Asuntos pendientes		

Realizar transformación y carga de los datos.

Tabla 5: Descripción Caso de Uso "Realizar transformación y carga de los datos"

Objetivo	Transformar y cargar los datos en el sistema.
Actores	Administrador ETL
Resumen	El caso de uso inicia cuando el administrador de ETL desea realizar la transformación y carga de los datos. El actor selecciona la fuente de información deseada, realiza las transformaciones convenientes y carga la información resultante en el MD, finalizando así el caso de uso.
Complejidad	Media
Prioridad	Crítico
Precondiciones	Los datos se encontraron correctamente extraídos en el área temporal y las estructuras del MD se encontraron disponibles para su uso.
Postcondiciones	Los datos del fichero 'Excel' correspondiente han sido transformados y cargados en el MD.
Flujo de eventos	
Flujo básico: Realizar Transformación y Carga de los datos	

	Actor	Sistema
1	El administrador de ETL selecciona las estructuras del área temporal que desea transformar.	
2	El administrador de ETL carga los datos seleccionados en memoria.	
3	El administrador de ETL aplica las transformaciones pertinentes y genera datos de auditoría.	
4	El administrador de ETL carga los datos en el MD.	
		5 Ejecuta la consulta. Finaliza el caso de uso.
Flujos alternos		
3. Si las transformaciones no pueden ser aplicadas.		
	Actor	Sistema
		3.1 El sistema muestra un mensaje de error y regresa al paso 2 del flujo normal de eventos. Finaliza el caso de uso.
Relaciones	CU Incluidos	No aplica
	CU Extendidos	No aplica
Requisitos no funcionales		
Asuntos pendientes		

2.7 Arquitectura

Para definir la arquitectura de un MD se debe tener en cuenta la forma de representar el origen de los datos, la comunicación, los procesos y la presentación final de la información al usuario. La arquitectura propuesta para la presente solución consta de cuatro niveles (ver figura 7):

- **Fuentes de datos:** se refiere al origen de los datos.
- **Subsistema de integración:** incluye los procesos que permiten que los datos de las fuentes sean extraídos, transformados y cargados hacia las fuentes destino.
- **Subsistema de almacenamiento:** es una BD relacional que contiene las tablas de dimensiones y hechos cargadas a través de los procesos de ETL.

- **Subsistema de visualización:** comprende las interfaces orientadas a usuarios que extraen información, facilitándoles la toma de decisiones.
- El subsistema de integración, se abastece de las diferentes fuentes de datos y se encarga de llevar a cabo los procesos que integran y transforman la información para su carga. Los usuarios que acceden a este subsistema son los encargados de la administración de dichos procesos.
- El subsistema de almacenamiento recibe la información manipulada durante la extracción, transformación y carga, en una BD soportada por el SGBD PostgreSQL 9.1 y administrada por los usuarios autorizados mediante la herramienta pgAdmin III.
- EL subsistema de visualización permite mostrar a los usuarios autorizados la información estandarizada en forma de reportes a través de la herramienta Pentaho Business Intelligence.

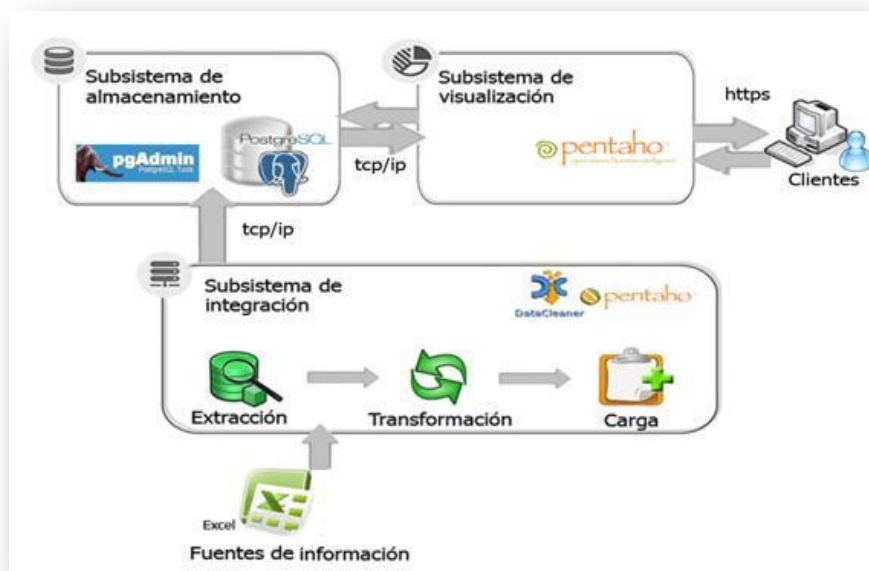


Figura 7: Arquitectura propuesta para la solución

2.8 Diseño del subsistema de almacenamiento

El proceso de diseño de un AD define el modelo conceptual, lógico, físico y de visualización del mismo (ver figura 8). En el conceptual se especifican los requerimientos del usuario mediante las definiciones de dimensiones, hechos y medidas que conforman el AD; en el modelo lógico se genera el modelo de datos; en el modelo físico se diseñan las transformaciones y en el de visualización se diseñan reportes candidatos por cada uno de los Libros de Trabajo (LT) definidos en el área de análisis.

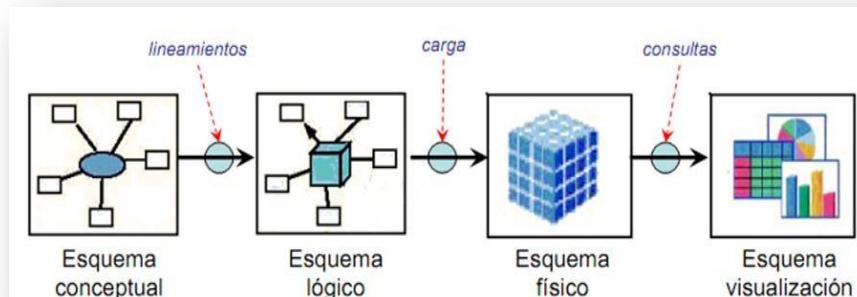


Figura 8: Diseño de un Almacén de Datos

2.8.1 Dimensiones

Las dimensiones son las características de un hecho, que permiten su posterior análisis en el proceso de toma de decisiones.

Las dimensiones definidas para el presente trabajo son las siguientes:

- dim_indicadores_pesca
- dim_indicadores_silvicultura
- dim_indicadores_otros_silvicultura
- dim_indicadores_ganaderia
- dim_indicadores_apicultura
- dim_indicadores_avicultura
- dim_indicadores_pollos_ceba
- dim_indicadores_gallinas_ponedoras
- dim_sector
- dim_indicadores_agricultura_no_cannera
- dim_indicadores_tierra
- dim_periodo_zafra
- dim_indicadores_cultivos_permanentes
- dim_categoria_zafra
- dim_indicadores_ganaderia_otros
- dim_temporal_anno

2.8.2 Hechos y medidas

Las tablas de hechos contienen las dimensiones y las medidas asociadas a estos. Las medidas son valores de datos numéricos que serán analizadas por los usuarios, son las variables de salida en el diseño, de ahí que representen lo contable que se necesita conocer, con un alto valor informativo. En el presente MD se definieron los siguientes hechos y medidas:

- hech_pesca (cantidad_captura,cantidad_dinamica_captura)
- hech_silvicultura (cantidad_plantaciones)
- hech_otros_silvicultura (cantidad)
- hech_ganaderia (cantidad_ganado)
- hech_apicultura (valor)
- hech_pollos_ceba (valor)
- hech_gallinas_ponedoras (valor)
- hech_avicultura (cantidad)
- hech_otros_agricultura (producción,superficie_cosechada)
- hech_tierra (estructura, distribución)
- hech_zafra (cantidad_zafra)
- hech_cultivos_permanentes (cantidad)
- hech_ganaderia_otros (cantidad)

2.8.3 Matriz bus

A continuación se muestra la matriz bus (ver figura 9), esta representa la relación existente entre las dimensiones y los hechos del MD Series históricas de agricultura, ganadería, silvicultura y pesca, evitando que exista solapamiento entre los hechos y posibilitando obtener un esquema más seguro.

HECHOS / DIMENSIONES	dim_temporal_año	dim_indicadores_pesca	dim_indicadores_silvicultura	dim_indicadores_otros_silvicultura	dim_indicadores_ganaderia	dim_indicadores_apicultura	dim_indicadores_pollos_ceba	dim_indicadores_gallinas_ponedoras	dim_indicadores_avicultura	dim_indicadores_agricultura_no_camera	dim_indicadores_agricultura_tierra	dim_cultivos_permanentes	dim_categoria_zafra
hech_pesca	x	x											
hech_silvicultura	x		x										
hech_otros_silvicultura	x			x									
hech_ganaderia	x				x			x					
hech_apicultura	x					x		x					
hech_pollos_ceba	x						x	x					
hech_gallinas_ponedoras	x							x	x				
hech_avicultura	x								x	x			
hech_otros_agricultura	x									x			
hech_tierra	x										x		
hech_zafra												x	x
hech_cultivos_permanentes	x											x	
hech_ganaderia_otros	x												x

Figura 9: Matriz bus

2.9 Diseño del subsistema de integración

El diseño del subsistema de integración lleva implícito un elemento fundamental, el diseño de las transformaciones o procesos de integración, que al ser ejecutadas permiten que los datos se encuentren disponibles en una tabla de salida. A continuación se muestra el diseño general que se tuvo en cuenta para implementar las diferentes transformaciones de los indicadores del MD Series históricas de agricultura, ganadería, silvicultura y pesca (ver figura 11).

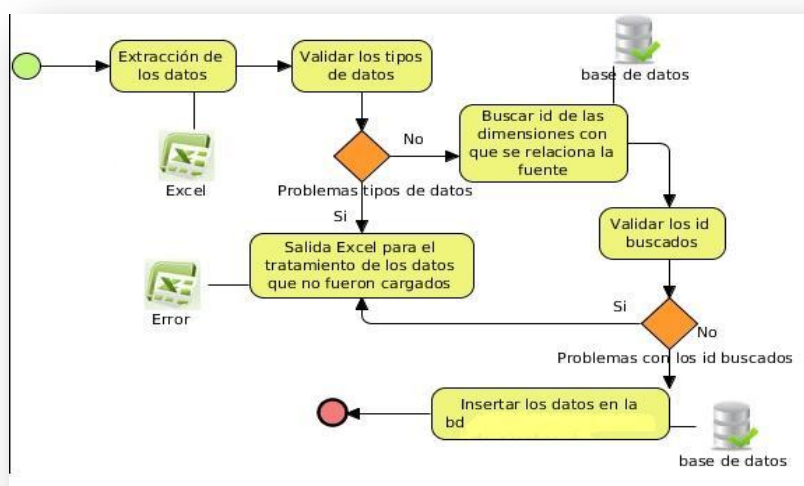


Figura 11: Diseño de las transformaciones

2.10 Diseño del subsistema de visualización

Los reportes candidatos responden a las necesidades de los usuarios, permitiendo identificar los reportes y demás salidas que debe producir el sistema, estos son implementados en la capa de BI. En el presente trabajo se diseñaron 31 reportes candidatos, agrupados en seis LT. A continuación se muestra un ejemplo del reporte candidato: “Obtener información de agricultura” (ver tabla 6). Para ver la totalidad de los reportes candidatos referirse al artefacto “Reportes candidatos” del expediente de proyecto.

Tabla 6: Reporte candidato: “Obtener información de agricultura”

Área de análisis (AA)	A.A Series agricultura
Libro de Trabajo (LT)	LT Agricultura
Reporte (Tabla de Salida – TS)	9.1- Distribución de la tierra del país y su utilización según formas de tenencia y tipos de empresas o entidades económicas en 31 de diciembre.

Descripción	En el reporte se visualizan totales relacionados con cultivos y con superficie, ambas clasificadas según el tipo de sector: estatal o no estatal (UBPC, CPA, CCS y privadas).
Elementos del reporte	<ul style="list-style-type: none"> • Sector (estatal y no estatal). • Año • Indicadores tierra
Frecuencia de emisión	Anual

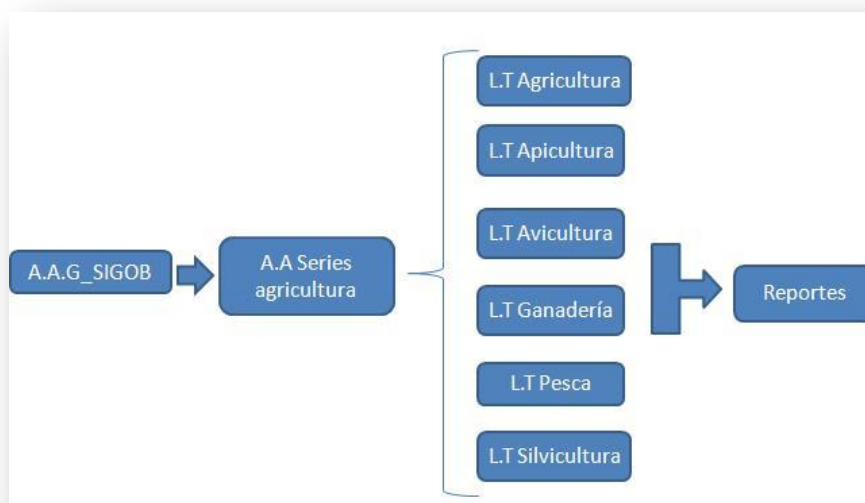


Figura 12: Arquitectura de información del Mercado de Datos

2.11 Políticas de seguridad

Para el acceso al MD se define un usuario por cada uno de los roles existentes en el sistema. Se decide la utilización del patrón RBAC (por sus siglas en inglés Role Based Access Control), lo que garantizará que cada usuario opere en el sistema según los permisos que se le definan al rol (ver tabla 7).

Tabla 7: Permisos y roles

Roles	Permisos	
	Lectura	Escritura
Sobre la Base de Datos		

Administrador	x	
Administrador ETL	x	x
Analista	x	
Sobre la aplicación		
Administrador	x	x
Analista	x	

2.11.1 Salva de la Base de Datos

Con el objetivo de garantizar la persistencia de la información se debe realizar un respaldo total de los datos del MD Series históricas de agricultura, ganadería, silvicultura y pesca con frecuencia anual. Dicha información será almacenada en el área de la dirección de informática en un banco de datos especial, encontrándose localizada en el edificio de la ONEI. Su seguridad y uso será responsabilidad del grupo de administración de redes de la ONEI.

Conclusiones

En el presente capítulo se realizó el análisis y diseño del MD Series históricas de agricultura, ganadería, silvicultura y pesca, obteniendo los siguientes resultados:

- Se definieron dos reglas del negocio.
- Se realizó el diseño del diagrama de CUS, obteniéndose seis CUI y siete CU funcionales, posibilitando que el sistema cumpla con las funciones requeridas.
- Se diseñó el modelo de datos identificando 16 tablas dimensionales y 13 tablas de hechos, garantizando el correcto funcionamiento del sistema.
- Se definieron las políticas de respaldo y recuperación de la información, garantizando su conservación y constancia.
- Se diseñaron las transformaciones, con el fin de poblar la BD.
- Se diseñaron los reportes candidatos y los cubos multidimensionales en correspondencia con cada una de las tablas de hechos del MD y con las necesidades de información.

CAPÍTULO 3: IMPLEMENTACIÓN DEL MERCADO DE DATOS

Introducción

En este capítulo se implementan los procesos definidos en el diseño, como el subsistema de almacenamiento, definiendo la estructura de los datos, esquemas y tablas de la BD, así como el subsistema de integración, a través del trabajo y los flujos de transformación. Se realiza la implementación del subsistema de visualización de los datos, mediante la estructura de navegación y los reportes candidatos.

3.1 Implementación del subsistema de almacenamiento

3.1.2 Implementación del modelo de datos

Para el desarrollo del sistema propuesto en la investigación se definieron los siguientes esquemas: el esquema *dimensiones*, que recoge todas las tablas de dimensiones que son comunes para el almacén de datos central, el esquema *mart_agricultura_series* que recoge las tablas de dimensiones y hechos propios del MD y el esquema *metadatos*, que incluye las tablas *md_jobs* y *md_transformation*. La solución cuenta con 16 tablas de dimensiones, 13 tablas de hechos y dos tablas para los metadatos, mostradas a continuación (ver tabla 8):

Tabla 8: Esquemas y tablas

Esquemas	Tablas
dimensiones	dim_tempral_anno
mart_agricultura_series	dim_indicadores_pesca
mart_agricultura_series	dim_indicadores_silvicultura
mart_agricultura_series	dim_indicadores_otros_silvicultura
mart_agricultura_series	dim_indicadores_ganaderia
mart_agricultura_series	dim_indicadores_apicultura
mart_agricultura_series	dim_indicadores_avicultura
mart_agricultura_series	dim_indicadores_pollos_ceba
mart_agricultura_series	dim_indicadores_gallinas_ponedoras
mart_agricultura_series	dim_sector
mart_agricultura_series	dim_indicadores_agricultura_no_cannera
mart_agricultura_series	dim_indicadores_tierra
mart_agricultura_series	dim_periodo_zafra
mart_agricultura_series	dim_indicadores_cultivos_permanentes
mart_agricultura_series	dim_categoria_zafra
mart_agricultura_series	dim_indicadores_ganaderia_otros

mart_agricultura_series	hech_pesca
mart_agricultura_series	hech_silvicultura
mart_agricultura_series	hech_otros_silvicultura
mart_agricultura_series	hech_ganaderia
mart_agricultura_series	hech_apicultura
mart_agricultura_series	hech_pollos_ceba
mart_agricultura_series	hech_gallinas_ponedoras
mart_agricultura_series	hech_avicultura
mart_agricultura_series	hech_otros_agricultura
mart_agricultura_series	hech_tierra
mart_agricultura_series	hech_zafra
mart_agricultura_series	hech_cultivos_permanentes
mart_agricultura_series	hech_ganaderia_otros
metadatos	md_jobs
metadatos	md_transformation

3.2 Implementación del subsistema de integración

3.2.1 Implementación de los procesos de Extracción, Transformación y Carga

El proceso de ETL se inicia al extraer los datos desde los sistemas fuentes. El formato de dichos datos se encuentra en ficheros de tipo “Excel”, separados por modelos en distintos directorios. Luego de haber realizado la extracción de los datos se procede a realizar las transformaciones pertinentes, estas constituyen un elemento esencial dentro de la implementación del proceso ETL. En este paso la información es validada y adaptada al modelo de datos desarrollado previamente.

El último de los subprocesos de ETL realizados para el desarrollo de la solución, es la carga de la información deseada, este proceso consiste en migrar los datos que han sido transformados para posteriormente realizar la implementación de la capa de visualización. La siguiente figura muestra un ejemplo de una de las transformaciones desarrolladas para poblar el MD Series históricas de agricultura, ganadería, silvicultura pesca (ver figura 13).

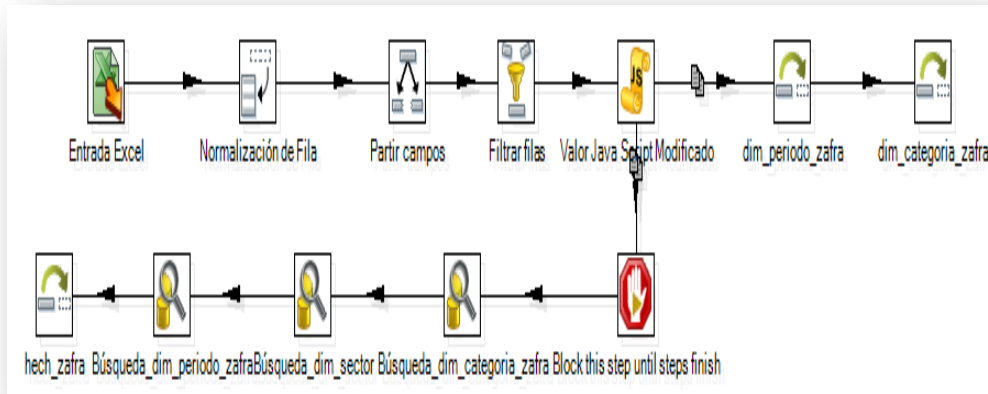


Figura 13: Transformación correspondiente al hecho zafra

3.2.2 Implementación del trabajo

Luego de concluida la realización de todas las transformaciones necesarias para la carga de los datos, se realiza la implementación del trabajo, este permitirá ejecutar todas las transformaciones que han sido diseñadas anteriormente en orden lógico, primero las dimensiones y luego los hechos. Al ejecutar el siguiente trabajo se ejecutan cada una de las transformaciones definidas en el orden anteriormente descrito (ver figura 14).

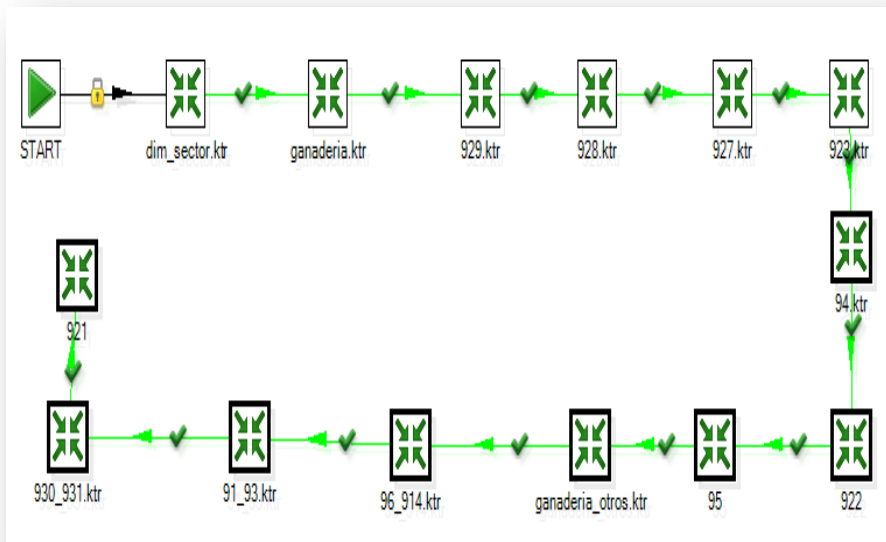


Figura 14: Trabajo para cargar los hechos y dimensiones del Mercado de Datos

3.3 Implementación del subsistema de visualización

3.3.1 Implementación de los cubos OLAP

La implementación de los cubos OLAP se realiza utilizando la herramienta Pentaho Schema Workbench 3.2.0. Esta permite generar un fichero de configuración “.XML”, en el cual se definen los cubos y las dimensiones con sus niveles de jerarquía, así como la conexión con el MD que contiene los datos para el cubo multidimensional.

En el presente trabajo se modelaron 13 cubos, con las características correspondientes a cada una de las tablas de hechos y dimensiones definidas. A continuación se muestra un ejemplo del diseño de los cubos (ver figura 15):

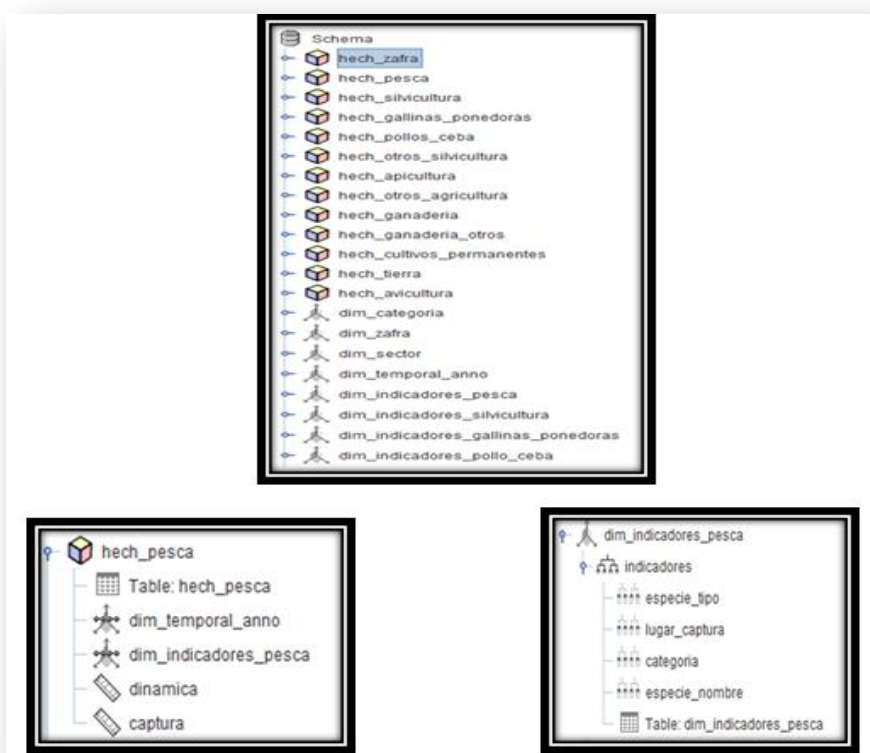


Figura 15: Cubo OLAP

Arquitectura de información

A continuación se detallan los elementos que conforman la estructura de navegación de la información que se presenta en la capa de visualización del MD (ver figura 16):



Figura 16: Arquitectura de la información del Mercado de Datos

3.3.2 Implementación de los reportes

Los reportes candidatos o tablas de salidas como también se les conoce son aquellos que contienen los valores asociados a los indicadores de interés para el cliente. Dichos reportes fueron seleccionados luego de realizar un análisis de las series que recogen toda la información referente al área de agricultura, ganadería, silvicultura y pesca de la ONEI. La implementación de los reportes fue a través de consultas MDX, que constituyen en los sistemas OLAP el equivalente a las consultas SQL en las BD relacionales. A continuación se muestra una vista de análisis implementada a través de consultas MDX (ver figura 17).

Indicadores	Dinámica de la captura									
	Año									
	● 2002	● 2003	● 2004	● 2005	● 2006	● 2007	● 2008	● 2009	● 2010	
Todos	59625.00	1948.41	2086.56	2088.33	2308.47	2147.19	1818.24	2397.15	1679.89	
▣ Pescado	42909.40	1386.15	1194.48	920.53	1630.85	1477.82	1221.06	1368.79	962.74	
▣ Cobo	997.10	37.34	148.27	116.67	50.67	176.03	69.79	133.18	94.68	
▣ Ostión	1300.20	101.02	90.11	88.38	106.03	97.94	79.31	159.58	119.90	
▣ Almeja	374.60	77.36	152.76	91.03	92.73	102.89	24.76	310.92	102.64	
▣ Langosta	7969.50	66.06	144.38	76.75	75.45	108.55	119.84	72.03	108.08	
▣ Camarón de mar	1308.00	114.62	100.29	109.03	95.91	30.60	114.19	112.32	124.00	
▣ Camaronicultura	1910.40	71.31	27.06	640.23	184.18	82.69	102.86	93.47	87.54	
▣ Otras Especies (incluye Morralla)	2855.80	98.55	229.21	45.71	72.65	70.67	86.43	146.86	80.31	

Figura 17: Ejemplo de reporte. Libro de Trabajo "Pesca"

Conclusiones

Después de realizar la implementación del MD Series históricas de agricultura, ganadería, silvicultura y pesca se arribaron a las siguientes conclusiones:

- Quedó implementado el subsistema de almacenamiento, partiendo del modelo de datos físico, el cual cuenta con tres esquemas: *dimensiones*, compuesto por una tabla de dimensiones, *mart_agricultura_series* que contiene 31 tablas, correspondientes a los hechos y dimensiones propios del MD y el esquema *metadatos* que incluye las tablas *md_jobs* y *md_transformation*.
- Se implementó el subsistema de integración mediante la realización del trabajo y las transformaciones necesarias para extraer, transformar y cargar los datos provenientes de la fuente hacia el MD.
- Se desarrolló el subsistema de visualización, implementando los cubos OLAP, quedando definidos 13 cubos, 16 dimensiones y una medida calculable, determinándose seis LT para un total de 31 vistas de análisis.

CAPÍTULO 4: VALIDACIÓN DEL MERCADO DE DATOS

Introducción

En este capítulo se valida la solución, mediante la utilización de las listas de chequeo, los casos de prueba y la carta de aceptación del cliente.

4.1 Pruebas

Las pruebas de software son procesos que permiten verificar y revelar la calidad de un determinado producto. Son utilizadas para identificar posibles fallos de implementación, calidad y usabilidad en la solución [16]. Para determinar el nivel de calidad se deben efectuar pruebas que permitan comprobar el grado de cumplimiento de los requisitos iniciales del sistema.

Las experiencias alcanzadas en el desarrollo de software han demostrado que un fallo puede representar elevados costos, lo que ha hecho posible que empresas y organizaciones inviertan en verificar minuciosamente el resultado de cada prueba, para así poder descubrir posibles síntomas de defectos antes de dar por concluido un determinado producto. Para realizar la comprobación de la calidad en la solución "MD Series históricas de agricultura, ganadería, silvicultura y pesca" se utilizó el Modelo V, definido por el Centro de Calidad para Aplicaciones Tecnológicas (CALISOFT), este es utilizado en DATEC con el fin de crear un estándar de verificación para que los productos cumplan con las especificaciones del negocio. En la siguiente imagen se muestra como el Modelo V relaciona las actividades de prueba con el análisis y el diseño (ver figura 18).

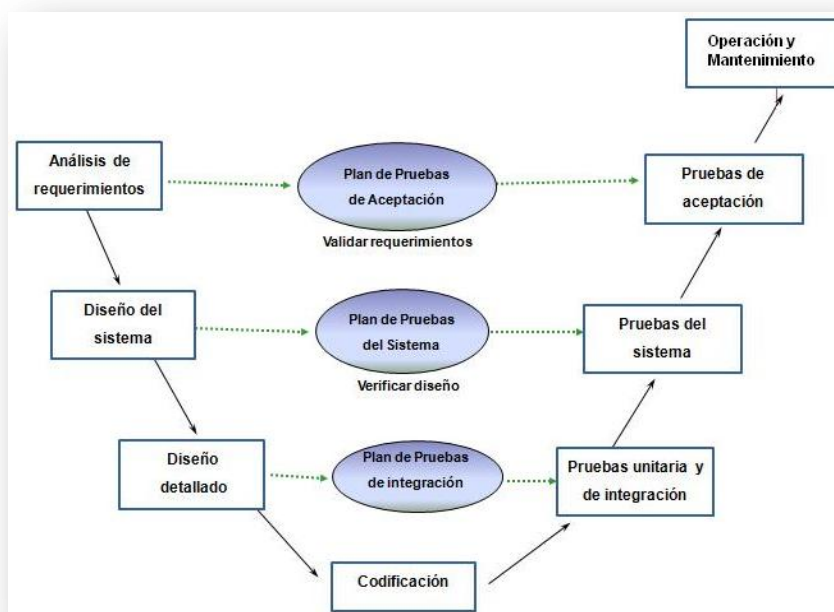


Figura 18: Modelo V

El Modelo V es una representación de dos cascadas enfrentadas y relacionadas, con su vértice en la codificación como punto en común. Propone una cascada a la izquierda, con las actividades relacionadas al desarrollo y una a la derecha con las actividades del aseguramiento de la calidad del software. Mediante este modelo se describe a un nivel muy alto de abstracción las fases del ciclo de desarrollo en las que se involucra la prueba. Para la validación del MD Series históricas de agricultura, ganadería, silvicultura y pesca se aplicaron diferentes tipos de pruebas, especificadas a continuación:

Pruebas unitaria: es el proceso de probar los componentes individuales de la solución. El propósito es identificar diferencias entre la especificación de los artefactos y el comportamiento real de cada módulo.

Pruebas de integración: es el proceso en el cual los componentes son agregados para crear otros más grandes. Es la prueba realizada para mostrar que aunque los componentes hayan pasado satisfactoriamente las pruebas de unidad, la integración de los componentes es incorrecta.

Pruebas de sistema: se refiere al comportamiento del sistema de manera integrada. Durante la etapa de las pruebas unitarias y de integración deben haberse identificado la mayoría de las No Conformidades (NC) presentes en la aplicación. La prueba de sistema es aplicada generalmente para probar los RNF de la solución.

Pruebas de aceptación: se realizan para probar que el sistema cumpla en su totalidad con los requerimientos especificados por el cliente.

4.2 Herramientas para validar el Mercado de Datos

4.2.1 Listas de chequeo

Para realizar las pruebas al MD Series históricas de agricultura, ganadería, silvicultura y pesca se aplicaron cuatro listas de chequeo a los artefactos correspondientes a los procesos de ETL.

Listas de chequeo: constituyen un mecanismo para el control de los riesgos, tienen como función básica detectar condiciones peligrosas que puedan atentar contra la calidad del producto.

Para elaborar la lista de chequeo se tuvieron en cuenta elementos de evaluación que son importantes una vez realizado el proceso de ETL, permitiendo recoger los puntos eficientes e ineficientes que posea dicho proceso. La lista de chequeo contiene diferentes indicadores a evaluar los cuales se encuentran distribuidos en tres secciones fundamentales (ver tabla 9):

- Estructura del documento: abarca todos los aspectos definidos por el expediente de proyecto o el formato establecido por el proyecto.
- Indicadores definidos: abarca todos los indicadores a evaluar durante la etapa de desarrollo del mercado.
- Semántica del documento: contempla todos los indicadores a evaluar respecto a la ortografía, redacción y demás.

Tabla 9: Lista de chequeo diseñada para validar el MD

Estructura del documento					
Peso	Indicadores a evaluar	Eval	(NP)	Cantidad de elementos afectados	Comentarios
crítico	1. ¿Los documentos están acorde a los estándares definidos por el proyecto?				
crítico	2. ¿Contiene las secciones obligatorias definidas en el expediente de proyecto?				
Elementos definidos por el modelo de desarrollo					
Peso	Indicadores a evaluar	Eval	(NP)	Cantidad de elementos afectados	Comentarios
crítico	1. ¿Se realizó un estudio preliminar de la entidad cliente?				
crítico	2. ¿Se identificaron los requisitos de información y las reglas del negocio?				
crítico	3. ¿Se realizó el diseño del modelo de datos correspondiente al mercado de datos Series históricas de agricultura, ganadería, silvicultura y pesca en conjunto con el cliente?				
crítico	4. ¿Se realizó el proceso de ETL correspondiente al mercado de datos Series históricas de agricultura, ganadería, silvicultura y pesca?				

crítico	5. ¿Se realizó la implementación de él/los trabajo(s) para el mercado de datos Series históricas de agricultura, ganadería, silvicultura y pesca?				
crítico	6. ¿Los trabajos se pueden ejecutar desde cualquier ordenador?				
crítico	7. ¿Se le dan tratamiento a los errores que ocurren durante el proceso de ETL?				
crítico	8. ¿Se realizó el proceso de Inteligencia de Negocio (BI) correspondiente al mercado de datos Series históricas de agricultura, ganadería, silvicultura y pesca?				
crítico	9. ¿Los reportes que se muestran en la capa de visualización se corresponden con las necesidades del negocio identificadas?				
	10. ¿La aplicación realizada apoya el proceso de toma de decisiones para el área Series históricas de agricultura, ganadería, silvicultura y pesca?				
crítico	11. ¿Se realizaron los diseños de los casos de prueba?				
crítico	12. ¿Se realizaron las pruebas de aceptación?				
crítico	13. ¿Se realizó el despliegue de la aplicación?				
crítico	14. ¿Se le da soporte y mantenimiento a la aplicación?				

Semántica del documento					
Peso	Indicadores a evaluar	Eval	(NP)	Cantidad de elementos afectados	Comentarios
crítico	1. ¿Se han identificado errores ortográficos en los entregables?				
crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?				
	3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?				

4.2.2 Casos de prueba

Para realizar las pruebas al MD Series históricas de agricultura, ganadería, silvicultura y pesca se aplicaron seis casos de prueba, para ver la totalidad de los casos de pruebas referirse al artefacto "Casos de prueba" del expediente de proyecto (ver figura 19 y figura 20).

Casos de prueba: tienen como propósito validar los CUI definidos en el diagrama de CUS, aunque pueden existir mayor cantidad de casos de prueba en dependencia de la complejidad de los CUI que hayan sido identificados. La aplicación de estas pruebas permite demostrar que las vistas de análisis definidas en el MD satisfacen los requisitos de información identificados inicialmente, garantizando así el cumplimiento de uno de los principales objetivos del sistema.

Escenario	Descripción	VARIABLES DE ENTRADA	VARIABLES DE SALIDA	Respuesta del sistema	Flujo central
EC 1.1 9.15-Existencia de ganado vacuno según sexo y categorías	Permite visualizar los reportes con las variables presentes en los mismos.	Año	Cantidad	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. Se autentica. Se entra al sistema. Se selecciona la el área de análisis (AA) Agricultura. Se selecciona el libro de trabajo (LT) Prueba. Se selecciona el reporte: " 9.15-Existencia de ganado vacuno según sexo y categorías"
		Otros indicadores de ganadería			
EC 1.2 9.16-Nacimientos y muertes del ganado vacuno	Permite visualizar los reportes con las variables presentes en los mismos.	Año	Cantidad	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. Se autentica. Se entra al sistema. Se selecciona el área de análisis (AA) Agricultura. Se selecciona el libro de trabajo (LT) Prueba. Se selecciona el reporte: "9.16-Nacimientos y muertes del ganado vacuno".
		Indicadores ganadería			
		Sector			

Figura 19: Caso de prueba CUI "Presentar información de ganadería"

No	Nombre de campo	Clasificación	Valor Nulo	Descripción
1	Año	Campo de texto	No	Muestra los años por lo que se presenta la información del reporte.
2	Sector	Lista desplegable	No	Muestra el tipo de sector por el que se presenta la información del reporte(estatal y no estatal).
3	Indicadores ganadería	Lista desplegable	No	Muestra indicadores de ganadería.
4	Otros indicadores de ganadería	Lista desplegable	No	Muestra información referente a otros indicadores de ganadería.

Figura 20: Variables Caso de prueba CUI "Presentar información de ganadería"

4.3 Evaluación de los resultados

Aplicadas las pruebas descritas anteriormente al MD Series históricas de agricultura, ganadería, silvicultura y pesca, se obtuvieron los siguientes resultados:

- La aplicación de las listas de chequeo permitió detectar dos NC en la estructura del documento, ninguna no conformidad en elementos definidos por la metodología y dos NC en la semántica del documento, detectadas de forma general en la revisión de los artefactos, todas resueltas satisfactoriamente (ver figura 21).

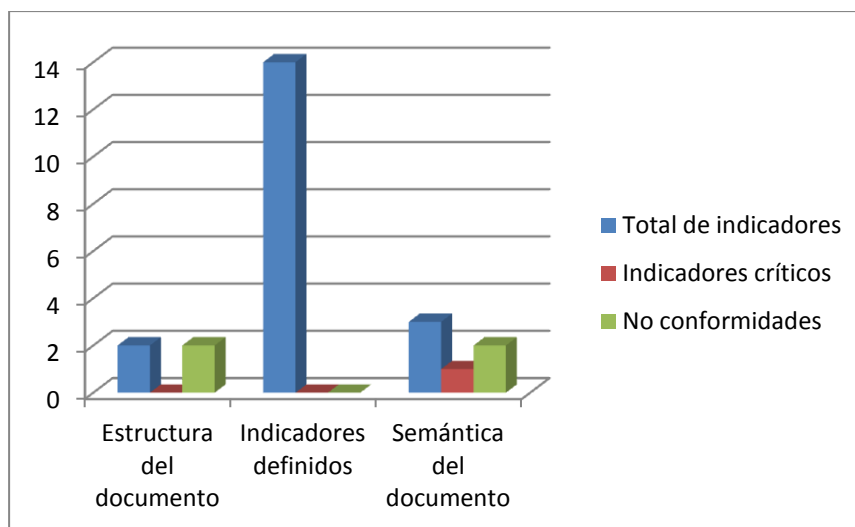


Figura: 21 Resultado de la aplicación de las listas de chequeo

- Mediante la aplicación de los seis casos de prueba se detectaron 12 NC, seis de alta prioridad, cuatro de prioridad media y dos de baja prioridad, todas resueltas en su totalidad (ver figura 22).

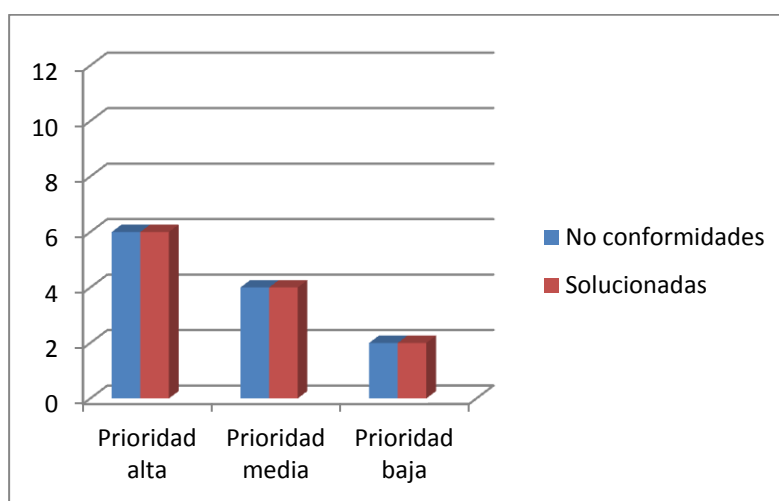


Figura: 22 Resultado de la aplicación de los casos de prueba

Una vez finalizado el desarrollo del mercado, se realizó la prueba de aceptación por la especialista Elena Leonila Fernández García (representante de la ONEI en la universidad), quien confirmó que el MD Series históricas de agricultura, ganadería, silvicultura y pesca satisface todas las necesidades de información identificadas con anterioridad. Dichos resultados quedaron oficializados con la carta de aceptación del cliente.

Conclusiones

En el capítulo se realizó la validación de la propuesta del MD para el área agricultura, ganadería, silvicultura y pesca donde:

- Se aplicaron seis casos de prueba para validar la solución propuesta, detectándose 12 NC, seis de alta prioridad, cuatro de prioridad media y dos de baja prioridad.
- Se realizaron las listas de chequeo al MD Series históricas de agricultura, ganadería, silvicultura y pesca, detectándose dos NC en la estructura del documento, ninguna no conformidad en elementos definidos por la metodología y dos NC en la semántica del documento, detectadas de forma general en la revisión de artefactos correspondientes a los procesos de ETL, todas resueltas satisfactoriamente.

CONCLUSIONES

El estudio de las etapas de desarrollo de un MD, proporcionó la elaboración del presente trabajo, cuyo resultado fue el “MD Series históricas de agricultura, ganadería, silvicultura y pesca”, el cual contribuyó a la integración de los datos de la institución y arrojó los siguientes resultados:

- La fundamentación de las metodologías, herramientas y tecnologías utilizadas en la investigación, permitió un correcto diseño e implementación de la solución presentada así como una organización estructurada en el proceso de desarrollo del MD Series históricas de agricultura, ganadería, silvicultura y pesca.
- El análisis permitió la identificación de las necesidades del cliente y contribuyó a la obtención del diagrama de CUS. El diseño realizado tributó a la construcción de los subsistemas que componen el MD, logrando una adecuada estructuración de la información en función de apoyar los análisis sobre el área de agricultura, ganadería, silvicultura y pesca.
- La implementación de los subsistemas que componen la solución contribuyó a la creación de tres esquemas en el SGBD PostgreSQL, que organizan 16 tablas de dimensiones y 13 tablas de hechos, lo que permitió separar las tablas de dimensiones comunes de las específicas del MD.
- Se implementaron 13 transformaciones que permitieron la carga de los datos hacia el MD Series históricas de agricultura, ganadería, silvicultura y pesca, 13 cubos OLAP y 31 reportes que garantizan la disponibilidad de la información de manera organizada, facilitando el proceso de toma de decisiones.
- Las pruebas unitarias, de integración, de sistema y de aceptación permitieron validar el MD Series históricas de agricultura, ganadería, silvicultura y pesca, garantizando que cumple con las necesidades del cliente.

RECOMENDACIONES

- Proponer un mecanismo para el tratamiento de errores de los ficheros fuentes que permita a los usuarios insertar los datos ya arreglados, sin necesidad de cargar toda la información nuevamente.
- Utilizar el contenido de la investigación como base de referencia para el desarrollo de futuros mercados de datos.

REFERENCIAS BIBLIOGRÁFICAS

- [1] Hernandez López, Ing. Asnioby. Almacenes de Datos aplicada a la Seguridad Ciudadana. Tesis (Máster en Informática Aplicada) Ciudad Habana. Universidad de las Ciencias Informáticas, 2009
- [2] http://www.sinnexus.com/business_intelligence/datawarehouse.aspx
- [3] Sinnexus. [En línea] [Citado el: 14 de Octubre de 2011.] http://www.sinnexus.com/business_intelligence/datamart.aspx.
- [4] Diaz Morales, Themis Patricia y Bermúdez Rodríguez, José Salvador. Diseño de un Datawarehouse para los Ensayos Clínicos que se gestionan en el Centro de Inmunología Molecular. Tesis (Ingeniero en Ciencias Informáticas) Ciudad Habana. Universidad de las Ciencias Informáticas, 2010.
- [5] http://etl-tools.info/es/bi/proceso_etl.html
- [6] Qué es OLAP. OlapX. [En línea] OlapXSoftware.com, 2005. [Consulta: 28 de Noviembre del 2011.]. Disponible en: <<http://www.olapxsoftware.com/es/WhatIsOlap.asp>>
- [7] DEFINICION DE METODOLOGIA [en línea] (2008). [Consulta: 30 de noviembre 2011]. Disponible en: <http://definicion.de/metodologia/>
- [8] Montenegro Campos, Eleanys María y Noguera Libera, Maybel María. Mercado de datos Transporte para el Sistema de Información de Gobierno. Tesis (Ingeniero en Ciencias Informáticas) Ciudad Habana. Universidad de las Ciencias Informáticas, 2011.
- [9] Ricardo Dario, Ing. Bernabeu. Data Warehousing: Investigación y sistematización de conceptos. Hefesto: Metodología propia para la construcción de un Datawarehouse. Córdoba : s.n., 2009.
- [10] DATEC, Especialistas de. Metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocios en DATEC. Habana : s.n., 2010.
- [11] Leon, Eduardo. *Visual Paradigm*. [En línea] (2007). [Consulta: 12 de diciembre del 2011.]. Disponible en: <http://slion2000.blogspot.com/2007/04/visual-paradigm-una-herramienta-de-lo.html>
- [12] DESCARGAS DE SOFTWARE. Sitios de descargas de software. *Visual Paradigm*. [en línea] (2007).[Consulta: 10 de diciembre del 2011.] Disponible en: http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML_%28M%C3%8D%29_14720_p/
- [13] <http://archives.postgresql.org/pgsql-es-fomento/2009-07/msg00000.php>
- [14] OFICINA NACIONAL DE ESTADISTICA [en línea] (2006). [Consulta: 23 de noviembre 2011]. Disponible en: <http://one.cu>
- [15]. Sobre la disciplina de Prueba. Conferencia 7 de Ingeniería de Software II. 2010-2011.
- [16] PRUEBAS DE SOFTWARE [en línea] (2009). [Consulta 10 mayo 2012] Disponible en: <<http://www.slideshare.net/aracelij/pruebas-de-software/>>.

BIBLIOGRAFÍA

- 1- ALMACENES DE DATOS. [en línea] (2009). [Consulta: 20 de diciembre del 2011.]. Disponible en: <http://www.encyclopediaspana.com/Almac%C3%A9n_de_datos.html>
- 2- AMERICAN AIRLINES. [en línea] (2008). [Consulta: 24 de diciembre del 2011.]. Disponible en: <<http://www.sybase.com/detail?id=210272>>
- 3- BRITISH AIRWAYS. [en línea] (2009). [Consulta: 25 de diciembre del 2011.]. Disponible en: <<http://www.highbeam.com/doc/1G1-80781004.html>>
- 4- DATAMART. [en línea] (2007). [Consulta: 26 de diciembre del 2011.]. Disponible en: <<http://www.dataprix.com/datawarehouse-vs-datamart>>
- 5- DATAMART. Ventajas y Desventajas [Consulta: 15 de diciembre de 2011]. Disponible en: <www.datawarehouse.com>
- 6- HERNÁNDEZ LÓPEZ, Ing. Asnioby. Almacenes de Datos aplicada a la Seguridad Ciudadana. Tesis (Máster en Informática Aplicada) Ciudad Habana. Universidad de las Ciencias Informáticas, 2009.
- 7- KIMBALL, Ralph y ROSS, Margy. The Data Warehouse Lifecycle Toolkit. 2a. ed. Canadá: Wiley Publishing, Inc, 2002.
- 8-DATEC, Especialistas de. Metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocios en DATEC. Habana : s.n., 2010.
- 9-Díaz Morales, Themis Patricia y Bermúdez Rodríguez, José Salvador. Diseño de un Datawarehouse para los Ensayos Clínicos que se gestionan en el Centro de Inmunología Molecular. Habana: s.n., 2010. Tesis (Ingeniero en Ciencias Informáticas).
- 10-Galindo González, Lic. Carlos y Pérez Vázquez, Dr. Ramiro. Gestipolis. [En línea] 27 de Agosto de 2009. [Consultado el: 6 de Noviembre de 2011.] <http://www.gestipolis.com/administracion-estrategia/almacenes-de-datos-y-microsoft-sql-server.htm>.
- 11-KLE. Transforming Knowledge into action! BI en la práctica. Artículos de BI en la práctica. 2010. <http://www.siskle.com/spanish/articulo04.html>.
- 12-Lanzillotta, Analia. Definición de OLAP. Tecnología OLAP. 2004. <http://www.mastermagazine.info/termino/6841.php>.
- 13- MODELOS AVANZADOS DE BASE DE DATOS. Almacenes de Datos y Bases de Datos XML. Universidad de Castilla-La Mancha (Escuela Superior de Informática). Disponible en: <<http://alarcos.inf-cr.uclm.es/doc/bbddavanzadas/08-09/FUNCIONALIDAD%204.pdf>>
- 14- SANCHEZ, Néstor. [en línea] (2008). [Consulta: 14 de enero de 2012.]. Disponible en: <<http://www.e-continua.com/documentos/indicadores2008.pdf>>
- 15- SINEXXUS. Sinergia e Inteligencia de Negocio. 2007. [Consulta: 29 de Noviembre del 2011.]. Disponible en: <http://www.sinnexus.com/business_intelligence/olap_avanzado.aspx>

- 16- SIRCAS. Sistema Interactivo de Referenciación Ambiental Sectorial. *Sistema Interactivo de Referenciación Ambiental Sectorial*. [en línea] [Consulta: 12 de febrero de 2012.]. Disponible en: <<http://www.sirac.info/Curtiembres/html/indicadores.asp>>.
- 17- SQL MAX CONNECTIONS. Data Warehousing [en línea] (2009). [Consulta: 12 de diciembre 2011]. Disponible en: <<http://www.sqlmax.com/dataw1.asp>>
- 18- http://www.sinnexus.com/business_intelligence/datawarehouse.aspx
- 19- Sinnexus. [En línea] [Citado el: 14 de Octubre de 2011.] http://www.sinnexus.com/business_intelligence/datamart.aspx.
- 20- http://etl-tools.info/es/bi/proceso_etl.html
- 21- <http://archives.postgresql.org/pgsql-es-fomento/2009-07/msg00000.php>
- 22- Sobre la disciplina de Prueba. Conferencia 7 de Ingeniería de Software II. 2010-2011.
- 23- PRUEBAS DE SOFTWARE [en línea] (2009). [Consulta 10 mayo 2012] Disponible en: <<http://www.slideshare.net/aracelij/pruebas-de-software/>>.

ANEXO # 1 Carta de aceptación del cliente:**ACTA DE ACEPTACIÓN**

En La Habana, a los 4 días del mes de junio del 2012

De una parte, la Oficina Nacional de Estadísticas e Información (ONEI) de Cuba, representado en este acto por **Elena Leonila Fernández García**, quien a los fines y efectos derivados del presente documento se denominará como "**El cliente**", y de otra Parte, el centro de Tecnologías de Gestión de Datos, conocido de forma abreviada como **DATEC** de la Universidad de las Ciencias Informáticas (UCI), representada en este acto por **José Salvador Bermúdez Rodríguez**, que a los fines y efectos derivados del presente documento se denominará **DATEC**.

Primero: Que en cumplimiento de los acuerdos, han sido efectuadas las actividades que se describen, **Las partes DECLARAN:**

CONSIDERANDO: Que se han efectuado las actividades siguientes:

1. – Diseño del mercado de datos: **Series históricas de agricultura, ganadería, silvicultura y pesca.**
2. – Proceso de extracción, transformación y carga de los datos.
3. – Implementación de las vistas de análisis OLAP.

CONSIDERANDO: Que las actividades realizadas han sido desarrolladas con la calidad requerida y bajo las condiciones pactadas y aprobadas por **Las Partes**.

CONSIDERANDO: Que las actividades que se han ejecutado cumplen con los requerimientos de **El Cliente**.

CONSIDERANDO: Que DATEC ha entregado la documentación que avala la ejecución de este acto al **El Cliente**.

POR TANTO: **Las Partes** acuerdan formalizar mediante la presente Acta, Aceptadas las actividades que han sido ejecutadas en esta fecha.

Y para que así conste, se extiende la presente Acta en dos (2) ejemplares, rubricados por **Las Partes**.

Por El Cliente

Elena Leonila Fernández García

Nombre y Apellidos

Por DATEC

José Salvador Bermúdez Rodríguez

Nombre y Apellidos


ANEXO # 2 Requisitos de información:**Sistema de Información de Gobierno**

En el presente documento se exponen los requisitos de información definidos para el mercado de datos Series históricas agricultura, ganadería, silvicultura y pesca perteneciente al proyecto Almacén de datos para el Sistema de Información de Gobierno, el cual se está desarrollando para la Oficina Nacional de Estadísticas e Información (ONEI) en el departamento Almacenes de datos del Centro de Tecnologías de Gestión de Datos (DATEC) de la Universidad de las Ciencias Informáticas (UCI).


Requisitos de información:

- 1- Obtener la distribución de la tierra por sector, año e indicadores de la agricultura.
- 2- Obtener el por ciento estructural de la tierra por sector, año e indicadores de la agricultura.
- 3- Obtener la cantidad de caña de azúcar por periodo de zafra, sector e indicadores de la zafra.
- 4- Obtener la superficie existente sembrada de cultivos permanentes por sector, año e indicadores de la agricultura no cañera.
- 5- Obtener la superficie cosechada de cultivos seleccionados por sector, año e indicadores de la agricultura no cañera.
- 6- Obtener la producción de cultivos seleccionados por sector, año e indicadores de la agricultura no cañera.
- 7- Obtener el rendimiento agrícola por sector, año e indicadores de la agricultura no cañera.
- 8- Obtener la existencia de ganado vacuno por año e indicadores del ganado vacuno.
- 9- Obtener los nacimientos de ganado vacuno por sector, año.
- 10- Obtener las muertes de ganado vacuno por sector, año.
- 11- Obtener la producción de leche de vaca por sector, año.
- 12- Obtener la existencia promedio de vacas en ordeño por sector, año.
- 13- Obtener el rendimiento anual por vaca en ordeño por sector, año.


Ing. Wendy Romalde Ruiz
Jefe de Proyecto


Liniuska Cardero Dieguez
Autora


Lic. Elena L. Fernández García
Cliente



Livan López González
Autor

ANEXO # 3 Requisitos de información


- 14-Obtener las entregas a sacrificio de ganado vacuno por año e indicadores ganado v.
- 15-Obtener el ganado porcino por sector, año e indicadores ganado porcino.
- 16-Obtener las entregas a sacrificio de ganado porcino por año e indicadores sacrificio ganado porcino.
- 17-Obtener la existencia total de aves por sector, año e indicadores avicultura.
- 18-Obtener las gallinas ponedoras por sector, año e indicadores gallinas ponedoras.
- 19-Obtener los pollos de ceba por sector, año e indicadores de los pollos de ceba.
- 20-Obtener la existencia de ganado por sector, año e indicadores de ganadería.
- 21-Obtener la existencia de ganado ovino por sector, año.
- 22-Obtener la existencia de ganado caprino por sector, año.
- 23-Obtener el ganado ovino-caprino por sector, año e indicadores ovino-caprino.
- 24-Obtener valores de apicultura por sector, año e indicadores apicultura.
- 25-Obtener las plantaciones forestales realizadas por año e indicadores silvicultura.
- 26-Obtener las semillas procesadas por año.
- 27-Obtener la producción de posturas por año.
- 28-Obtener la reconstrucción de bosques por año.
- 29-Obtener el mantenimiento silvicultural por año.
- 30-Obtener los tratamientos silviculturales por año.
- 31-Obtener las fajas verdes por año.
- 32-Obtener las trochas corta fuegos por año.
- 33- Obtener la Captura por grupos de especies por año, indicadores de pesca
- 34- Obtener la Dinámica de la captura por grupos de especies por año, indicadores de pesca

Para hacer constar la validez de la presente información firman los siguientes involucrados:


Ing. Wendy Romalde Ruiz
Jefe de Proyecto


Liniuska Cárdeno Dieguez
Autora

Lic. Elena L. Fernández García
Cliente


Livan López González
Autor

GLOSARIO DE TÉRMINOS

Almacén de Datos: es una base de datos corporativa caracterizada por integrar y depurar información de una o más fuentes distintas, permitiendo su análisis desde diversas perspectivas.

Dimensión: característica de un hecho que permite su análisis posterior en el proceso de toma de decisiones y brinda una perspectiva adicional a un hecho dado.

ETL: proceso que permite extraer, transformar y cargar los datos de un AD.

Inteligencia de negocio: conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la organización) en información estructurada, para su explotación directa o para su análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio.

Lista de chequeo: instrumento de medición y evaluación que consiste básicamente en un formulario de preguntas referentes al atributo de calidad que se está probando y de las características del documento en el caso de la documentación.

Mercado de Datos: es una base de datos departamental, que se especializa en el almacenamiento de los datos de un área específica, brindando una estructura óptima para analizar los procesos que tienen lugar dentro del departamento. Están orientados a temas específicos y contienen datos de solo una línea del negocio.

No conformidad: defecto, error o sugerencia que se le hace al equipo de desarrollo una vez encontrada alguna dificultad en lo que se está evaluando.

SGBD: Sistemas de Gestión de Bases de Datos. Es un conjunto de programas que permiten crear y mantener una base de datos, asegurando su integridad, confidencialidad y seguridad.

HECHO: operación que se realiza en el negocio la cual está estrechamente relacionada con el tiempo y es objeto de análisis para la toma de decisiones.

JDBC: protocolo de conexión de Java a base de datos (del inglés Java Data Base Connectivity).