

Universidad de las Ciencias Informáticas

FACULTAD 6



Título: Mercado de datos Extensión universitaria para la Sala Situacional de la Universidad de las Ciencias Informáticas

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

Autores:

Melissa Fong Guerra
Ernesto Dueñas Rodríguez

Tutoras:

Ing. Themis Patricia Díaz Morales
Ing. Marisel Santana Rodríguez

La Habana, junio 2012.
“Año 54 de la Revolución”



“El secreto es la respuesta a todo lo que ha sido, todo lo que es y todo lo que será por siempre”.

Emerson

DECLARACIÓN DE AUTORÍA

Declaramos ser autores del presente trabajo “Mercado de datos Extensión universitaria para la Sala Situacional de la Universidad de las Ciencias Informáticas” y reconocemos a la Universidad de las Ciencias Informáticas (UCI) los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmamos la presente a los ____ días del mes de _____ del año 2012.

Melissa Fong Guerra

Firma de la Autora

Ernesto Dueñas Rodríguez

Firma del Autor

Ing. Themis Patricia Díaz Morales

Firma de la Tutora

Ing. Marisel Santana Rodríguez

Firma de la Tutora

DATOS DE CONTACTO:

Tutoras:

Ing. Themis Patricia Díaz Morales

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente:

Categoría científica:

Años de experiencia en el tema: 3

Años de graduado: 2

Correo Electrónico: tpdiaz@uci.cu

Ing. Marisel Santana Rodríguez.

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente:

Categoría científica:

Años de experiencia en el tema: 3

Años de graduado: 2

Correo Electrónico: msantana@uci.cu

Le doy mis mayores agradecimientos a los seres más importantes de mi vida y los que han confiado toda su fe en que hoy yo sea ingeniera, a mi mamá Gricell por ser mi mundo y universo, la mejor amiga y confidente, por depositar toda su confianza siempre en mis estudios y por enseñarme todos los días que la victoria solo es conseguida por aquellos que se valen del sacrificio, a mi papá Francisco por ser desde que era pequeña mi orgullo y el ejemplo que siempre quise seguir de superación y modestia.

Agradezco a mi querido tata Jacques que siempre me protegió como un padre y aún lo hace. A mi nona Gladys como yo cariñosamente le llamo por brindarme con su sonrisa todas sus experiencias y mimarme como una abuela.

A mis abuelas Nelida por decirme lo orgullosa que siempre se sentía de mí y Catalina por velar de mi, a mi tío Perico que me apoyó como un abuelo y me aconsejó que estudiara mucho para ser en el futuro una mujer independiente, les agradezco a estos tres ángeles que a pesar de no estar físicamente, en vida constantemente me apoyaron.

Le agradezco al amor de mi vida, compañero incondicional y fiel amigo Alain por amarme y apoyarme cuando más lo he necesitado y menos lo he merecido. Por estar hecho "A la medida" de esta igualmente compañera incondicional.

En general le agradezco a toda mi familia, vecinos, amigas y amigos de la universidad en especial a Yohana y Yeilen por de una forma u otra preocuparse y aconsejarme en todo momento.

Les agradezco a mis tutoras Themis y Marisel por darnos su voto de confianza y depositar en nosotros todos sus esfuerzos, y a mi dúo de tesis por tener ese carisma peculiar que lo caracteriza y que ha hecho mucho más ameno este trabajo.

Melissa

A mi mamá Norma por todo el amor que cada día me regala, por encontrar en ella el apoyo y la comprensión que necesito para seguir adelante.

A mi abuela Haydee por dedicarme su vida sin esperar nada a cambio. Por siempre estar ahí para mí, por ese cariño incomparable que no tiene límites.

A mi tía Alba por demostrarme en cada momento el concepto de familia y por estimularme a alcanzar metas más altas.

A Melissa por los buenos momentos que pasamos juntos, por aguantar mis malas crianzas y por apoyar este gran sueño.

A mi dúo de tesis por su energía positiva y la confianza que depositó siempre en mi trabajo.

A mis tutoras Themis y Marisel por impulsar mi autoconfianza, por su ayuda constante, por su paciencia.

A todos mis compañeros, en especial a Lia, Yohia, Pino, Andry, Jorge y David por brindarme su amistad y por tantos momentos imborrables.

Ernesto

A mi familia le dedico mi tesis con todo mi amor y cariño por ser ellos el motor impulsor de lo que es hoy mis logros.

Melissa

A toda mi familia, que siempre han esperado lo mejor de mí.

A mi papá Joseito y mi abuelo Ricardo que aunque hoy, me es imposible su compañía, están más presentes que nunca.

Ernesto

RESUMEN

La presente investigación surge por la necesidad de gestionar toda la información del área de Extensión universitaria de la Universidad de las Ciencias Informáticas (UCI), donde se recopilan todos los datos relacionados con las actividades extracurriculares de profesores y estudiantes, como los eventos culturales, deportivos y recreativos. Se pretende crear un repositorio donde la información se encuentre centralizada, estandarizada y pueda ser de fácil acceso para su consulta. Para su desarrollo se documentan las diferentes herramientas informáticas, tecnología y metodología a utilizar. Se realiza el análisis, diseño e implementación del subsistema de almacenamiento, integración y visualización, así como la aplicación de las listas de chequeo y de los casos de prueba, con el objetivo de obtener un producto que cumpla con las necesidades del cliente. Finalmente se obtuvo un Mercado de Datos (MD) que almacena la información de las vicerrectorías de Extensión universitaria y Residencia, que forma parte del Almacén de Datos (AD) Sala Situacional de la UCI (SSUCI), el cual permite hacer análisis estadísticos históricos de los principales indicadores de estas áreas, facilitando la toma de decisiones a los directivos de la universidad.

Palabras claves: almacén de datos, extensión universitaria, información, mercado de datos, repositorio, toma de decisiones.

ÍNDICE

RESUMEN.....	IV
INTRODUCCIÓN.....	1
CAPÍTULO 1: Fundamentos teóricos de los Almacenes de Datos.....	4
Introducción.....	4
1.1 ¿Qué es la Extensión universitaria?	4
1.2 Almacenes de Datos	5
1.2.1 Revisión conceptual.....	5
1.2.2 Características principales de los Almacenes de Datos.....	5
1.2.3 Ventajas y desventajas de los Almacenes de Datos	8
1.3 Mercado de Datos	9
1.3.1 Características	9
1.4 Etapas de desarrollo de un Almacén de Datos.....	9
1.4.1 Análisis y Diseño.....	10
1.4.2 Extracción, transformación y carga de los datos	13
1.4.3 Inteligencia de Negocios.....	14
1.5 Metodología para el desarrollo del Mercado de Datos	16
1.5.1 Ciclo de vida Kimball.....	16
1.5.2 Metodología de DATEC	17
1.6 Herramientas informáticas para el desarrollo del Mercado de Datos.....	19
1.6.1 Herramienta de modelado	19
1.6.2 Sistema Gestor de Bases de Datos.....	20
1.6.3 Herramienta para el perfilado y limpieza de los datos.....	21
1.6.4 Herramientas para los procesos de integración de datos	22
1.6.5 Herramientas para la Inteligencia de negocio	23
Conclusiones del capítulo	25
CAPÍTULO 2: Análisis y diseño del mercado de datos Extensión universitaria.....	26
Introducción.....	26
2.1 Necesidades de información	26
2.1.1 Requisitos de información.....	26
2.1.2 Requisitos funcionales	28
2.1.3 Requisitos no funcionales	28
2.2 Reglas del negocio.....	29
2.3 Diagrama de Casos de uso del sistema	29
2.3.1 Actores	30
2.4 Casos de uso de información.....	30

2.5	Casos de uso	33
2.6	Arquitectura	34
2.7	Diseño del subsistema de almacenamiento.....	35
2.7.1	Dimensiones	36
2.7.2	Hechos	37
2.7.3	Modelo de datos	39
2.8	Diseño del subsistema de integración	41
2.8.1	Perfilado de los datos	41
2.9	Diseño del subsistema de visualización.....	42
2.10	Esquema de seguridad	44
2.10.1	Políticas de respaldo y recuperación.....	45
	Conclusiones del capítulo	45
	CAPÍTULO 3: Implementación del mercado de datos Extensión universitaria	46
	Introducción	46
3.1	Implementación del subsistema de almacenamiento	46
3.1.1	Estándares de codificación	46
3.1.2	Implementación del modelo de datos físico.....	46
3.2	Implementación del subsistema de integración	47
3.2.1	Implementación de las transformaciones	47
3.2.2	Implementación de los trabajos	49
3.3	Implementación del subsistema de visualización	50
3.3.1	Implementación de los cubos OLAP	50
3.3.2	Arquitectura de información	51
3.3.3	Implementación de los reportes candidatos	52
	Conclusiones del capítulo	53
	CAPÍTULO 4: Validación del mercado de datos Extensión universitaria	54
	Introducción	54
4.1	Pruebas aplicadas al MD Extensión universitaria	54
4.2	Herramientas para validar el MD Extensión universitaria	56
4.3	Resultados de las pruebas.....	57
	Conclusiones del capítulo	60
	CONCLUSIONES	61
	RECOMENDACIONES.....	62
	REFERENCIAS BIBLIOGRÁFICAS	63
	BIBLIOGRAFÍA	66
	ANEXOS	67

GLOSARIO69

INTRODUCCIÓN

El desarrollo vertiginoso de las tecnologías experimentó un acelerado cambio durante el transcurso de la década de los años 70, en la que se fue creando la gran estructura informática que hoy se conoce y cubre diversos campos de la sociedad. Desde entonces con el estudio, diseño y puesta en práctica de los sistemas informáticos a través del uso de hardware y software, se ha conseguido dar paso a la conversión, almacenamiento, protección, transmisión y recuperación de la información. Las Tecnologías de la Información y las Comunicaciones (TICs) han jugado un papel imprescindible en la transformación del modelo social y económico, lo que ha traído como consecuencia importantes cambios en esferas como la salud, el deporte y la educación.

A nivel mundial la mayoría de las empresas han automatizado sus procesos o se encuentran involucradas en ello, en una época en que la información es la clave para obtener una ventaja competitiva en el mundo de los negocios. Por esto, las organizaciones buscan cada día soluciones tecnológicas que les brinden mayor rapidez en el manejo de sus datos.

Cuba durante los últimos años ha aumentado el nivel informático que poseía, adaptándose a las nuevas tecnologías, por lo que se han creado centros productivos que contribuyen a la informatización del país. La Universidad de Ciencias Informáticas (UCI) es uno de los centros especializados en crear soluciones, que eleven el nivel cultural de la población y su calidad de vida. El Centro de Tecnologías de Gestión de Datos (DATEC) brinda soluciones relacionadas con el almacenamiento de datos. Dentro de este se encuentra el departamento Almacenes de Datos, el cual está realizando el proyecto Sala Situacional de la UCI (SSUCI) que gestiona la información referente a las diferentes áreas de la universidad.

Una de estas áreas es la de Extensión universitaria que está dividida en dos grandes vicerrectorías, la de Residencia y la de Extensión universitaria. Esta última se subdivide en cuatro direcciones: Extensión cultural, Deporte, Televisión universitaria y Recreaciones, encargadas de gestionar las actividades recreativas y de divulgación. Por otra parte, la vicerrectoría de Residencia se encarga de las áreas de alojamiento y de salud de la comunidad universitaria. La realización de proyectos para el desarrollo socio-cultural de la universidad y las actividades extracurriculares es una labor que se realiza de conjunto por directivos y estudiantes, bajo los lineamientos de la máxima dirección y el consejo universitario, presidido por el rector.

Esta área por toda la gestión extracurricular que procesa, acumula un gran volumen de información que se encuentra disgregada, imposibilitando su homogeneidad. Además, la información no está estandarizada, tornándose engorroso el proceso para su manejo por parte de los directivos de la institución. Existen limitaciones para realizar análisis temporales y estadísticos, pues no se dispone de

datos históricos que eviten la pérdida de información con el transcurso del tiempo, dificultando el proceso de toma de decisiones.

Por toda la situación anteriormente descrita, se plantea como **problema a resolver**: ¿Cómo contribuir al proceso de toma de decisiones en el área de Extensión universitaria a través de la Sala Situacional de la Universidad de las Ciencias Informáticas?

La investigación tiene como **objeto de estudio** los almacenes de datos, enmarcado en el **campo de acción** mercado de datos del área de Extensión universitaria para la sala situacional de la Universidad de las Ciencias Informáticas.

El **objetivo general** de la investigación es desarrollar el mercado de datos Extensión universitaria para la Sala Situacional de la Universidad de las Ciencias Informáticas, que contribuya a la toma de decisiones.

En correspondencia con el objetivo general, se han desglosado los siguientes **objetivos específicos**:

- ✓ Fundamentar la selección de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.
- ✓ Realizar el análisis y diseño del mercado de datos del área de Extensión universitaria.
- ✓ Implementar el mercado de datos del área de Extensión universitaria.
- ✓ Validar el mercado de datos del área de Extensión universitaria.

Para el cumplimiento de estos objetivos específicos se realizaron las siguientes **tareas de investigación**:

- ✓ Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.
- ✓ Levantamiento de requisitos.
- ✓ Descripción de los casos de uso del mercado de datos.
- ✓ Definición de los hechos, las medidas y las dimensiones del mercado de datos.
- ✓ Diseño del modelo de datos.
- ✓ Definición de la arquitectura del mercado de datos.
- ✓ Diseño del subsistema de integración.
- ✓ Diseño del subsistema de visualización.
- ✓ Diseño de los casos de pruebas.
- ✓ Implementación del modelo de datos.

- ✓ Implementación del subsistema de integración.
- ✓ Implementación del subsistema de visualización.
- ✓ Aplicación de las listas de chequeo.
- ✓ Aplicación de los casos de prueba

Estructura del Trabajo de diploma

El presente Trabajo de Diploma está estructurado como se muestra a continuación: resumen, introducción, cuatro capítulos, conclusiones, recomendaciones, referencias bibliográficas, bibliografía, anexos y glosario de términos.

Capítulo 1: Fundamentos teóricos de los Almacenes de Datos.

En este capítulo se abordan definiciones y conceptos importantes sobre la Extensión universitaria, Almacén de Datos y Mercado de Datos, así como características, ventajas y desventajas de su utilización. Asimismo, se documentan las metodologías, herramientas y tecnologías para el desarrollo de un Mercado de Datos y se explican las etapas del desarrollo de un Almacén de Datos para su mejor entendimiento.

Capítulo 2: Análisis y diseño del mercado de datos Extensión universitaria.

En este capítulo se realiza un análisis del negocio, con el propósito de comprender los principales aspectos de relevancia para la organización, se diseñan los subsistemas de almacenamiento, integración, visualización y se define el esquema de seguridad del Mercado de Datos.

Capítulo 3: Implementación del mercado de datos Extensión universitaria.

En este capítulo se realiza la implementación de cada uno de los subsistemas que conforman el Mercado de Datos, almacenamiento, integración y visualización.

Capítulo 4: Validación del mercado de datos Extensión universitaria.

En este capítulo se aplican diferentes tipos de pruebas de forma manual, tanto a la documentación como al sistema final. Las pruebas se realizan a través de las listas de chequeo y los casos de prueba, con el fin de comprobar si los resultados obtenidos a partir de la entrada de datos satisfacen los resultados esperados.

CAPÍTULO 1: Fundamentos teóricos de los Almacenes de Datos

Introducción

En este capítulo se abordan definiciones y conceptos importantes sobre la Extensión universitaria, Almacén de Datos (AD) y Mercado de Datos (MD), así como características, ventajas y desventajas de su utilización. Asimismo, se documentan las metodologías, herramientas y tecnologías para el desarrollo de un MD y se explican las diferentes etapas del desarrollo de un AD para su mejor entendimiento.

1.1 ¿Qué es la Extensión universitaria?

La Extensión universitaria comprende todas aquellas actividades que se realizan en la universidad que no son de carácter docente ni investigativo, las que se denominan como “actividades de extensión”. Roberto Rodríguez Gómez la define en su escrito “El significado de la extensión universitaria en el presente” como: “...la Extensión universitaria, debe proponerse con fines fundamentales, proyectar dinámica y coordinadamente la cultura y vincular a todos los habitantes universitarios con la propia universidad; estimulando el desarrollo social, elevando el nivel espiritual, intelectual y técnico de la nación, proponiendo imparcial y objetivamente ante la opinión pública, las soluciones fundamentales a los problemas de interés general” (1).

Extensión universitaria es la interacción entre la universidad y los demás componentes del cuerpo social, a través de la cual ésta asume y cumple su compromiso de participación, en el proceso social de creación de la cultura y de liberación y transformación radical de la comunidad nacional (2). Es la respuesta a la pregunta ¿de qué manera se relaciona la universidad con su entorno, además de formar profesionales y desarrollar investigación especializada?

Gestión de la información en el área Extensión universitaria de la UCI

El proceso de almacenamiento de la información en esta área varía en dependencia de la vicerrectoría, esto provoca que existan grandes volúmenes de datos de diferentes lugares dentro de la universidad y en distintos formatos, como consecuencia se dificulta el análisis y se cometen errores estadísticos. Con el objetivo de resolver este problema se creó en conjunto con los especialistas de la vicerrectoría de Extensión seis modelos en formato Excel que estandarizan las fuentes de datos, por otro lado en la vicerrectoría de Residencia la información se almacena en un sistema que solo tiene desplegado el módulo de alojamiento y no satisface las necesidades informativas de los directivos.

1.2 Almacenes de Datos

Desde su creación se ha ido posicionando como la vía más acertada para la realización de análisis de información histórica. Es una potente herramienta para la recuperación efectiva de las más complejas consultas y de servir como base para la toma de decisiones.

1.2.1 Revisión conceptual

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 la tabla hace referencia a las concepciones de tres autores:

Tabla 1: Conceptos de almacén de datos

“...una copia de datos transaccionales, específicamente estructurados para la consulta y el análisis” (3).
Colección almacenada de datos corporativos originados en diversos sistemas transaccionales dentro de una organización. Un almacén de datos es un repositorio único de datos cuyo propósito específico es soportar la toma de decisiones en un negocio, no las operaciones del mismo, por ello se encuentra organizado de una forma muy distinta a las bases de datos tradicionales (4).
Un almacén de datos es un conjunto de datos integrados, orientados hacia una materia, que varían con el tiempo y que no son transitorios, los cuales apoyan el proceso de toma de decisiones de una administración (5).

En la presente investigación se asumirá la tercera definición ya que es el concepto más completo definiendo las características de los AD.

1.2.2 Características principales de los Almacenes de Datos

Las principales características que definen a los AD son (6):

- ✓ Orientado a temas.
- ✓ Integrado.
- ✓ De tiempo variante.
- ✓ No volátil.

Orientado a temas: el diseño del AD está orientado en cuanto a temas globales y por las principales áreas temáticas de la empresa (Figura 1). Esta cualidad contrasta con el clásico método que era

orientado al proceso y funcionamiento de las aplicaciones utilizadas en sistemas operacionales más antiguos.



Figura 1: Orientado a temas. Los datos se organizan por áreas temáticas.

Integrado: los datos necesitan ser almacenados en los AD de una forma globalmente aceptable y singular, aunque el programa operacional los almacene de una forma distinta. Los datos deben de ser consistentes siempre dentro del AD e integrados de distintas fuentes de datos operacionales (Figura 2).



Figura 2: Integrado. Cuando los datos se cargan al AD desde las fuentes de datos, son integrados quedando de la forma deseada.

De tiempo variante: las características de los datos contenidos en los AD son distintas a aquellos en el ambiente operacional (Figura 3). En el ambiente operacional, al momento de acceder a los datos estos son exactos a aquellos que se espera recibir. Sin embargo, en los AD los datos son exactos en algún momento del tiempo. Las variantes del tiempo se pueden notar de tres formas:

- Límite de tiempo: el margen de tiempo del AD es mucho mayor en cuanto a los datos (puede contener datos entre 5 y 10 años de almacenamiento). Por otro lado, en el ambiente operacional el margen de tiempo de almacenamiento de los datos es mucho menor, por lo que la información a la cual se accede es mucho más “fresca” (60 - 90 días); ya que un programa de aplicación para trabajar eficientemente debe llevar la mínima cantidad de datos necesarios para realizar las transacciones.
- Clave de estructura: los datos en los AD contienen un elemento de tiempo (Ej.: día, semana, mes y año).
- Actualizaciones: los datos una vez almacenados correctamente en los AD no pueden ser alterados, por lo tanto no se pueden actualizar.



Figura 3: De tiempo variante. Los informes que se generan reflejan las variaciones que se han producido en los datos a lo largo del tiempo.

No volátil: la manipulación de datos en los AD es mucho más simple. Solo existen dos tipos de operaciones que se llevan a cabo en los AD: cuando se cargan inicialmente los datos y cuando se acceden a él (Figura 4). Los datos almacenados no se modifican ni actualizan nunca, solo se añaden nuevos datos.



Figura 4: No volátil. La información no se modifica ni se elimina, cuando se carga al almacén se convierte en información de sólo lectura, manteniéndolo para futuras consultas. En una BD operacional los datos son actualizados registro por registro.

1.2.3 Ventajas y desventajas de los Almacenes de Datos

La implementación de un AD puede beneficiar a una organización ya que (7):

- ✓ Aumenta la productividad de los técnicos de dirección.

- ✓ Hacen más fácil el acceso a una gran variedad de datos a los usuarios finales.
- ✓ Se obtiene una base de datos histórica y clasificada por temas.
- ✓ Integra información procedente de múltiples sistemas externos.
- ✓ Facilita la toma de decisiones estratégicas.

Sin embargo, utilizar AD también trae consigo algunos problemas como (8):

- Infravaloración de los recursos necesarios para la captura, carga y almacenamiento.
- Infravaloración del esfuerzo necesario para su diseño (integración, consolidación) y creación.
- Incremento continuo de los requerimientos de los usuarios.
- Privacidad de los datos.

1.3 Mercado de Datos

Un MD almacena la información de un área o departamento específico dentro del negocio y un conjunto de mercados de datos forman un AD. Un MD es una solución que, compartiendo tecnología con el AD (pero con contenidos específicos, volumen de datos más limitado y un alcance histórico menor), permita dar soporte a una empresa pequeña, un departamento o área de negocio de una empresa grande (9).

Los datos pueden ser agrupados, explorados y propagados de diferentes formas para que diversos grupos de usuarios realicen la explotación de los mismos de la forma más conveniente según sus necesidades. Brevemente se puede decir que un MD es un pequeño AD centrado en un tema o un área de negocio específico dentro de la organización.

1.3.1 Características

Los MD se caracterizan por disponer de (10):

- Poco volumen de datos: la existencia de una menor cantidad de datos implican que se procesan antes tanto las cargas de datos como las consultas.
- Se centran en los requisitos de los usuarios asociados a un departamento o área de negocio concreto.
- Costo inferior al de la creación de un AD completo.
- 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 un AD.

1.4 Etapas de desarrollo de un Almacén de Datos

Independientemente de la metodología que se seleccione, el desarrollo de un AD transita por tres etapas fundamentales las cuales serán descritas a continuación.

1.4.1 Análisis y Diseño

Toda construcción necesita una base sólida para que perdure en el tiempo y se obtenga la calidad esperada. En esta etapa se efectúa un estudio del negocio y se identifican las necesidades del cliente. Con este paso se tiene el necesario conocimiento del negocio; sus objetivos, prioridades, la frecuencia con que se maneja la información y otros puntos de utilidad que servirán para la implementación de un AD.

Después de haber realizado la especificación del negocio se identifican los requisitos necesarios para lograr un producto con calidad que satisfaga las necesidades informáticas del cliente. Estos se recogen en la especificación de requisitos, el cual registra los requerimientos informativos, funcionales y los no funcionales definidos (3).

Al finalizar el análisis se procede a realizar el diseño, es en dicha etapa donde se realiza el modelo dimensional. En este modelo, a diferencia de los sistemas de BD que presentan sus estructuras diseñadas mediante el modelo Entidad-Relación, los AD se diseñan mediante un Modelo Dimensional. Este posee la misma información que el Diagrama de Entidad Relación (DER) pero la organiza de forma diferente. Una de sus características principales es que no necesita una predefinición de los reportes, debido a que se diseñan de forma tal que cubra las posibles variantes que los usuarios necesiten consultar.

El **modelo dimensional** divide el mundo de los datos en dos grandes conjuntos: las medidas y las descripciones del entorno de estas medidas (Figura 5). Las medidas generalmente numéricas se almacenan en las tablas de hechos y las descripciones de los entornos que son textuales se almacenan en las tablas de dimensiones. Las **tablas de hechos** son las tablas primarias en el modelo dimensional y contiene los valores del negocio. Cada tabla contiene dos o más llaves extranjeras que acoplan con sus respectivas tablas de dimensiones.

Las **tablas de dimensiones** son las compañeras integrales de las tablas de hechos. En el modelo dimensional, las tablas de dimensiones poseen varios atributos que en su conjunto definen una fila. Los atributos de las dimensiones sirven como fuente primaria de las restricciones de las consultas, agrupaciones y las etiquetas de los reportes (11).

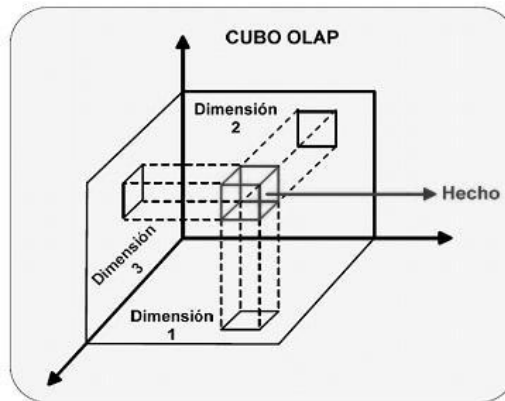


Figura 5: Estructura espacial del modelado dimensional. Los hechos representan la captura de un evento dentro del negocio y las dimensiones las perspectivas de análisis con las que se describe ese evento.

Esquemas de modelado (12)

Los esquemas de modelado son la forma en que se va a estructurar el depósito de datos y simplifican el entendimiento de los mismos por parte de los usuarios. Definir qué esquema se utilizará para el almacenamiento de los datos es muy importante, ello ayuda a un buen modelado dimensional. La elección depende del negocio y los requerimientos del cliente.

Esquema de modelado estrella (star): en el esquema en estrella una sola tabla de hechos está relacionada a cada tabla de dimensión (Figura 6). Las tablas de dimensiones son enlazadas a la tabla de hechos mediante referencias de una llave foránea. La llave primaria en la tabla de hechos se compone de una relación de las llaves primarias de las tablas de dimensiones. Se le denomina estrella por su similitud con una estrella natural.

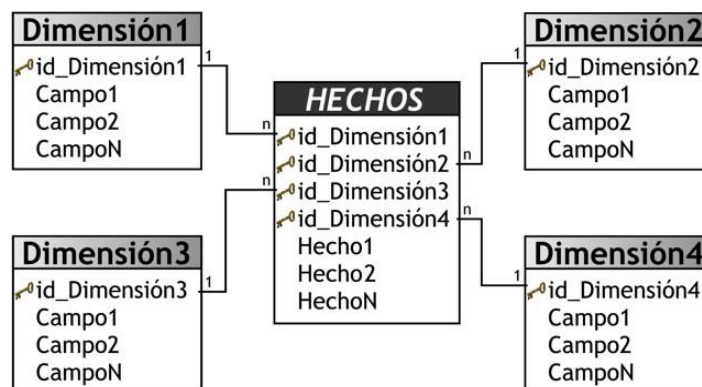


Figura 6: Esquema estrella. Una tabla de hechos y una tabla para cada dimensión

Características:

- Posee los mejores tiempos de respuesta.
- Su diseño es fácilmente modificable.

- Existe paralelismo entre su diseño y la forma en que los usuarios visualizan y manipulan los datos.
- Simplifica el análisis.
- Facilita la interacción con herramientas de consulta y análisis.

Esquema copo de nieve (snowflaked): surgen por modificaciones realizadas al esquema estrella, con el objetivo de ahorrar espacio de almacenamiento. En el esquema copo de nieve las tablas de dimensiones son normalizadas para lograr presentar la información sin redundancia (Figura 7).

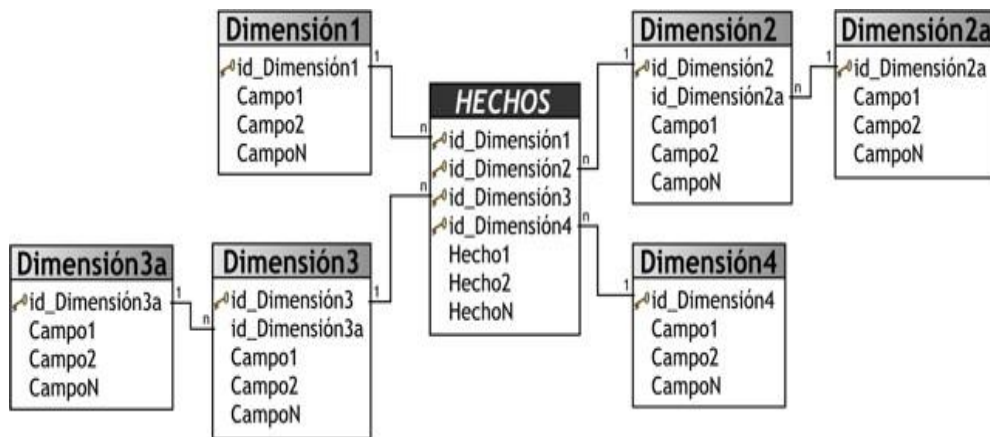


Figura 7: Esquema copo de nieve. Las tablas de las dimensiones se encuentran normalizadas.

Características:

- Posee mayor complejidad en su estructura.
- Hace una mejor utilización del espacio.
- Es muy útil en tablas de dimensiones de muchas tuplas.
- Las tablas de dimensiones están normalizadas, por lo que requiere menos esfuerzo de diseño.
- Puede desarrollar clases de jerarquías fuera de las tablas de dimensiones, que permiten realizar análisis de lo general a lo detallado y viceversa.

Esquema constelación de hechos: los esquemas en estrella y copo de nieve pueden generalizarse con la inclusión de distintas tablas de hechos que comparten todas o algunas de las dimensiones (Figura 8). De esta forma se pueden utilizar diversas medidas, separadas en diferentes tablas de hechos, definidas por las mismas dimensiones.

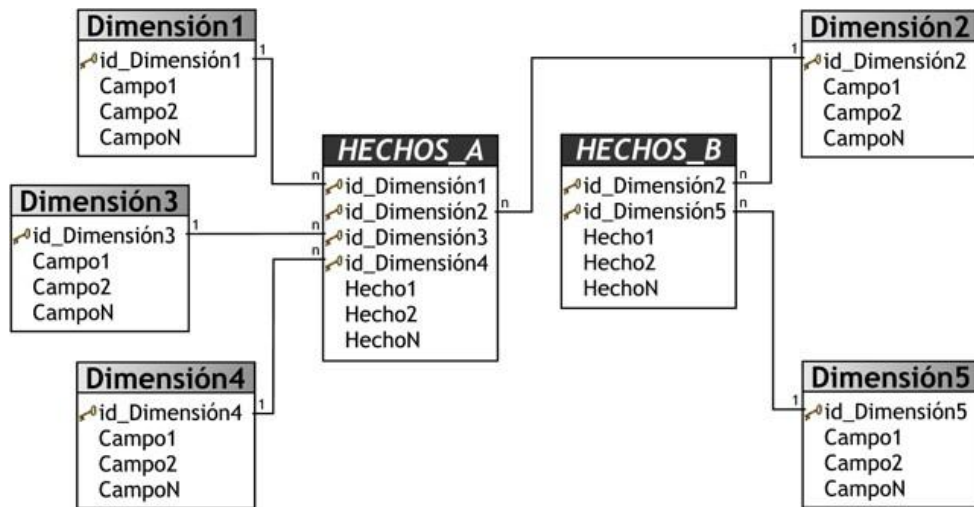


Figura 8: Esquema constelación de hechos. Esquemas en estrella o copo de nieve en el que los hechos comparten dimensiones.

Características:

Su diseño y cualidades son muy similares a las del esquema en estrella, pero posee una serie de diferencias con él, que son precisamente las que lo destacan y caracterizan. Entre ellas se pueden mencionar:

- Permite tener más de una tabla de hechos, por lo cual se podrán analizar más aspectos claves del negocio con un mínimo esfuerzo adicional de diseño.
- Contribuye a la reutilización de las tablas de dimensiones, ya que una misma tabla de dimensión puede utilizarse para varias tablas de hechos.

1.4.2 Extracción, transformación y carga de los datos

Al concluir con la etapa de análisis y diseño, se da paso al proceso de extracción, transformación y carga de los datos (ETL por sus siglas en inglés), donde se extraen los datos que proceden de distintas fuentes de información. Estos datos se transforman, pues no siempre se encuentran homogéneos y estandarizados. Por último, se da paso a cargarlos en el depósito de datos o sistema de destino (13).

Extracción: la primera parte del proceso ETL consiste en extraer los datos desde los sistemas de origen. La mayoría de los proyectos de almacenamiento de datos fusionan datos provenientes de diferentes sistemas de origen. Cada sistema separado puede usar una organización diferente de los datos o formatos distintos. Los formatos de las fuentes normalmente se encuentran en bases de datos relacionales o ficheros Excel, pero pueden incluir bases de datos no relacionales u otras estructuras diferentes.

Transformación: Esta fase es la encargada de convertir aquellos datos inconsistentes en un conjunto de datos compatibles y congruentes, para que puedan ser cargados en el AD. Estas acciones se realizan debido a que pueden existir diferentes fuentes de información y es vital conciliar un formato y forma única, definiendo estándares, para que todos los datos que ingresen al AD estén integrados. Puede involucrar la división o la combinación de registros fuentes y la corrección de errores para que los datos queden de la forma deseada.

Carga: Este proceso es el responsable de cargar la estructura de datos del AD con aquellos datos que han sido transformados y que residen en el almacenamiento intermedio y los datos de los OLTP que tienen correspondencia directa con el depósito de datos. Se debe tener en cuenta que antes de mover los datos al AD es necesario analizarlos con el propósito de asegurar su calidad, ya que es un factor clave que no debe dejarse de lado (12).

1.4.3 Inteligencia de Negocios

Todo el proceso del análisis de los datos se realiza a través de técnicas de análisis que extraen de los datos la información que se necesita para una posterior toma de decisiones y que se muestre al cliente de forma más entendible, estas son (14):

- ✓ Reportes y consultas (del inglés Queries and reports).
- ✓ Cubos OLAP (del inglés On Line Analytical Processing).
- ✓ Minería de Datos (del inglés Data Mining).

Reportes y consultas

Esta técnica no es más que el uso de herramientas destinadas a la producción de consultas y reportes. Estas ofrecen a los usuarios la posibilidad de generar informes del área de interés del negocio que se esté analizando, a través de pantallas gráficas intuitivas. El usuario únicamente debe seleccionar las opciones que le brindan para especificar los elementos de datos, sus condiciones, criterios de agrupación y demás atributos que se consideren significativos.

Minería de Datos

Es un mecanismo de explotación consistente en la búsqueda de información valiosa en grandes volúmenes de datos capturados a lo largo de las operaciones del negocio. La minería de datos se centra en satisfacer la necesidad de descubrir el por qué, para luego predecir y pronosticar las posibles acciones con cierto factor de confianza para cada predicción. Es el análisis de archivos y bitácoras de transacciones, trabaja a nivel del conocimiento con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones útiles para la toma de decisiones.

OLAP

Los sistemas OLAP son BD orientadas al procesamiento analítico. Este análisis suele implicar, generalmente, análisis complejos de grandes volúmenes de datos para llegar a extraer algún tipo de información útil: tendencias de ventas, patrones de comportamiento de los consumidores y elaboración de informes complejos. Esta técnica es la usada por el departamento DATEC y existen tres modelos:

- ✓ **Procesamiento Analítico Relacional en Línea (ROLAP por sus siglas en inglés):** almacena los datos en un motor relacional. Utiliza una arquitectura de tres niveles que permite el análisis de una enorme cantidad de datos:
 - El nivel de base de datos usa bases de datos relacionales para el manejo, acceso y obtención de los datos.
 - El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios.
 - El motor ROLAP se integra con niveles de presentación, a través de los cuales los usuarios realizan los análisis OLAP
- ✓ **Procesamiento Analítico Multidimensional en Línea (MOLAP por sus siglas en inglés):** esta implementación OLAP almacena los datos en una BD multidimensional. Provee excelente rendimiento y compresión de los datos, tiene el mejor tiempo de respuesta, muy adecuado para cubos por su rápida respuesta.
- ✓ **Procesamiento Analítico Híbrido en Línea (HOLAP por sus siglas en inglés):** es una solución que incluye las implementaciones anteriores (MOLAP y ROLAP). Las agregaciones de los datos son almacenadas en una estructura multidimensional usada por MOLAP y la BD fuente en una BD relacional. Los cubos almacenados como HOLAP son más pequeños que los MOLAP y responden más rápido que los ROLAP. Es generalmente usado para cubos que requieran rápida respuesta para sumas de grandes cantidades de datos.

Para el desarrollo del MD se escogió como modo de almacenamiento de datos al ROLAP porque los usuarios finales pueden realizar sus análisis multidimensionales a través de este modo, que transforma dinámicamente sus consultas a consultas SQL. Luego estas consultas se ejecutan en las bases de datos relacionales y sus resultados se relacionan mediante tablas cruzadas y conjuntos multidimensionales para devolver los resultados a los usuarios. Además, esta arquitectura accede directamente a los datos almacenados en el AD, es capaz de usar datos anteriormente calculados si estos están disponibles, generar dinámicamente los resultados desde los datos elementales y soporta técnicas de optimización de accesos para mejorar el tiempo de respuesta de las consultas.

1.5 Metodología para el desarrollo del Mercado de Datos

La metodología desempeña un papel fundamental en la elaboración de cualquier producto informático, ya que se puede categorizar como el conjunto de pasos, métodos y procedimientos que rigen una investigación para alcanzar los objetivos planteados. Para la elaboración y desarrollo de un AD, existen diferentes metodologías que se enmarcan en dos enfoques o vertientes principales. La propuesta de Ralph Kimball (principal promotor del enfoque dimensional para el diseño de almacenes de datos) y la de Bill H. Inmon (conocido por muchos como el padre del AD). Ambos son las personalidades más influyentes en el área de los AD y responsables de las dos vertientes a las que se hace referencia.

La visión de Inmon se basa principalmente en un enfoque descendente (del inglés top-down), además plantea la creación de un repositorio de datos corporativo como fuente de información consolidada, persistente, histórica y de calidad. Al ser construido descendentemente, el MD se nutre del AD corporativo, convirtiéndose en un complejo empresarial de BD relacional.

La propuesta de Kimball se basa principalmente en un enfoque ascendente (del inglés bottom-up), debido a que plantea que se debe crear por departamentos un conjunto de MD independientes, orientados a los temas que estén relacionados con él. “El Almacén de Datos es la unión de todos los Mercados de Datos de una entidad” (15).

Existen en el mundo diferentes metodologías para el desarrollo de AD entre las que se encuentran Metodología Hefesto, Desarrollo de almacenes de datos dirigidos por modelos (Trujillo) y el Ciclo de vida Kimball.

1.5.1 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 se definen los requisitos del negocio. Luego de esto, el proyecto se enfoca en tres líneas concurrentes: tecnología, datos y aplicaciones de BI. El ciclo de vida culmina con el despliegue y mantenimiento del producto (11).

Ventajas de su utilización (15):

- ✓ Crea los conceptos de hechos y dimensiones, lo que es muy importante en el proceso de la toma de decisiones y a su vez proporciona mayor agilidad en el proceso de desarrollo.
- ✓ Propone ir construyendo el AD a través de la construcción de los MD departamentales, lo que constituye una buena estrategia y coincide con la división lógica de las empresas, entidades y organismos.
- ✓ Es una metodología madura y reconocida por los usuarios dedicados al tema, además de tener bien definidas sus etapas, actividades, roles y artefactos.

Sin embargo, su uso también trae algunas desventajas (15):

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

1.5.2 Metodología de DATEC

El **Modelo para el desarrollo de soluciones de Almacenes Datos e Inteligencia de Negocios en DATEC** se basa en el ciclo de vida Kimball y en la propuesta realizada por Leopoldo Zenaido Zepeda Sánchez en su tesis de doctorado, en la cual plantea incluir los casos de uso para guiar el proceso de desarrollo (16).

Cuenta con 4 fases principales (15):

1. Requerimientos y gestión del proyecto.
2. Arquitectura.
3. Diseño e implementación.
4. Implantación y operaciones.

Durante su ciclo de vida se tienen los siguientes flujos de trabajo (15):

- ✓ **Estudio preliminar y planeación:** se realiza el estudio de la entidad cliente, la planeación del proyecto, se definen los objetivos, el alcance preliminar, los costos estimados y otras actividades.
- ✓ **Requerimientos:** se realiza en dos direcciones, una mediante la identificación de las necesidades de información y reglas del negocio; y la otra con un levantamiento detallado de las fuentes de datos a integrar. Luego se procede a la definición de los requerimientos.

- ✓ **Arquitectura:** se definen las estructuras de almacenamiento, se diseñan las reglas de ETL y se define la arquitectura de información que regirá el desarrollo de la solución.
- ✓ **Implementación:** se diseña físicamente el repositorio de datos, se crean las estructuras de almacenamiento, el área temporal de almacenamiento, se ejecutan las reglas de transformación y se configuran e implementan las herramientas de Inteligencia de Negocio (BI por sus siglas en inglés) para la obtención de los elementos que se acordaron con el cliente final.
- ✓ **Prueba:** se realizan las pruebas al sistema desde las pruebas de unidad hasta las de aceptación con el cliente final.
- ✓ **Despliegue:** se realiza un despliegue piloto en el cual se configuran los servidores, se instalan las herramientas y se carga una muestra de los datos para demostrar que el sistema funciona. Posterior a la aceptación del cliente se realiza la carga de los datos, la capacitación y transferencia tecnológica. El resultado fundamental es la solución desplegada en el entorno real en correcto funcionamiento.
- ✓ **Soporte y mantenimiento:** comienza cuando la solución está implantada y en explotación, y se ejecuta según el contrato firmado y las condiciones de soporte establecidas.
- ✓ **Gestión y administración del proyecto:** a lo largo del ciclo de vida se realizan actividades de control, gestión y chequeo del desarrollo, de los gastos, las utilidades, los recursos y demás actividades por parte del grupo de dirección del proyecto.

La utilización de esta metodología puede traer muchas ventajas como son (15):

- ✓ Los productos son más comprensibles para los usuarios.
- ✓ La solución completa se puede implementar en poco tiempo.
- ✓ Es resistente y tolerante ante los cambios.

En la presente investigación se definió como metodología a utilizar el Modelo para el desarrollo de soluciones de Almacenes Datos e Inteligencia de Negocios en DATEC, que permite mitigar las desventajas identificadas en la Metodología de Kimball y ajustarse a las condiciones y características de producción de DATEC y de la UCI. Las particularidades que presenta este modelo de adaptación es la identificación de requerimientos de información y a su vez la trazabilidad que tienen estos en todo el ciclo de desarrollo del MD. También la inclusión de una etapa de pruebas que fortalece la calidad con que se despliega la solución propuesta. Ajusta sus fases, actividades y artefactos a las propuestas del programa de mejora para alcanzar el nivel 2 de CMMI.

1.6 Herramientas informáticas para el desarrollo del Mercado de Datos

Para realizar el modelado del negocio, diseño, implementación y puesta en marcha del sistema se hace necesario el uso de herramientas y tecnologías que permitan el desarrollo de la aplicación. En los siguientes epígrafes se muestra un análisis de las herramientas usadas para la realización del MD del área de Extensión universitaria.

1.6.1 Herramienta de modelado

Al aplicar la Ingeniería de Software durante el proceso de desarrollo de un producto informático es necesario modelar artefactos, para esto existen varias herramientas automatizadas que utilizan técnicas de diseño y metodologías bien definidas. En la presente investigación se utiliza para el diseño Visual Paradigm for UML 8.0, ya que es multiplataforma y permite su uso en cualquier sistema operativo. Está disponible en varios idiomas, además es fácil de instalar y fácil de actualizar.

Visual Paradigm for UML 8.0

El Visual Paradigm for UML (VP) es una herramienta profesional que soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue. El Lenguaje Unificado de Modelado (UML por sus siglas en inglés) ayuda a una rápida construcción de aplicaciones mejores y a un menor coste. Permite modelar todos los tipos de diagramas de clases, código inverso, generar código desde diagramas y generar documentación. Proporciona abundantes tutoriales, demostraciones interactivas y proyectos de UML (17). La UCI cuenta con la licencia para su uso.

Entre sus principales características se encuentran (18):

- ✓ Soporte de UML versión 2.1.
- ✓ Diseño centrado en casos de uso y enfocado al negocio que genera un software de mayor calidad.
- ✓ Uso de un lenguaje estándar común a todo el equipo de desarrollo que facilita la comunicación.
- ✓ Disponibilidad en múltiples plataformas.
- ✓ Contiene facilidades para redactar especificaciones de casos de uso del sistema.
- ✓ Sincronización entre diagramas de entidad relación y diagramas de clases.
- ✓ Generación de documentos.
- ✓ Integración con distintos ambientes integrados de desarrollo.

1.6.2 Sistema Gestor de Bases de Datos

Los SGBD se utilizan para el manejo de los datos, por lo que se convierten en un elemento clave de la información. Estos sistemas van a permitir que se pueda realizar la eliminación y actualización de registros, así como la combinación con otras BD y la generación de informes impresos (19).

PostgreSQL 9.1

PostgreSQL es un sistema de gestión de bases de datos objeto-relacional basado en el proyecto Postgres de la Universidad de Berkeley. Es una derivación libre de este proyecto y utiliza el lenguaje SQL92/SQL99. Fue el pionero en muchos de los conceptos existentes en el sistema objeto-relacional actual, incluido más tarde en otros sistemas de gestión comerciales. Incluye características de la orientación a objetos, como puede ser: la herencia, tipos de datos, funciones, restricciones, disparadores, reglas e integridad transaccional. A pesar de esto, no es un sistema de gestión de bases de datos puramente orientado a objetos (20).

A continuación en la tabla se muestra algunas características que presenta el SGBD PostgreSQL 9.1.

Tabla 2: Características del SGBD PostgreSQL

Característica	Descripción
Instalación ilimitada	PostgreSQL no puede ser demandado por violar acuerdos de licencia, puesto que no hay costo asociado a la licencia del software.
Soporte	Existe una importante comunidad de profesionales y entusiastas de PostgreSQL de los que su compañía puede obtener beneficios y contribuir.
Ahorros considerables en costos de operación	PostgreSQL ha sido diseñado y creado para tener un mantenimiento y ajuste mucho menor que otros productos, conservando todas las características, estabilidad y rendimiento.
Extensible	El código fuente está disponible para todos sin costo. Si su equipo necesita extender o personalizar PostgreSQL de alguna manera pueden hacerlo con un mínimo esfuerzo, sin costos adicionales.

Se seleccionó PostgreSQL teniendo en cuenta sus características pues ofrece estabilidad, rendimiento y es capaz de manejar grandes volúmenes de información, además está respaldado por la comunidad de PostgreSQL para la obtención de información y uso de la herramienta. Con la selección de esta herramienta se cumple con las políticas de migración a software libre de la UCI.

PgAdmin III 1.14.0

En la presente investigación se escoge PgAdmin III en su versión 1.14.0 pues está diseñado para responder a las necesidades de la mayoría de los usuarios, desde escribir simples consultas SQL hasta desarrollar bases de datos complejas. La interfaz gráfica soporta todas las características de PostgreSQL y hace simple la administración. Está disponible en más de una docena de lenguajes y para varios sistemas operativos, incluyendo Microsoft Windows, Linux, FreeBSD, Mac OSX y Solaris. Soporta versiones de servidores de BD PostgreSQL 7.3 o superiores (21).

PgAdmin III es una herramienta de código abierto para la administración de BD que incluye:

- ✓ Interfaz administrativa gráfica.
- ✓ Herramienta de consulta SQL (con un EXPLAIN gráfico).
- ✓ Editor de código procedural.
- ✓ Agente de planificación SQL/shell/batch.

1.6.3 Herramienta para el perfilado y limpieza de los datos

Los datos de origen en muchas ocasiones se encuentran con errores. Las herramientas para la limpieza de los datos permiten mitigar estos errores con el fin de que se almacenen de forma estandarizada.

DataCleaner 1.5.4

DataCleaner es una aplicación Open Source para el análisis, perfilado, transformación y limpieza de datos. Estas actividades ayudan a mejorar la calidad de los datos (22).

Dentro de sus características se incluyen:

- ✓ Validación de los datos: el validador le dará un resultado que puede ser interpretado como bueno o malo, ya que valida los datos.
- ✓ Compatibles con diferentes tipos de BD: Oracle, MySQL, PostgreSQL, Firebird y SQLite
- ✓ Soporta acceso de lectura a muchos tipos de AD:
 - BD compatibles con JDBC (oficialmente probadas y compatibles: Oracle, MySQL, PostgreSQL, Firebird, SQLite, HSQLDB, Derby/javadb).
 - Excel (.xls) hojas de cálculo.
 - Archivos XML.

La Licencia Pública General Menor (por sus siglas en inglés LGPL) le permite que cualquiera pueda utilizar el software para todos los efectos, pero ninguna de las modificaciones realizadas en el código debe ser aportado a la comunidad. Es una aplicación muy fácil de usar, genera sofisticados informes y gráficos que permiten a los usuarios determinar de un vistazo el nivel de calidad de los datos, identificar y analizar la estructura del origen de datos y combinar resultados y gráficos, creando vistas fáciles de interpretar.

1.6.4 Herramientas para los procesos de integración de datos

La informatización como parte del avance tecnológico trae consigo a gran escala la utilización de modernos medios de integración de la información, entre los que se encuentra el Pentaho Data Integration. Herramienta que permite la extracción, transformación y carga de los datos cubriendo sus necesidades.

Pentaho Data Integration 4.2.1 (PDI)

Es una de las soluciones más extendidas y mejor valoradas del mercado. Permite realizar transformaciones y trabajos de una forma muy sencilla e intuitiva. Es aplicable a diversos tipos de BD como son SQL server, PostgreSQL, MySQL, Microsoft Access. A parte de ser open source y sin costes de licencia, las características básicas de esta herramienta son (22):

- ✓ Entorno gráfico de desarrollo.
- ✓ Uso de tecnologías estándar: Java, XML, JavaScript.
- ✓ Fácil de utilizar.
- ✓ Multiplataforma: Windows, Macintosh, Linux.
- ✓ Basado en dos tipos de objetos: Transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones).

Pentaho incluye cuatro herramientas (23):

- ✓ **Spoon:** es una interfaz gráfica de usuario que permite diseñar transformaciones y trabajos que se pueden ejecutar con las herramientas Pan y Kitchen.
- ✓ **Pan:** es un motor de transformación de datos que realiza muchas funciones tales como lectura, manipulación y escritura de datos hacia y desde varias fuentes de datos.
- ✓ **Chef:** permite diseñar la carga de datos incluyendo un control de estado de los trabajos.
- ✓ **Kitchen:** es un programa que ejecuta los trabajos diseñados.

Las soluciones de Pentaho están escritas en java y tienen un ambiente de implementación también basado en java. Eso hace que Pentaho sea una solución muy flexible para cubrir una amplia gama de necesidades empresariales.

Entre sus principales ventajas se encuentran (23):

- ✓ Abre, limpia e integra la información y la pone a disposición del usuario
- ✓ Tiene una gran cantidad de usuarios por ser una de las herramientas libres más antiguas, los cuales comparten muchos consejos y trucos.
- ✓ Provee una sola versión de todos los recursos de información, que es uno de los más grandes desafíos para las organizaciones.

1.6.5 Herramientas para la Inteligencia de negocio

Las tecnologías de BI son herramientas que permiten analizar y manipular la información de las empresas. Con un entendimiento más acelerado de los datos se pueden tomar mejores y más rápidas decisiones. Proporcionan el impulso clave hacia el logro de los objetivos propuestos. El líder actual en cuanto a soluciones de BI es el Pentaho, pues ofrece sus soluciones propias y cuenta con una amplia variedad de recursos para desarrollar, mantener y explotar un proyecto de inteligencia de negocio.

Pentaho Schema Workbench 3.2.1

La herramienta Pentaho Schema Workbench es una interfaz de diseño que le permite crear y probar esquemas de cubos OLAP visualmente. El motor de Mondrian procesa las solicitudes de MDX (del inglés MultiDimensional eXpressions) con el esquema ROLAP (OLAP Relacional). Estos archivos son los modelos de esquemas XML (lenguaje de marcas extensibles) de metadatos que se crean en una estructura específica utilizada por el motor de Mondrian. Los modelos XML se pueden considerar como las estructuras en forma de cubo que utilizan hechos existentes y tablas de dimensiones.

Ofrece las siguientes funcionalidades (24):

- ✓ Editor de esquema integrado con el origen de datos subyacente para su validación.
- ✓ Permite la ejecución de consultas MDX contra el esquema y la BD.
- ✓ La navegación por la BD subyacente.

Mondrian OLAP Server 3.0.4

Mondrian es un motor Hybrid OLAP que combina la flexibilidad de los motores ROLAP con una caché que le proporciona velocidad. El núcleo del servidor Mondrian es similar a JDBC, pero exclusivo para OLAP. Este proporciona la conexión a la BD y ejecuta las sentencias 'SQL'. Entre sus principales características se encuentra las facilidades que brinda para el análisis de grandes volúmenes de información, que se encuentren almacenados en BD. OLAP es la tecnología que permite organizar la información en una estructura dimensional que nos posibilita desplazándonos por sus dimensiones. Permite además realizar consultas al AD posibilitando que los resultados sean presentados mediante un navegador, de modo que el usuario pueda realizar las actividades típicas de navegación. Mondrian utiliza MDX, como lenguaje de consulta, que fue un lenguaje propuesto por Microsoft. Puede ser

integrado independientemente en cualquier otra plataforma y de hecho es el componente, junto con Data Integration, que más se utiliza independientemente (25).

Pentaho BI Server 3.10

Proporciona los servicios básicos, incluidos autenticación, registro, auditoría, servicios web y motor de reglas. La plataforma también incluye un motor de solución que integra reportes, análisis y componentes de minería de datos. Su diseño modular y la arquitectura basada en “plug-in” permiten a todos o parte de la plataforma estar inmersa en aplicaciones de terceros por los usuarios finales, así como fabricantes de equipos originales. Cuenta con una interfaz de usuario donde se encuentran disponibles todos los informes, vistas OLAP y cuadros de mando integral. Cuenta además con una consola de administración que permite gestionar y supervisar la aplicación y los usuarios.

Entre sus principales ventajas se encuentran (26):

- ✓ Aplicación extensible, adaptable y configurable.
- ✓ Integración con procesos de negocio.
- ✓ Proporciona mucha libertad al usuario y los desarrolladores para crear contenidos nuevos.
- ✓ Administra y programa reportes.

Apache Tomcat 6.0.29

Tomcat (también llamado Jakarta Tomcat o Apache Tomcat) es un servidor de código abierto, desarrollado en un entorno abierto, participativo y publicado bajo la licencia Apache versión 2. Es la intención de ser una colaboración de los mejores desarrolladores de su clase en todo el mundo. Puede funcionar como servidor web por sí mismo. Dado que Tomcat fue escrito en Java, funciona en cualquier sistema operativo que disponga de la máquina virtual Java.

Es actualmente un servidor web muy utilizado a la hora de trabajar con Java en entornos web; también puede especificarse como el manejador de las peticiones de JSP (del inglés Java Server Pages) y servlets recibidas por servidores web populares, como el servidor Apache HTTP de la Fundación de software de Apache o el servidor Microsoft Internet Information Server (IIS).

Entre sus principales ventajas se encuentran (27):

- ✓ Implementado de Servlet 2.5 y JSP 2.1.
- ✓ Soporte para Unified Expression Language 2.1.
- ✓ Diseñado para funcionar en Java SE 5.0 y posteriores.

Conclusiones del capítulo

Luego de esta investigación se arribaron a las siguientes conclusiones:

- Se decidió utilizar el Modelo para el desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio en DATEC como metodología para el desarrollo de la solución.
- Se seleccionó la herramienta Visual Paradigm for UML en su versión 8.0 como herramienta de modelado, además del PostgreSQL como SGBD en su versión 9.1 y PgAdminIII como administrador del SGBD en su versión 1.14.0.
- Se decidió utilizar la herramienta informática Data Cleaner para el perfilado y limpieza de los datos en su versión 1.5.4 y PDI en su versión 4.2.1 para la implementación del proceso de ETL.
- Para el diseño de los cubos OLAP se decidió utilizar la herramienta Pentaho Schema Workbench en su versión 3.2.1 y como servidor web se propone utilizar el Apache Jakarta Tomcat en su versión 6.0.29.
- Para el análisis de la información se decidió utilizar el Pentaho BI Server en su versión 3.10 y como servidor Mondrian OLAP Server en su versión 3.0.4.

CAPÍTULO 2: Análisis y diseño del mercado de datos Extensión universitaria

Introducción

En este capítulo se realiza un análisis 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), Requisitos de Información (RI) y los Casos de Uso del Sistema (CUS). Se realiza el diseño de los subsistemas de almacenamiento, integración y visualización.

2.1 Necesidades de información

Para el desarrollo de toda herramienta informática se llevan a cabo diferentes etapas; sin lugar a duda una de las más importantes es el levantamiento de requisitos. Estos los debe proporcionar el cliente, pues es el usuario final y quien dirá si la herramienta cumple con el propósito para la que fue creada. Para el caso de los AD esta etapa de captura de requisitos se realiza luego de que el cliente provea las necesidades de información y con ellas los indicadores que le servirán de ayuda para la toma de decisiones en el futuro de su negocio.

Para dar cumplimiento a esta etapa en la presente investigación, se realizaron entrevistas con los especialistas de las distintas direcciones del área de Extensión universitaria, con el objetivo lograr un mejor entendimiento del negocio y definir las necesidades de información.

Luego de varias entrevistas, se decide clasificar la información en siete grupos principales que se corresponden con los temas de análisis que agrupan toda la información solicitada por el usuario final, ellos son:

- Dirección de Extensión cultural
- Dirección de Deporte
- Dirección de Televisión universitaria
- Dirección de Recreación
- Superación
- Proyectos socioculturales
- Alojamiento

2.1.1 Requisitos de información

Los requisitos de información son básicamente información que debe estar disponible para su consulta. Describen la información que debe almacenar el sistema para satisfacer las necesidades de los clientes (28). Atendiendo a la trazabilidad, un requisito de información es uno de los reportes que los clientes necesitan visualizar. Los mismos fueron clasificados según los temas de análisis definidos.

Vicerrectoría de Extensión universitaria

Dirección de Extensión Cultural:

RI. 1 Obtener cantidad de premios en Cultura, por facultad, por tipo de evento artístico, por manifestación, por premio artístico, por año.

Dirección de Deporte:

RI. 2 Obtener cantidad de deportistas aficionados, por deporte, por premio deportivo, por facultad, por tipo de evento deportivo, por año.

RI. 3 Obtener cantidad de premios en Deporte, por facultad, por tipo de evento deportivo, por año.

Dirección de Televisión Universitaria (TV):

RI. 4 Obtener cantidad de medios, por tipo de medios de comunicación, por año.

RI. 5 Obtener cantidad de trabajos publicados, por género literario, por tipo de medio de comunicación, por temática, por título, por año.

Dirección de Recreación:

RI. 6 Obtener cantidad de actividades centrales en la institución, por lugar, por año.

RI. 7 Obtener cantidad de actividades fuera de la UCI, por lugar, por año.

Superación:

RI. 8 Obtener cantidad de cursos por dirección, por tipos de cursos, por lugar, por año.

RI. 9 Obtener cantidad de trabajadores certificados, por dirección, por curso, por año.

Proyectos socioculturales:

RI. 10 Obtener cantidad de proyectos socioculturales, por dirección, por año.

RI. 11 Obtener cantidad de profesores, por dirección, por proyecto sociocultural, por año.

RI. 12 Obtener cantidad de estudiantes, por dirección, por proyecto sociocultural, por año.

RI. 13 Obtener cantidad de acciones realizadas, por lugar, por proyecto, por año.

RI. 14 Obtener cantidad de premios por proyecto, por área, por tipo de evento artístico, por año.

Vicerrectoría de Residencia

Alojamiento:

RI. 15 Obtener cantidad de residentes, por edificio, por apartamento, por cuarto, por dirección de residencia, por área, por género, por edad, por categoría, por año.

RI. 16 Obtener capacidad de residentes, por edificio, por apartamento, por cuarto, por año.

2.1.2 Requisitos funcionales

Los RF representan las capacidades y condiciones que el sistema debe cumplir (28) para dar respuesta a los requisitos de información. A continuación se muestran los identificados para la presente investigación:

- RF. 1 Validar usuario y contraseña.
- RF. 2 Adicionar rol.
- RF. 3 Eliminar rol.
- RF. 4 Modificar rol.
- RF. 5 Adicionar usuario.
- RF. 6 Eliminar usuario.
- RF. 7 Modificar usuario.
- RF. 8 Adicionar reporte.
- RF. 9 Eliminar reporte
- RF. 10 Modificar reporte.
- RF. 11 Extraer datos de la fuente.
- RF. 12 Realizar transformación de los datos.
- RF. 13 Cargar los datos al MD Extensión universitaria.
- RF. 14 Personalizar reporte.
- RF. 15 Exportar a otro formato.

Para más detalle consultar el artefacto 0113_Especificación de Requisitos de Software del Expediente de proyecto del MD Extensión universitaria.

2.1.3 Requisitos no funcionales

Los RNF son las propiedades o cualidades que el producto debe tener (28). A continuación se exponen los definidos para la investigación:

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 MD 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. Garantizar la persistencia de la información:

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

Se identificaron otros RNF, de soporte, restricciones de diseño, documentación de usuarios en línea, ayuda del sistema, interfaz, legales, de derecho de autor que se podrán apreciar con más detalle en el artefacto 0113_Especificación de Requisitos de Software de proyecto del MD Extensión universitaria.

2.2 Reglas del negocio

Las RN describen las políticas, normas, operaciones, definiciones y restricciones presentes en una organización y que son de vital importancia para alcanzar los objetivos (28). Durante el análisis fueron identificadas con el cliente las siguientes RN:

- RN. 1** Las direcciones están definidas de la siguiente manera: Dir. Extensión cultural, Dir. Deportes, Dir. Recreación y Dir. Televisión universitaria.
- RN. 2** El género de un edificio estará definido de la siguiente manera: H (Hembras), V (Varones), M (Mixto).
- RN. 3** Las facultades están definidas de la siguiente manera: Facultad 01, Facultad 02, Facultad 03, Facultad 04, Facultad 05, Facultad 06, Facultad 07, Facultad 08, Facultad 09, Facultad 10, Facultad 15.
- RN. 4** Las direcciones de residencia están definidas como: Dir. Residencia 1, Dir. Residencia 2, Dir. Residencia 3, Dir. Residencia de Profesores y Especialistas.

Se identificaron otras RN que se podrán apreciar con más detalle en el artefacto 0000058880-Reglas de negocio y transformación del Expediente de proyecto del MD Extensión universitaria.

2.3 Diagrama de Casos de uso del sistema

Los RF y RI identificados se agruparon en Casos de Uso Funcionales (CUF) y Casos de Uso de Información (CUI) respectivamente, representados en el diagrama de casos de uso del sistema (Figura 9), así como su relación con los actores del sistema.

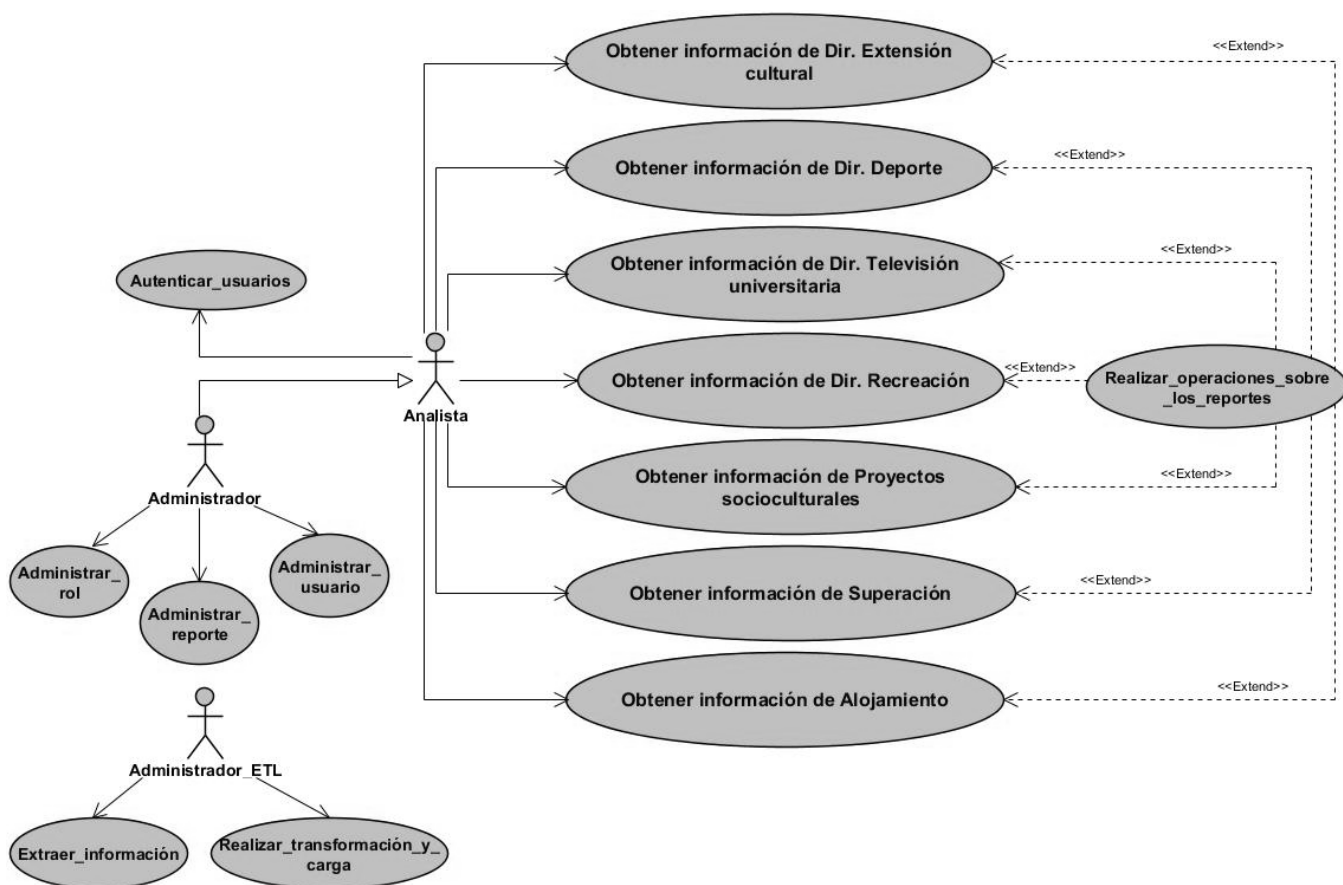


Figura 9: Diagrama de casos de uso del sistema. Relación de los actores con los casos de uso.

2.3.1 Actores

- ✓ **Administrador:** encargado de administrar los usuarios del sistema, asignarles sus respectivos roles, además de administrar dichos roles y los reportes.
- ✓ **Analista:** encargado de los casos de uso relacionados con la presentación y análisis de la información del MD.
- ✓ **Administrador de ETL:** encargado de llevar a cabo los procesos de extracción, transformación y carga de los datos del sistema fuente.

2.4 Casos de uso de información

Dependiendo de las necesidades del cliente, la información que se maneja fue agrupada mediante el criterio temas de análisis en Casos de Uso de Información (CUI). Los que se relacionan a continuación:

- CUI. 1** Obtener información de Dir. Extensión cultural: visualiza los reportes de los requisitos de información de la Dirección de Extensión cultural de la Vicerrectoría de Extensión universitaria.


- CUI. 2** Obtener información de Dir. Deporte: visualiza los reportes de los requisitos de información de la Dirección de Deporte de la Vicerrectoría de Extensión universitaria.
- CUI. 3** Obtener información de Dir. Televisión universitaria: visualiza los reportes de los requisitos de información de la Dirección de Televisión universitaria de la Vicerrectoría de Extensión universitaria.
- CUI. 4** Obtener información de Dir. Recreación: visualiza los reportes de los requisitos de información de la Dirección de Recreación de la Vicerrectoría de Extensión universitaria.
- CUI. 5** Obtener información de Proyectos socioculturales: visualiza los reportes de los requisitos de información de los Proyectos socioculturales de las distintas direcciones de la Vicerrectoría de Extensión universitaria.
- CUI. 6** Obtener información de Superación: visualiza los reportes de los requisitos de información de Superación de las distintas direcciones de la Vicerrectoría de Extensión universitaria.
- CUI. 7** Obtener información de Alojamiento: visualiza los reportes de los requisitos de información de alojamiento de la Vicerrectoría de Residencia.

A continuación en la tabla se describe el CUI **Obtener información de Dir. Extensión cultural**. Las demás descripciones de los CU se pueden encontrar en el artefacto 0114_Especificación de casos de uso de proyecto del MD Extensión universitaria.

Tabla 3: Descripción del CUI Obtener información de Dir. Extensión cultural.

Objetivo	Mostrar toda la información referente a la dirección de extensión cultural.
Actores	Especialista: (Inicia) seleccionando el AAG_Sala_Situacional_UCI.
Resumen	El caso de uso inicia cuando el actor desea hacer un análisis de la información relacionada con la Dirección de Extensión cultural desde diferentes perspectivas. El actor selecciona el reporte que desea ver, el sistema muestra la información contenida en él y las opciones de los posibles cambios que le puede hacer al reporte. Finaliza el caso de uso.
Complejidad	Media
Prioridad	Alta
Precondiciones	El usuario se autenticó correctamente. Mercado de datos poblado. Los reportes fueron creados.

Postcondiciones	Se muestran los datos.	
Flujo de eventos		
Flujo básico		
	Actor	Sistema
1	Selecciona el AAG_Sala_Situacional_UCI.	
2		Muestra AA_Vicer_Extensión_universitaria
3	Selecciona el AA_Vicer_Extensión_universitaria	
4		Muestra los libros de trabajo del AA_Vicer_Extensión_universitaria
5	Selecciona el LT_Dir_Extensión cultural.	
6		Muestra los reportes existentes en el LT_Dir_Extensión_cultural.
7	Selecciona reporte deseado	
8		Muestra la información contenida en el reporte seleccionado y brinda la posibilidad al actor de hacerle cambios al reporte para su análisis. Ir al CU Realizar_operaciones_sobre_los_reportes. Finaliza el caso de uso.
Opciones Obtener información de Dir. Extensión cultural		
Perspectivas de análisis		Posibles resultados
		Medidas
		Periodicidad
Variables de entrada relacionadas con el CU Obtener información de Dir. Extensión cultural	Variables de salida disponibles en el hecho Cultura:	Rango de tiempo en que se solicitan las variables de salida:
<ul style="list-style-type: none"> • Área • Manifestación • Año • Premio • Evento artístico 	<ul style="list-style-type: none"> • Cantidad de premios en cultura 	<ul style="list-style-type: none"> • Anual

Prototipo de interfaz de usuario		
Relaciones	CU Incluido	No aplica
	CU Extendido	CU Realizar operaciones sobre los reportes: Paso ocho del Flujo Básico.
Requisitos no funcionales	0113_Especificación de Requisitos de Software, epígrafe 3.2	
Asuntos pendientes	No aplica	

2.5 Casos de uso

Los RF fueron agrupados en CU. Los cuales son relacionados a continuación:

- CU. 1 Autenticar_usuario:** se realiza la autenticación de los usuarios en el sistema.
- CU. 2 Administrar_usuario:** se insertan, modifican o eliminan los usuarios que interactúan con el sistema.
- CU. 3 Administrar_rol:** se insertan, modifican o eliminan en el sistema.
- CU. 4 Administrar_reporte:** se eliminan, insertan o modifican los reportes que se visualizan.
- CU. 5 Realizar_operaciones_sobre_los_reportes:** se realizan las operaciones o funciones deseadas sobre los reportes.
- CU. 6 Extraer_información:** se realiza la extracción de los datos necesarios de las BD fuentes.
- CU. 7 Realizar_transformación_y_carga:** se realiza la transformación y carga de los datos necesarios para la construcción del MD.

A continuación en la tabla se describe el caso de uso **Autenticar usuario**. Las demás descripciones de los CU se pueden encontrar en el artefacto 0114_Especificación de casos de uso del expediente de proyecto del MD Extensión universitaria.

Tabla 4: Descripción del CU Autenticar usuario.

Objetivo:	Permitir que la persona se autentique y acceda al sistema si es un usuario del mismo.
Actores:	Analista, Administrador.

Resumen:	Inicia cuando el actor se conecta al sistema para acceder a las funcionalidades del mismo. Para ello debe proporcionar un usuario y contraseña. El caso de uso termina -una vez validado el usuario y contraseña- cuando el usuario logra acceder al sistema.
Complejidad:	Media
Prioridad:	Media
Precondiciones:	El sistema debe estar disponible. El usuario debe existir en la BD.
Poscondiciones:	El usuario ha sido autenticado en el sistema.
Flujo Normal de Eventos	
1. El usuario interactúa con el sistema.	
	2. El sistema muestra un formulario para introducir los datos.
3. El usuario introduce los datos.	
	4. El sistema valida los datos.
	5. El sistema le da los permisos para entrar a la aplicación. Finaliza el C.U.
Flujo Alternativo	
Acción del Actor	Respuesta del Sistema
	5.1 Si los datos no son válidos, muestra un mensaje de error y regresa al paso 2 del Flujo Normal de Eventos . Finaliza el caso de uso.

2.6 Arquitectura

Para representar 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 al usuario. La arquitectura lógica de este tipo de sistemas consta de cuatro niveles (29) (Figura 10):

- **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:** base de datos 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 y le facilitan la toma de decisiones.

El subsistema de integración se abastece de las diferentes fuentes de datos y se encarga de 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 los procesos de ETL.

El subsistema de almacenamiento se nutre de los datos cargados en los procesos de ETL. La información es almacenada en una BD en el SGBD PostgreSQL y se administra por los usuarios autorizados a través de la herramienta PgAdmin III.

El subsistema de visualización permite mostrar la información estandarizada en forma de reportes a través de la herramienta Pentaho Business Intelligence. A dicho subsistema acceden los usuarios autorizados para obtener los reportes.

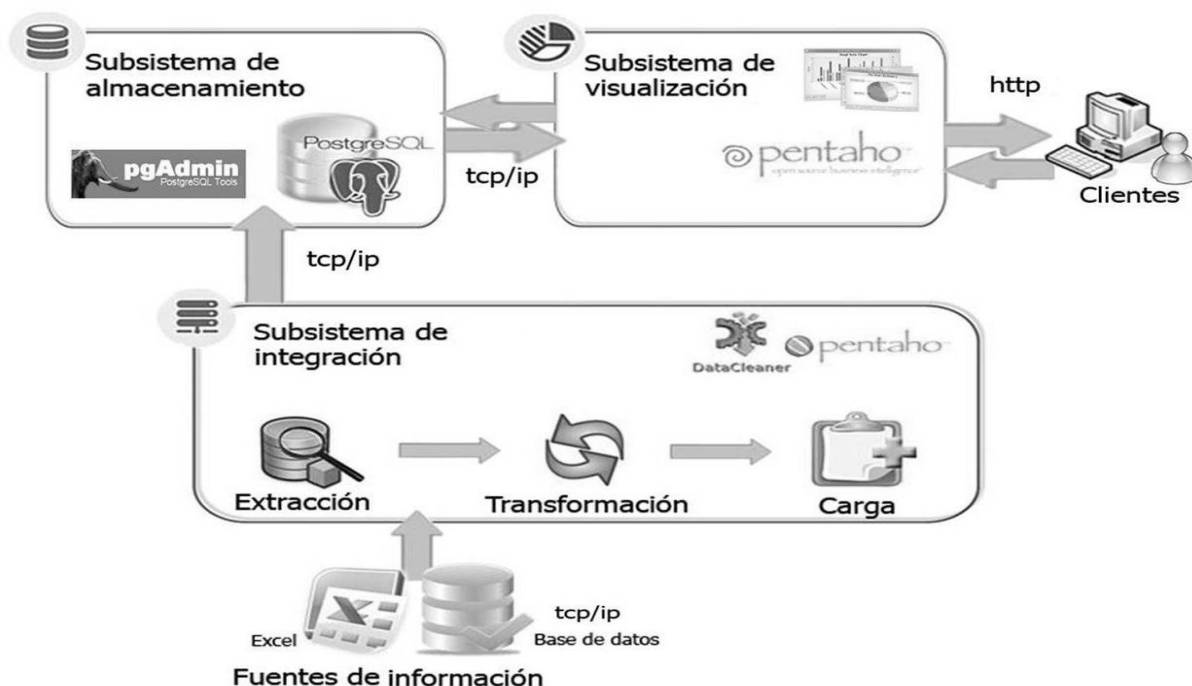


Figura 10: Arquitectura de la solución. Consta de cuatro niveles: las fuentes de datos y los subsistemas de integración, almacenamiento y visualización.

2.7 Diseño del subsistema de almacenamiento

Para el desarrollo y el correcto funcionamiento del MD, en esta etapa se realiza el modelo dimensional. Este contiene las tablas de las dimensiones y hechos identificadas en el negocio, así como las relaciones que existen entre estas (29).

2.7.1 Dimensiones

Cada dimensión está descrita por un conjunto de atributos organizados en jerarquías, atendiendo a las dependencias funcionales entre estos atributos y para realizar un análisis más detallado de los datos (10). Cada dimensión identificada en el estudio de la investigación del MD Extensión universitaria junto con los niveles jerárquicos se muestra a continuación:

D.1 Dimensión área (dim_area_extension): esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a las áreas de universidad.

Jerarquía: área

D.2 Dimensión manifestación (dim_manifestacion): esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a las manifestaciones con las que se participa en los eventos culturales.

Jerarquía: manifestación

D.3 Dimensión premios (dim_tipo_premios): esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los premios otorgados a los estudiantes.

Jerarquía: premios_tipo ->premios_subtipo

D.4 Dimensión evento (dim_evento): esta dimensión describe los de valores bajo los cuales puede clasificarse la información atendiendo a los eventos en los que participan los estudiantes.

Jerarquía: evento_tipo ->evento_nombre

D.5 Dimensión dirección (dim_direccion): esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a las direcciones en las que se divide área.

Jerarquía: direc_nombre

D.6 Dimensión curso (dim_curso): esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los cursos de superación que reciben los trabajadores.

Jerarquía: curso_tipo ->curso_nombre ->curso_duracion

D.7 Dimensión lugar (dim_lugar): esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los lugares donde se realizan actividades en general.

Jerarquía: lugar_nombre ->localización

D.8 Dimensión deporte (dim_deporte): esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los deportes en los que compiten los estudiantes.

Jerarquía: deporte_nombre

D.9 Dimensión publicación (dim_publicacion): esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a las publicaciones de los diferentes medios de comunicación.

Jerarquía: publicacion_titulo ->publicacion_tematica ->publicacion_genero_literario

D.10 Dimensión medios de comunicación (dim_medios_comunicacion): esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los medios de comunicación de la universidad.

Jerarquía: medios_tipo ->medios_nombre

D.11 Dimensión proyecto (dim_proyecto): esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los proyectos de las direcciones.

Jerarquía: proyecto_nombre

D.12 Dimensión tiempo año (dim_tiempo_anno): esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los años en lo que se almacena información.

Jerarquía: anno

D.13 Dimensión género (dim_genero): esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los géneros residenciales que existen en la Vicerrectoría de residencia.

Jerarquía: genero_tipo

D.14 Dimensión inmueble (dim_inmueble): esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los inmuebles residenciales que existen en la Vicerrectoría de residencia.

Jerarquía: direccion_residencia ->direccion_edificio ->direccion_apto ->direccion_cuarto

Para más detalle ver el artefacto Especificaciones del Modelo de Datos Dimensional del Expediente de proyecto del MD Extensión universitaria.

2.7.2 Hechos

El hecho posee datos numéricos que representa una actividad específica del negocio. Figuran el patrón de interés o el evento dentro de una empresa que necesita ser analizado. Los hechos son implícitamente definidos por la combinación de valores de las dimensiones (10). Cada hecho identificado en el estudio de la investigación del MD Extensión universitaria se muestra a continuación.

H1. Hecho cultura (hech_cultura): en este hecho se almacena toda la información relacionada con indicadores definidos en la Dirección de Extensión cultural.

H2. Hecho deportista (hech_deporte): en este hecho se almacena toda la información relacionada con indicadores definidos en la Dirección de Deporte.

H3. Hecho medios (hech_medios): en este hecho se almacena toda la información relacionada con indicadores definidos para los medios de comunicación en la Dirección de Televisión universitaria.

H4. Hecho proyecto acciones (hech_proyecto_acciones): en este hecho se almacena toda la información relacionada con los indicadores de las acciones realizadas por los proyectos socioculturales en las diferentes direcciones del área.

H5. Hecho proyectos premios (hech_proyecto_premios): en este hecho se almacena toda la información relacionada con los indicadores definidos relacionados con los premios otorgados a algunos proyectos socioculturales en las diferentes direcciones del área.

H6. Hecho proyectos socioculturales (hech_proyectos_socioculturales): en este hecho se almacena toda la información relacionada con indicadores definidos a los proyectos socioculturales que se efectúan en las diferentes direcciones del área.

H7. Hecho publicaciones tv (hech_publicaciones_tv): en este hecho se almacena toda la información relacionada con indicadores definidos a las publicaciones en los medios de la Dirección de Televisión universitaria.

H8. Hecho recreación (hech_recreacion): en este hecho se almacena toda la información relacionada con indicadores definidos a las actividades dentro y fuera de la universidad que gestiona la Dirección de Recreación.

H9. Hecho superación (hech_superacion): en este hecho se almacena toda la información relacionada con indicadores definidos a los cursos certificados en las diferentes direcciones del área.

H10. Hecho Alojamiento (hech_alojamiento): en este hecho se almacena toda la información relacionada con indicadores definidos a los residentes y la capacidad de los inmuebles en la Vicerrectoría de Residencia.

Para más detalle, ver el artefacto Plantilla Especificaciones del Modelo de Datos Dimensional que se encuentra en la carpeta Arquitectura y diseño del Expediente de proyecto del MD Extensión universitaria.

Matriz bus

La matriz bus representa la relación existente entre las dimensiones y los hechos del MD Extensión universitaria. Constituye una validación del análisis realizado evitando que exista solapamiento entre los hechos. Las columnas de la matriz representan los hechos identificados en el MD y las filas las dimensiones utilizadas. Las celdas marcadas con una X indican que la fila de dimensión está relacionada con la columna del hecho (Tabla 5).

Tabla 5: Matriz bus. Representa la relación entre los hechos y las dimensiones. Evita solapamiento entre los hechos.

Dimensiones/Hechos	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10
área extensión	X	X			X			X		X
curso									X	
deporte		X								
dirección	X	X				X			X	
evento	X	X			X					
género										X
inmueble										X
lugar				X				X	X	
manifestación	X									
medios comunicación			X				X			
proyecto				X	X	X				
publicación							X			
premios	X	X								
año	X	X	X	X	X	X	X	X	X	X

2.7.3 Modelo de datos

Una vez definidos todos los elementos descritos anteriormente, así como sus relaciones se da paso al diseño del modelo dimensional del área de Extensión universitaria (Figura 11). El principal objetivo es que los datos del negocio queden representados en una estructura lógica, evidenciando las relaciones que existen entre las tablas de hechos y las dimensiones.

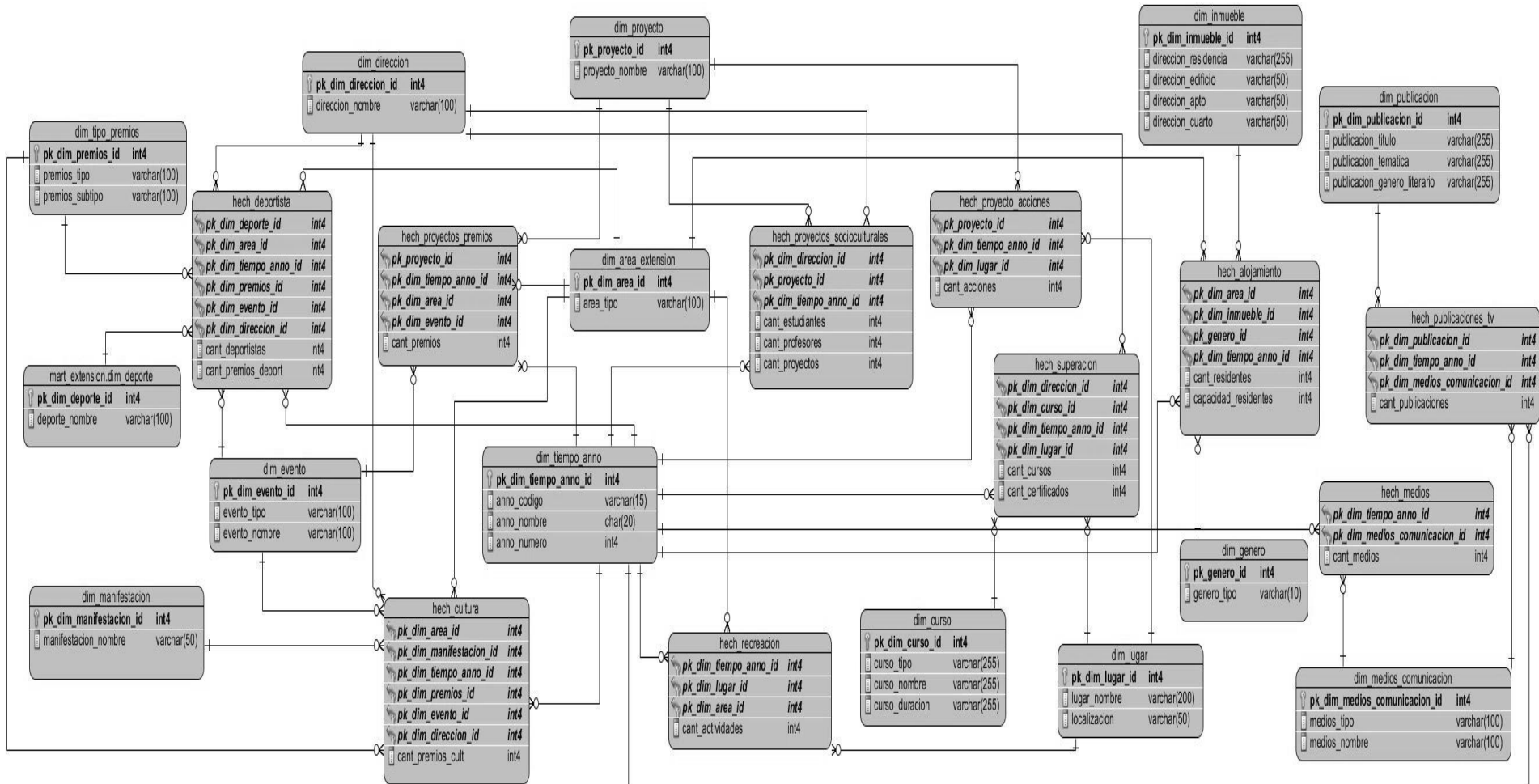


Figura 11: Modelo de datos dimensional. Representa la relación entre los hechos y las dimensiones. Evidencia una topología constelación de hechos.

2.8 Diseño del subsistema de integración

La implementación del subsistema de integración depende mucho del buen diseño que se haga, para ello se requiere efectuar el perfilado de los datos para una integración con mayor calidad y el diseño de las transformaciones (29).

2.8.1 Perfilado de los datos

El perfilado de los datos se realiza con el objetivo de identificar valores distintos, nulos y duplicados que existan en las fuentes de datos. Permite conocer el estado en que se encuentran los datos contenidos en las fuentes y definir nuevas reglas de transformación las cuales se implementarán en los procesos de ETL consistente (30)

Al realizar el análisis de los resultados arrojados por el perfilado de los datos de los ficheros fuentes, se concluye que:

- ✓ Los datos encontrados son: textos, números enteros y fechas.
- ✓ No existen valores negativos.
- ✓ Se encontraron 11 valores vacíos.
- ✓ El valor entero mínimo encontrado es 1.
- ✓ La cantidad mínima de caracteres encontrados es 0.

Para más información del perfilado de los datos de la presente investigación consultar el artefacto Perfil de los Datos del Expediente de proyecto del MD Extensión universitaria.

Diseño de los procesos de integración

Un vez diseñado el subsistema de almacenamiento y conocida la estructura de los datos se procede a realizar el diseño de las transformaciones del MD Extensión universitaria (Figuras 12-13).

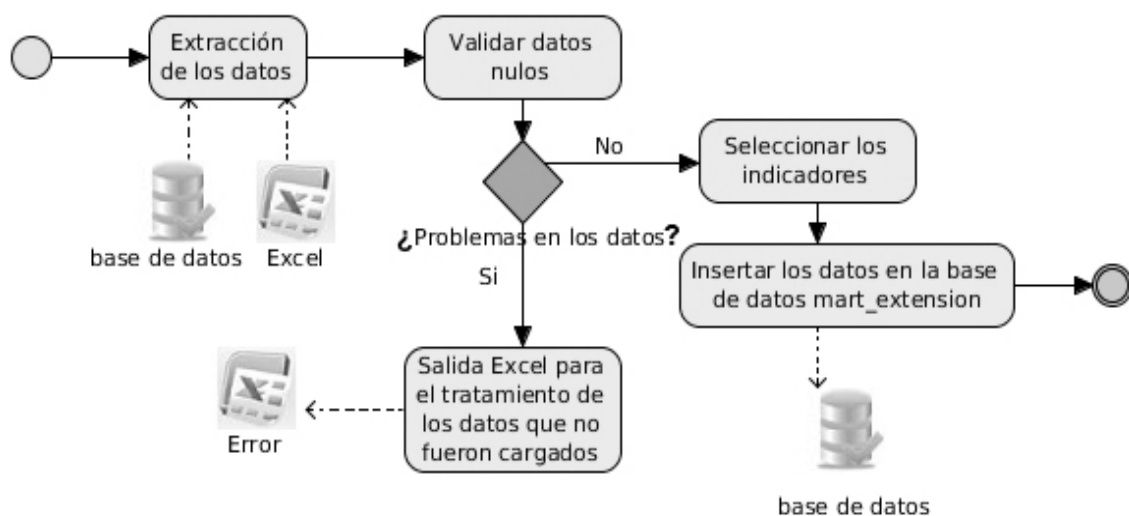


Figura 12: Diseño general de las transformaciones para la carga de las dimensiones.

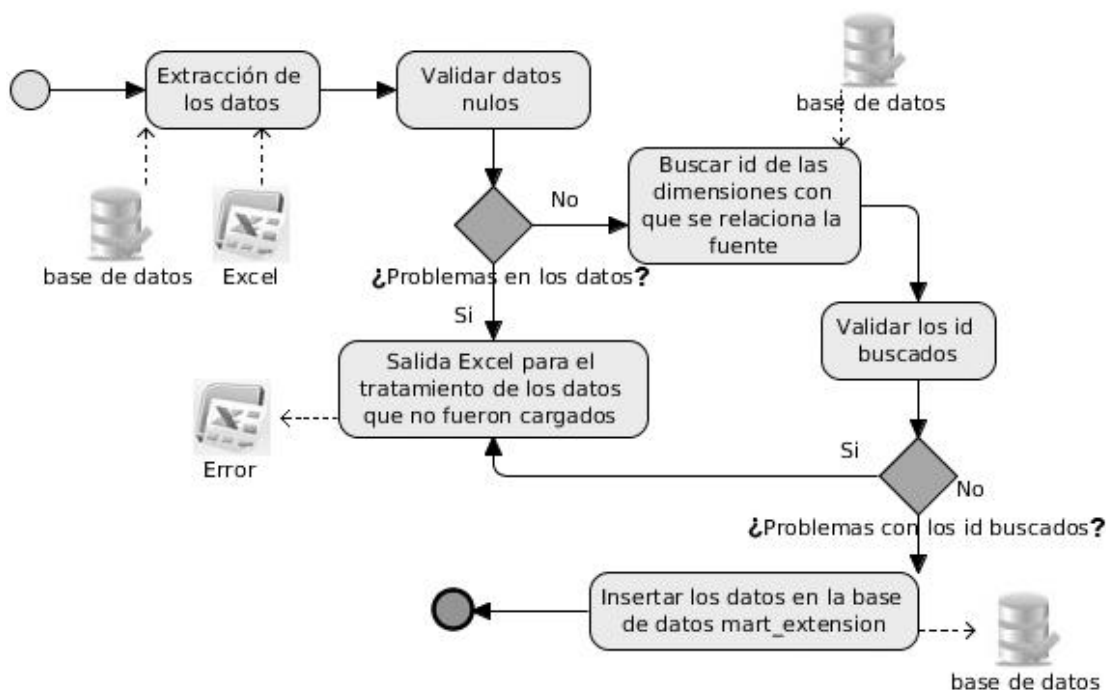


Figura 13: Diseño general de las transformaciones para la carga de las dimensiones.

2.9 Diseño del subsistema de visualización

El diseño del subsistema de visualización comprende la creación de los cubos OLAP. Estos proveen 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. En los cubos se definen las dimensiones, las medidas y los niveles de jerarquía de cada dimensión. Este subsistema tiene como funcionalidades principales generar reportes candidatos, es decir vistas de análisis que contribuyan al proceso de toma de decisiones.

Diseño de la arquitectura de información y reportes candidatos

En el Área de Análisis General Sala Situacional de la UCI (SSUCI) según la necesidad de los clientes se propone el Área de Análisis (AA) Extensión universitaria, la cual contendrá el AA **Vicer. Extensión universitaria** y el AA **Vicer. Residencia**, agrupando la información referente a los indicadores en libros de trabajo.

En la primera A.A se encuentran todos los datos referentes a las direcciones relacionadas con esta vicerrectoría: Cultura, Deporte, Televisión universitaria y Recreación, así como los proyectos socioculturales y la superación de cada una de ellas. En la segunda se localizan los datos referentes a los residentes que se encuentran en la universidad y las capacidades de los inmuebles. El mapa de navegación que se presenta a continuación constituye una representación del diseño definido para la arquitectura de información del MD (Figura 14).

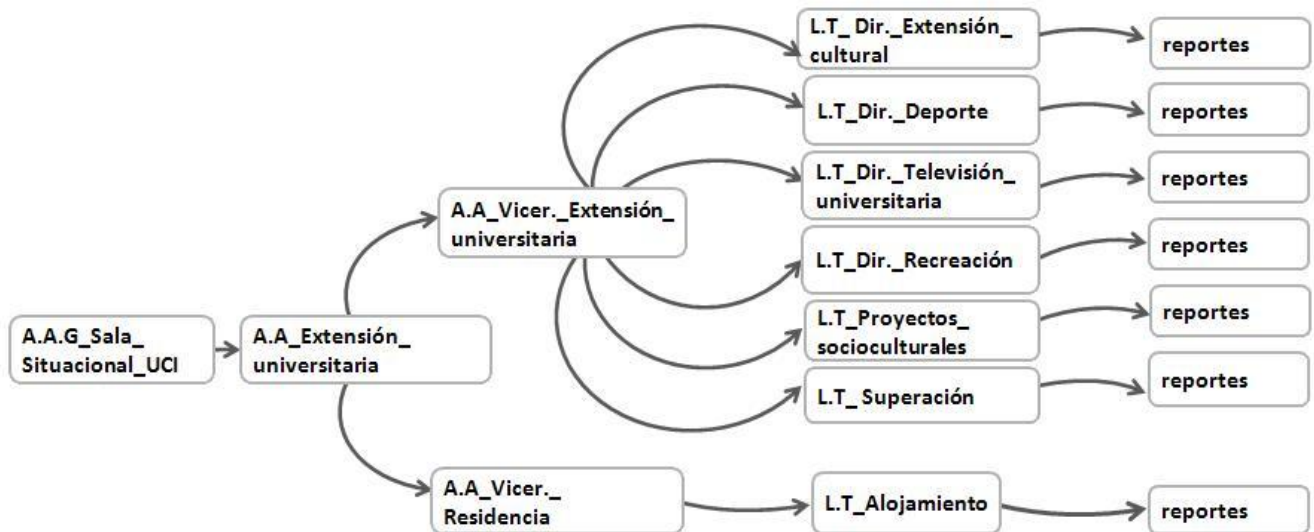


Figura 14: Diseño de la arquitectura de información del MD Extensión universitaria.

A continuación se describe la distribución de los libros de trabajo según el AA en la que está contenido:

A.A Vicer. Extensión universitaria

L.T Extensión cultural: contiene tres reportes con los que se logra realizar un análisis general de los datos en función del indicador cantidad de premios en cultura.

L.T Dir. Deporte: contiene cuatro reportes con los que se logra realizar un análisis general de los datos en función de los indicadores cantidad de deportistas y cantidad de premios en cultura.

L.T Dir. Televisión universitaria: contiene dos reportes con los que se logra realizar un análisis general de los datos en función de los indicadores cantidad de medios y cantidad de trabajos publicados.

L.T Dir. Recreación: contiene un reporte con el que se logra realizar un análisis general de los datos en función de los indicadores cantidad de actividades centrales en la institución y cantidad de actividades fuera de la UCI.

L.T Superación: contiene cuatro reportes con los que se logra realizar un análisis general de los datos en función de los indicadores cantidad de cursos y cantidad de trabajadores certificados.

L.T Proyectos socioculturales: contiene cinco reportes con los que se logra realizar un análisis general de los datos en función de los indicadores cantidad de proyectos socioculturales, cantidad de profesores y estudiantes, cantidad de acciones realizadas y cantidad de premios.

A.A Vicer. Residencia

L.T Alojamiento: contiene dos reportes con los que se logra realizar un análisis general de los datos en función de los indicadores cantidad de residentes y capacidad de residentes.

2.10 Esquema de seguridad

La seguridad de todo sistema es un elemento esencial en cuanto a la protección e integridad de los datos, por ello en el MD Extensión universitaria está determinada por los niveles de acceso al sistema. Dicha seguridad se rige por la definición de un usuario por cada rol que maneje el sistema, de esta forma se controla el acceso a la capa de visualización del MD, así como a las acciones que estos pueden realizar sobre los datos y la interacción con la BD manifestándose un mejor control del manejo de la información (31).

Seguridad en la aplicación y en la base de datos

Para la interacción de los usuarios con la aplicación y la BD se definen los siguientes roles con la relación de los permisos (Tabla 6):

Tabla 6: Permisos de los roles en la aplicación y en la BD.

Roles	Permiso	
	Lectura	Escritura
Aplicación		
Administrador	x	x
Analista	x	
BD		
Administrador BD	x	x
Administrador ETL	x	x

La herramienta Pentaho BI contiene su propia seguridad basada en la infraestructura proporcionada por el Sistema de Seguridad Acegi, dividida en cuatro áreas (37).

- **Seguridad de acceso a datos de objetos:** incluye usuarios, contraseñas, autorizaciones permitidas, recursos web y protección a datos.
- **Autenticación:** tiene que ver con el procesamiento de información interactiva de inicio de sesión (por ejemplo nombre de usuario y contraseña).
- **Autorización de recursos web (URL):** brinda protección a las URL para indicar a cada usuario si puede o no acceder a una determinada página. Esto es decidido por el administrador de recursos web, el cual le brinda a cada usuario autenticado un permiso de seguridad, delimitando las páginas a las que tiene acceso y a las que no.
- **Autorización a objetos del dominio:** en el sistema, los únicos objetos del dominio protegidos por la plataforma son los objetos de repositorio otorgados al usuario autenticado. Es responsabilidad de los objetos del dominio autorizar las operaciones solicitadas por este.

2.10.1 Políticas de respaldo y recuperación

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

- ✓ **Periodicidad de las salvas:** las salvas de toda la información contenida en la BD del área de Extensión universitaria se realizan anualmente, así lo define la organización, verificando en todo momento que exista una copia escrita de la información presente en el servidor.
- ✓ **Tablas involucradas:** las tablas que se involucran en la realización son las diez tablas de hechos identificadas en el proceso de análisis y las 14 dimensiones relacionadas.
- ✓ **Backups o salvas existentes:** En esta área no existen “backups” actualmente. Se prevee la realización de reemplazos de estos anualmente, así como también el chequeo de su estado mensualmente.

Conclusiones del capítulo

Finalizada la etapa de análisis y diseño del MD se puede concluir que:

- Se identificaron 15 RI y 19 RF agrupados en siete CUI y CU respectivamente, sirviendo de base para elaborar el diagrama de Casos de Uso del Sistema. De igual manera se identificaron nueve RN permitiendo definir las reglas de transformación que serán utilizadas en los procesos de integración
- El modelo dimensional diseñado representa las relaciones entre las diez tablas de hechos, las 14 dimensiones y las 15 medidas identificadas para el MD.
- Se diseñaron dos transformaciones, una para la carga de los hechos y otra para la carga de las dimensiones, con el fin de poblar la BD.
- Se diseñaron los 20 reportes candidatos en correspondencia con las necesidades de información de los clientes.
- Se definieron los roles y permisos para el MD, lo que contribuye a la seguridad de la aplicación.

CAPÍTULO 3: Implementación del mercado de datos Extensión universitaria

Introducción

Una vez realizado el diseño del MD y teniendo como guía la metodología utilizada, se procede a realizar la implementación del sistema. Esta nueva etapa en el proceso de desarrollo implica la implementación de cada uno de los subsistemas que conforman el MD, subsistemas de almacenamiento, integración y visualización.

3.1 Implementación del subsistema de almacenamiento

En la implementación del subsistema de almacenamiento se definen los estándares de codificación de las estructuras, lo que facilita su comprensión. Se desarrolla además la estructura física del MD.

3.1.1 Estándares de codificación

Con el objetivo de organizar como se van a denominar la estructura del AD, se formaliza un modelo, norma, patrón o un estándar de codificación. Esta acción permite a los desarrolladores entender cada una de las estructuras de los mercados.

Se propone seguir con la misma estructura para la clasificación teniendo en cuenta si una tabla es una dimensión o un hecho. A modo que en una tabla que se declara dimensión se ubican las letras “dim” como prefijo del carácter “_” y el nombre de la dimensión, ejemplo “dim_manifestacion”, por otro lado para las tablas de hechos, como prefijo se ubican las letras “hech”, ejemplo hech_cultura.

En el caso de los atributos de las dimensiones se siguió la misma política en todas. Cuando se refiere a la llave de las dimensiones se le denominó “pk_dim_dimension_id”, en caso que fuera algún código del negocio se le especificó “dimension_codigo” y alguna descripción como “dimension_descripcion”. A modo de generalizar se puede decir que todos los atributos se nombraron como “dimension_nombre”, excepto cuando la dimensión posee más de un nivel jerárquico, en este caso se siguió la estrategia de nombrarlos “nombre_de_la_jerarquía_clase”, ejemplo en la “dim_tiempo_anno” el atributo “anno_codigo”. Con relación a las medidas se nombraron de la forma “cant_medida” ejemplo “cant_premios_cult”.

Luego de este proceso se ha logrado una estandarización de la nomenclatura utilizada para la denominación de las tablas (dimensiones y hechos), atributos y medidas dentro de la BD.

3.1.2 Implementación del modelo de datos físico

Para organizar las tablas de la BD del MD Extensión universitaria se define la utilización de tres esquemas. El esquema **dimensiones** que contiene las tablas de dimensiones comunes para el AD, el esquema **mart_extension** que contiene las tablas dimensiones y hechos propias del MD Extensión universitaria y el esquema **metadatos** las tablas que permiten el control de los cambios en el MD. Se

especifican 14 tablas de dimensiones, diez tablas de hechos y tres de metadatos, formando un total de 27 tablas en la BD (Tabla 7).

Tabla 7: Relación de tablas por esquemas del MD Extensión universitaria.

Esquemas	Tablas contenidas		
dimensiones	dim_tiempo_anno		
metadatos	transformación	trabajo	gest_cambio_datos_fuentes
mart_extension	dim_area_extension	dim_manifestacion	dim_tipo_premios
	dim_evento	dim_direccion	dim_curso
	dim_lugar	dim_deporte	hech_alojamiento
	dim_publicacion	dim_medios_comunicacion	hech_superacion
	dim_proyecto	dim_genero	dim_inmueble
	hech_recreacion	hech_cultura	hech_deportista
	hech_medios	hech_proyecto_acciones	hech_proyectos_premios
	hech_proyectos_socio culturales	hech_publicaciones_tv	

3.2 Implementación del subsistema de integración

El proceso de integración de los datos consta de tres etapas fundamentales relacionadas entre sí: extracción, transformación y carga de los datos. Para identificar y corregir los problemas se realiza la limpieza de los datos que permite llenar valores ausentes y corregir errores. Una vez que los datos son transformados se cargan, poblando las dimensiones y los hechos que conforman la estructura del MD Extensión universitaria.

3.2.1 Implementación de las transformaciones

El elemento básico de diseño de los procesos ETL es la transformación, la cual está compuesta por pasos enlazados entre sí a través de los saltos. Los pasos constituyen el elemento más pequeño dentro de las transformaciones, y a través de los saltos fluye la información entre los diferentes pasos. En la presente investigación se realizó un flujo de transformación para la carga de cada una de las tablas pertenecientes al esquema mart_extension. Para las dimensiones, la transformación se realizó a partir de la carga de los indicadores de cada uno de los modelos de las fuentes de datos, estos contienen los campos que son necesarios para poblar la BD para luego pasar a las transformaciones adecuadas.

Para realizar la carga de las dimensiones (Figura 15), el primer paso que se realiza es la extracción de los datos de las fuentes (Excel Cultura). Luego se ordenan las filas, pues es necesario que los campos por los que se va a agrupar estén ordenados. El componente “Filas Únicas” permite que se seleccione solo uno de los campos que se encuentren repetidos, evitando repeticiones en la BD destino.

Para los eventos artísticos y deportivos las fuentes no muestran un campo identificativo, es decir en ninguno aparece el tipo de evento en el que se participa, por ello se declara una constante para agregar valores fijos y que en el destino se vea una columna tipo de evento (en este caso Evento artístico). La sustitución de los valores nulos es otro paso dentro de esta transformación, permitiendo sustituir por un valor que facilita hacer una búsqueda de posibles errores, por ejemplo “-1” eliminando la inconsistencia en los datos. En caso de que se encuentre en el próximo paso, el valor “-1” se desvía hacia un paso de error, donde se le presentaran a los encargados de corregirlos para una próxima ejecución de la transformación, si la validación es un éxito se da paso a la carga de las tablas dimensiones.

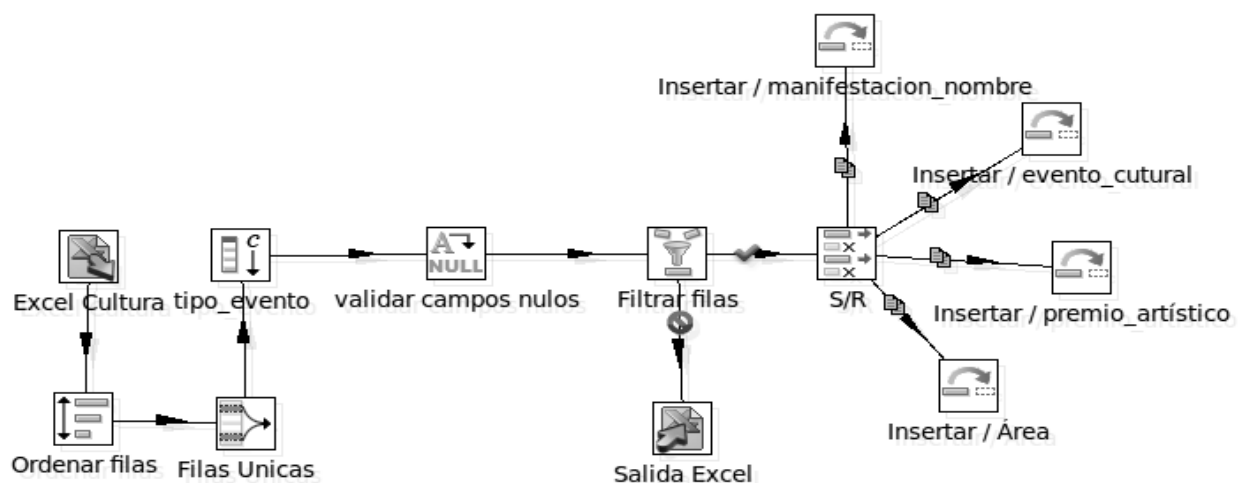


Figura 15: Transformación para cargar de las dimensiones manifestación, evento, premio y área.

La siguiente figura es un ejemplo de transformación para el hecho cultura (Figura 16), con el objetivo de visualizar la forma en que se cargaran los datos en la tabla hecho_cultura.

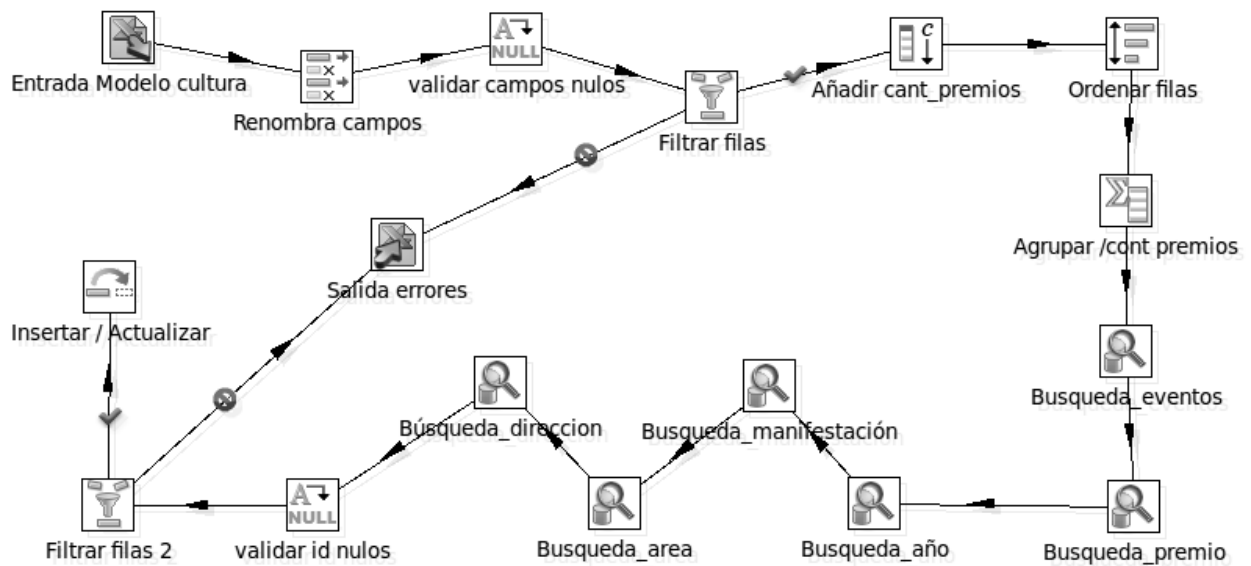


Figura 16: Transformación para realizar la carga del hecho cultura. Los componentes describen los pasos que se siguen para realizar la carga.

3.2.2 Implementación de los trabajos

Un trabajo o “Job” es un conjunto de tareas con el objetivo de realizar una acción determinada. Estos permiten ejecutar varias transformaciones o trabajos previamente diseñados y organizar una secuencia de ejecución de éstos. Los trabajos se encuentran en un nivel superior de las transformaciones.

En la presente investigación se realizaron tres trabajos, uno es destinado a ejecutar las transformaciones para la carga de las 14 dimensiones (Figura 17), otro para ejecutar las transformaciones para la carga de los diez hechos (Figura 18) y el general que ejecuta los dos trabajos anteriores.

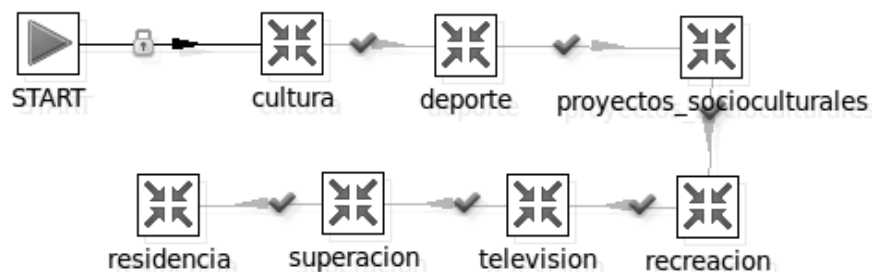


Figura 17: Trabajo que permite la ejecución de las transformaciones para la carga de las dimensiones.

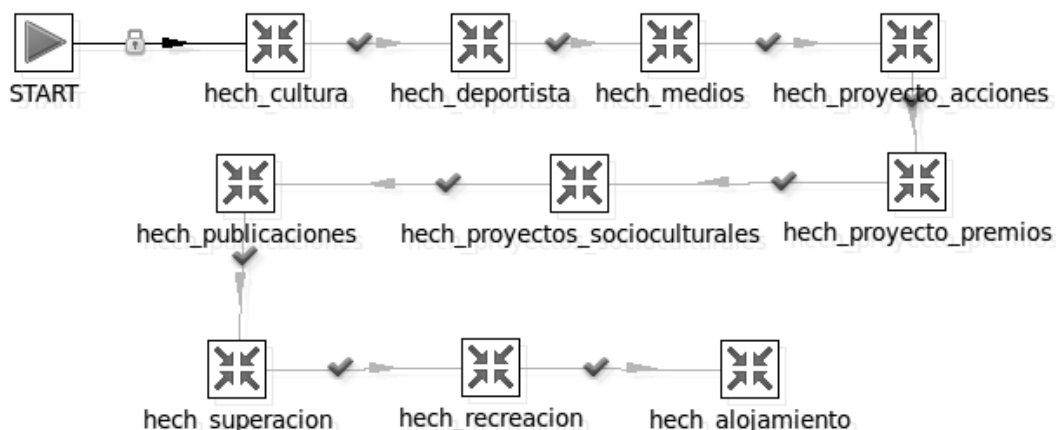


Figura 18: Trabajo que permite la ejecución de las transformaciones para la carga de los hechos.

3.3 Implementación del subsistema de visualización

Una vez realizada la carga de los datos y como parte de la implementación del subsistema de visualización, se implementan los cubos OLAP y los reportes candidatos, así como la configuración del control de acceso al sistema.

3.3.1 Implementación de los cubos OLAP

Los cubos proveen un mecanismo para buscar datos con rapidez y tiempo de respuesta, independientemente de la cantidad de datos en el cubo o la complejidad del procedimiento de búsqueda. Son subconjuntos de datos de un almacén de datos, organizado y sumado dentro de una estructura multidimensional. En los cubos OLAP la información es organizada de forma jerárquica, lo que posibilita llevar a cabo un análisis rápido (10).

En correspondencia con las tablas de hechos definidas para el área de Extensión universitaria se crearon diez cubos (Figura 19). El esquema cuenta con 14 dimensiones y 15 medidas que dan respuesta a los requisitos del cliente.

A continuación se da una breve descripción de los cubos con sus dimensiones y medidas.

El cubo **Cultura** cuenta con seis dimensiones y una medida.

El cubo **Superación** cuenta con cuatro dimensiones y dos medidas.

El cubo **Deportista** cuenta con seis dimensiones y dos medidas.

El cubo **Recreación** cuenta con tres dimensiones y una medida.

El cubo **Publicaciones_tv** cuenta con tres dimensiones y una medida.

El cubo **Proyectos_socioculturales** cuenta con tres dimensiones y tres medidas.

El cubo **Medios** cuenta con dos dimensiones y una medida.

El cubo **Proyecto_acciones** cuenta con tres dimensiones y una medida.

El cubo **Proyectos_premios** cuenta con cuatro dimensiones y una medida.

El cubo **Alojamiento** cuenta con cuatro dimensiones y dos medidas.

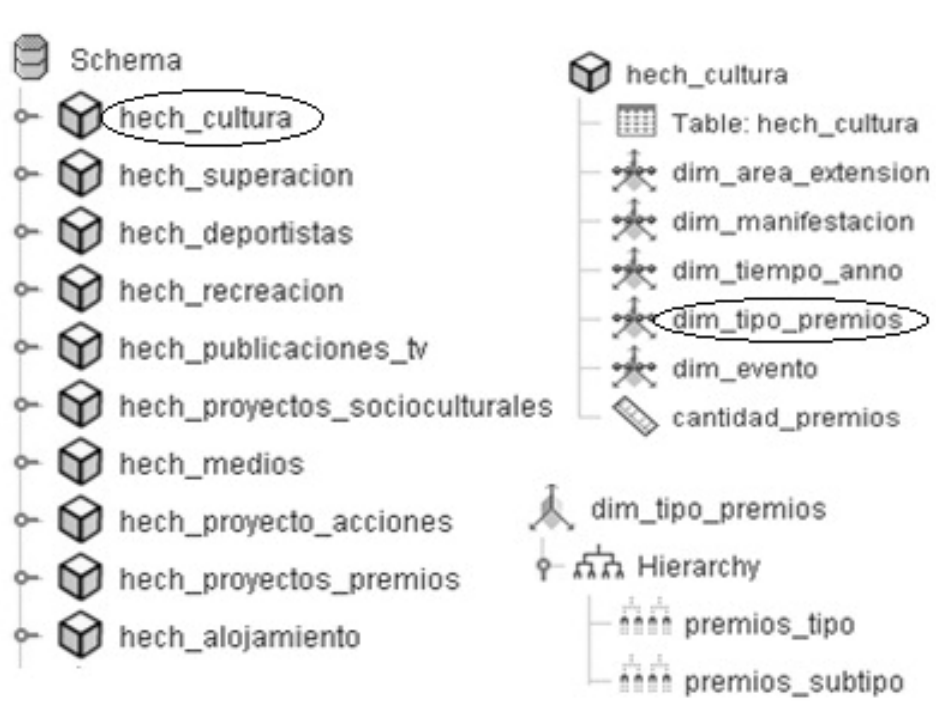


Figura 19: Representación de diez cubos del MD Extensión universitaria. El cubo hech_cultura a la derecha, contiene cinco dimensiones y una medida. La dim_tipo_premio muestra la jerarquía que la compone.

3.3.2 Arquitectura de información

La arquitectura de información contribuye a definir el entorno de análisis de las áreas de Extensión universitaria. El AA_Extension_universitaria contenida en el AAG_Sala_Situacional_UCI, está dividida en dos AA. El AA_Vicer._Extension_universitaria, contiene siete libros de trabajo y 18 reportes o tablas de salida, mientras que el AA_Vicer._Residencia contiene un LT y dos reportes, conformando un total de 20 reportes que satisfacen las necesidades del cliente (Figura 20).

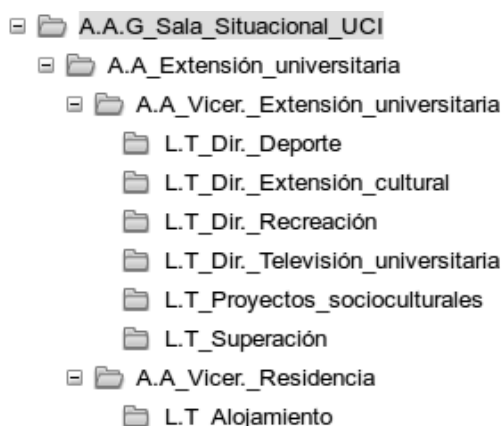


Figura 20: Arquitectura de información del MD Extensión universitaria. Las áreas de análisis contienen los libros de trabajo, en los que se ubican los reportes.

3.3.3 Implementación de los reportes candidatos

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 fuentes que recogen toda la información referente al área de Extensión universitaria. Fueron implementados en la herramienta Pentaho BI Server, mediante consultas MDX. Una consulta MDX es muy similar a una consulta SQL, ya que devuelve un conjunto de celdas, que son un subconjunto de celdas del cubo original. En la siguiente figura se muestra un ejemplo (Figura 21).

		Año							
Eventos	Manifestaciones	2004	2005	2006	2007	2008	2009	2010	2011
Festival Facultad	Danza			81	94	72	68	60	50
	Música			90	126	80	111	124	63
	Teatro			60	83	57	50	84	58
Festival UCI	Artes plásticas	9	17	40	49	66	48	27	41
	Audiovisuales					21	10	12	12
	Danza	14	18	37	20		30	24	22
	Literatura	6	15	28	36	36	26	19	21
	Música	13	24	30	26		22	43	39
	Teatro	4	13	13	20	26	21	26	25

Figura 21: Reporte correspondiente a la cantidad de premios en la dirección de Cultura, permitiendo hacer análisis por eventos y manifestaciones en un rango de tiempo.

Conclusiones del capítulo

Finalizada la etapa de implementación del MD se puede concluir que:

- Las reglas del negocio identificadas durante la fase de análisis y las reglas de transformación arrojadas por el perfilado de datos fueron aplicadas en las transformaciones de las dimensiones y los hechos.
- Se realizó la extracción, transformación y carga de los datos, quedando poblada la BD.
- Se crearon en la BD un total de 27 tablas: 14 de dimensiones, diez hechos y tres tablas de metadatos organizadas en tres esquemas, quedando implementado el modelo físico.
- Se diseñaron diez cubos OLAP, dos AA y siete libros de trabajo, así como 20 reportes candidatos para dejar implementado el subsistema de visualización.

CAPÍTULO 4: Validación del mercado de datos Extensión universitaria

Introducción

Se aplican diferentes tipos de pruebas de forma manual, tanto a la documentación como al sistema final. Las pruebas se realizan a través de las listas de chequeo y los casos de prueba, con el fin de comprobar si los resultados obtenidos a partir de la entrada de datos satisfacen los resultados esperados.

4.1 Pruebas aplicadas al MD Extensión universitaria

El desarrollo de un software involucra una serie de acciones de producción en las que existen altas posibilidades de que se cometan errores a la hora de su implementación. Se pueden manifestar desde el primer momento en que los objetivos o requerimientos especificados estén de forma errónea e imperfecta, de la misma forma en los pasos siguientes del diseño y desarrollo. Producto al interés humano de trabajar y comunicarse de forma perfecta, la elaboración o desarrollo del software ha de ir acompañado de una actividad que garantice la calidad (32).

Implementad el MD se pasa a la etapa de Pruebas. Las pruebas son acciones en las que un sistema o elemento es ejecutado bajo condiciones o requerimientos especificados. Los resultados obtenidos se captan para ser observados y registrados, realizando evaluaciones de algún aspecto en específico. Las pruebas de software son un elemento con carácter crítico en cuanto a la calidad del producto y figura como la revisión final de las especificaciones del diseño y la implementación.

La prueba es un proceso de ejecución de un programa con la intención de descubrir un error. Un buen caso de prueba es aquel que tiene una alta probabilidad de mostrar un error no descubierto hasta el momento. Una prueba tiene éxito si descubre un error no detectado hasta entonces (32).

Las experiencias de la comunidad desarrolladora han demostrado que un fallo en muchos software representa elevados costos, esto ha motivado que las empresas u organizaciones inviertan en verificar minuciosamente el resultado de cada prueba, para así poder descubrir posibles síntomas de defectos.

Para poner en marcha la calidad de la solución "MD Extensión universitaria" se utilizó el Modelo V, el mismo fue definido por CALISOFT (Centro de Calidad para Aplicaciones Tecnológicas) y se utiliza en DATEC con el fin de crear un estándar de comprobación para que los productos cumplan con las especificaciones del negocio (Figura 22)

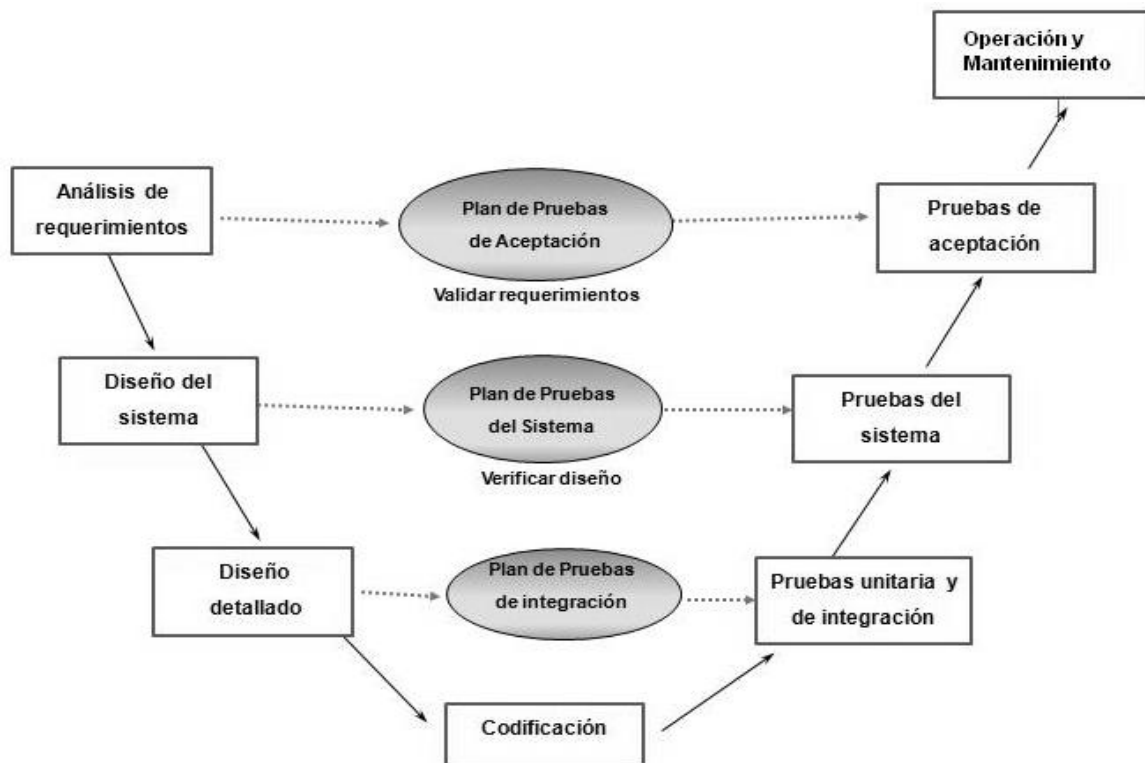


Figura 22: Modelo V. Permite realizar pruebas a lo largo del ciclo de desarrollo del software.

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 (10). A continuación se describen las pruebas realizadas al MD:

Prueba unitaria: se enfocan en un programa o un componente que desempeña una función específica que puede ser probada y que se asegura que funcione tal y como lo define la especificación del programa. Los programadores siempre prueban el código durante el desarrollo, por lo que las pruebas unitarias son realizadas solamente después de que el programador considera que el componente está libre de errores.

Prueba de integración: su objetivo es identificar errores introducidos por la combinación de programas o componentes probados unitariamente, además, verificar que las especificaciones de diseño sean alcanzadas. Componentes individuales son combinados con otros componentes para asegurar que la comunicación, enlaces y los datos compartidos ocurran apropiadamente. No son

verdaderamente pruebas de sistema porque los componentes no están implementados en el ambiente operativo.

Prueba de sistema: son usualmente conducidas para asegurar que todos los módulos trabajan como sistema, sin error. Es similar a la prueba de integración pero con un alcance mucho más amplio. Las pruebas del sistema examinan qué tan bien el sistema cumple con los requerimientos de la organización y su utilidad, seguridad y desempeño. También se realizan estas pruebas a la documentación del sistema.

Prueba de aceptación: son realizadas principalmente por los usuarios con el apoyo del equipo del proyecto. El propósito es confirmar que el sistema está terminado, que desarrolla puntualmente las necesidades de la organización y que es aceptado por los usuarios finales.

4.2 Herramientas para validar el MD Extensión universitaria

Casos de prueba

Los casos de prueba permiten la verificación de la calidad del software, son usados para la identificación de posibles fallos de implementación y para la comprobación del grado de cumplimiento de las especificaciones iniciales del sistema. En el MD Extensión universitaria se aplicaron ocho casos de prueba, uno para los procesos ETL y siete casos de pruebas basados en caso de uso, los cuales responden a los CUI en los que se encuentran agrupados los RI del MD.

Para más detalle consultar los artefactos pertenecientes a los casos de prueba, como por ejemplo DATEC-SSUCI-0111_CasoPrbasObtener información de Dir. Extensión cultural de la carpeta de 1.3 implementación y prueba del Expediente de proyecto MD Extensión universitaria.

Listas de chequeo

La lista de chequeo consta de una serie de preguntas, en forma de cuestionario, por la cual se verifica el grado de cumplimiento de reglas establecidas con un fin determinado. Se utilizará con el fin de evaluar la eficiencia del MD, midiéndose una serie de indicadores que se encuentran implicados en el proceso de la creación de la capa de integración y visualización del sistema, además de medir la calidad de los artefactos y documentos generados en la realización del producto. La lista de chequeo se divide en tres secciones fundamentales:

- **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 y redacción.

La estructura de la lista de chequeo está formada por los siguientes elementos:

- **Peso:** define si el indicador a evaluar es crítico o no. El mismo se define con una C si es crítico.
- **Indicadores a evaluar:** son los indicadores a evaluar en las secciones Estructura del documento, Semántica del documento e Indicadores definidos por la etapa.
- **Evaluación (Eval):** es la forma de evaluar el indicador en cuestión. El mismo se evalúa de 1 en caso de que exista alguna dificultad sobre el indicador y 0 en caso de que el indicador revisado no presente problemas.
- **N.P. (No Procede):** se usa para especificar que el indicador no es necesario evaluarlo en ese caso.
- **Cantidad de elementos afectados (CEA):** especifica la cantidad de errores encontrados sobre el mismo indicador.
- **Comentario (Comt):** especifica los señalamientos o sugerencias que quiera incluir la persona que aplica la lista de chequeo. Pueden o no existir señalamientos o sugerencias.

4.3 Resultados de las pruebas

Aplicadas las pruebas descritas anteriormente al MD Extensión universitaria, se arrojaron los siguientes resultados.

Pruebas unitarias y de integración

Las pruebas unitarias y de integración se aplicaron a los subsistemas que componen el MD identificándose 12 No Conformidades (NC) (Figura 23) que son detalladas a continuación.

En el subsistema de almacenamiento se identificaron tres NC con complejidad Alta.

- NC1** Cambiar el nombre de la dimensión premio por tipo o clasificación de premio.
- NC2** Unir en una sola dimensión la dimensión título, temática y género literario.
- NC3** Dividir la dimensión premio en dos dimensiones para poder ver los reportes por direcciones.

En el subsistema de integración se identificaron cinco NC con complejidad Media:

- NC4** No se utilizan variables de entorno internas para el enrutamiento de archivos .kjb.
- NC5** No tienen la implementación de la transformación para la carga de las variables de entorno definidas para las conexiones.
- NC6** Se señalaron algunos detalles en el diseño de integración de datos.
- NC7** Ponerles nombres sugerentes a los pasos utilizados.

NC8 Arreglar detalles en el diseño del subsistema integración de datos según los señalamientos.

En el subsistema de visualización se identificaron dos NC con complejidad Media:

NC9 Arreglar que se tienen datos fijos en el reporte BI.

NC10 Revisar con el cliente si se puede quitar el valor de los todos en los reportes.

Posteriormente se aplicaron cuatro listas de chequeo a los artefactos generados en el proceso de ETL, donde se identificaron cinco NC (Figura 23), de ellas una de complejidad Alta, ninguna de complejidad Media y cuatro de complejidad Baja.

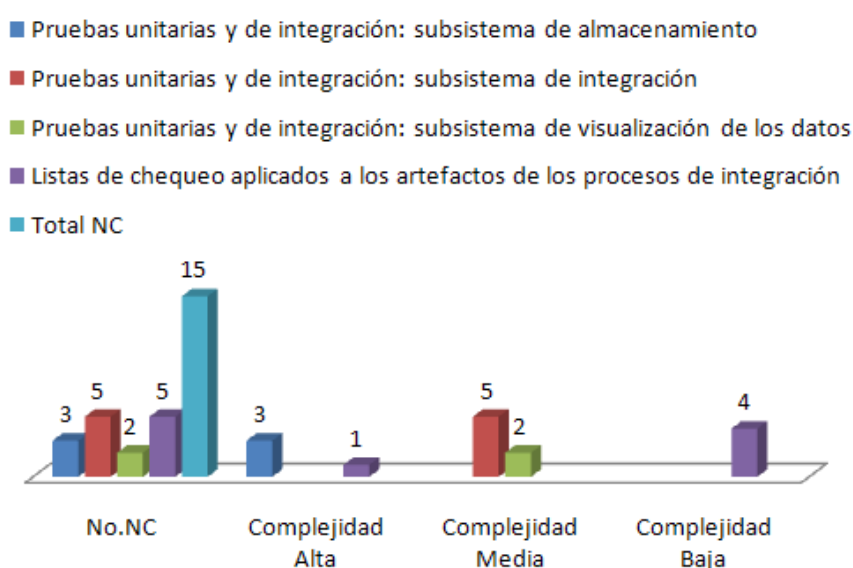


Figura 233: Resultado de la aplicación de las listas de chequeo y las pruebas unitarias y de integración según su complejidad.

Pruebas de sistema

Las pruebas se aplicaron por parte del grupo de calidad interno del departamento. Los siete casos de prueba basados en CU (Anexo 1) permitieron identificar un total de seis NC, tres de complejidad Baja, una de complejidad Media y dos de complejidad Alta (Figura 24).

NC1 Las descripciones de los escenarios del caso de prueba alojamiento está mal, debe haber una descripción de lo que realmente se hace.

NC2 No se corresponden el orden de los escenarios del caso de prueba con el orden de los reportes en la aplicación.

NC3 El flujo central descrito en los CP no se corresponde con la aplicación.

NC4 El nombre de las columnas de Variables de entrada y salida del CP están mal nombradas.

NC5 Las descripciones de las variables no están acorde a lo que muestra la aplicación.

NC6 Error ortográfico en las descripciones de los LT Extensión universitaria.

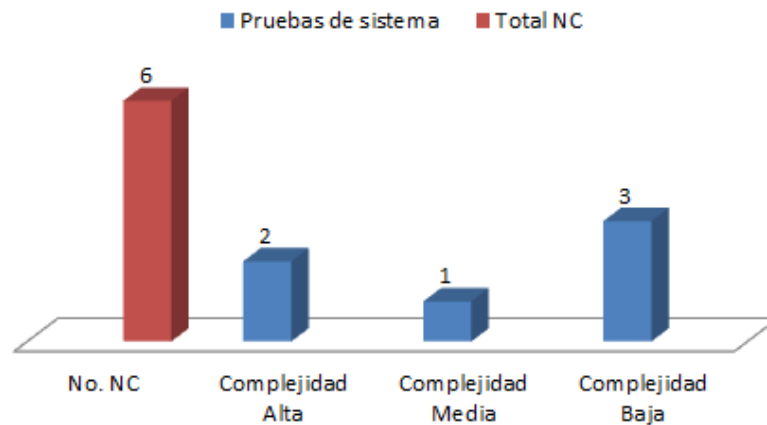


Figura 24: Resultado de las pruebas de sistema según su complejidad.

Se realizó una segunda iteración de las pruebas, que permitió comprobar que las NC fueron corregidas. Emitiéndose una carta de liberación interna por el jefe del grupo de calidad del departamento.

Pruebas de aceptación

El MD fue presentado por el equipo al especialista del área, que pudo comprobar que las funcionalidades implementadas satisfacen sus necesidades, emitiendo la carta de aceptación la cual valida que el sistema está listo para ser desplegado (Anexo 2). La siguiente gráfica muestra el resultado general de las pruebas aplicadas (Figura 25).



Figura 25: Resultado general de las pruebas aplicada al MD Extensión universitaria según su complejidad.

Conclusiones del capítulo

En este capítulo se diseñó las pruebas tanto a la documentación como al sistema del MD Extensión universitaria, se ha finalizado con la etapa de prueba por lo que se puede concluir que:

- Se diseñaron ocho casos de pruebas de ellos uno para ETL y siete basados en CU para las pruebas de sistema, con el fin de mejorar la organización y la calidad del MD.
- Mediante las pruebas unitarias y de integración se detectaron diez NC de las cuales cinco pertenecen al proceso de ETL con una complejidad Media, dos pertenecen al proceso de visualización con complejidad Media y tres relacionadas con el diseño del modelo de datos con complejidad Alta, todas resueltas satisfactoriamente.
- Se aplicaron las cuatro listas de chequeo al MD Extensión universitaria detectándose una NC en la estructura del documento, ninguna NC en elementos definidos por la metodología y dos NC en la semántica del documento, detectadas de forma general en la revisión de los artefactos y el documento de tesis. Todas se resolvieron satisfactoriamente.
- Se detectaron con las pruebas de sistema seis NC, tres de complejidad Baja, una con complejidad Media y dos con complejidad Alta.
- Mediante las pruebas de aceptación se comprobaron las funcionalidades requeridas con el especialista del área que emitió una carta de aceptación, lo cual permite desplegar el sistema.

CONCLUSIONES

Una vez concluida la investigación se puede afirmar que se desarrolló el Mercado de datos Extensión universitaria para la Sala Situacional de la UCI. Para ello:

- Se seleccionó como herramienta de modelado Visual Paradigm for UML; para el diseño de los diagramas de CUS y modelo de datos, como SGBD PostgreSQL y PgAdmin III para su administración; para el perfilado e integración de los datos el DataCleaner y el Pentaho Data Integration respectivamente. Para el diseño de los cubos OLAP se utilizó el Schema Workbench, para la implementación de los reportes el Pentaho BI Server, como motor OLAP de esta suite, Mondrian y Apache Tomcat como servidor web.
- En la etapa de análisis se identificaron siete CUI que contribuyeron a la obtención del diagrama de casos de uso del sistema. El diseño del subsistema de almacenamiento permitió obtener el modelo físico a partir del modelo lógico donde se identificaron diez hechos, 14 dimensiones y 15 medidas.
- Se diseñaron e implementaron los subsistemas que componen la solución, lo cual contribuyó a la creación de tres esquemas en el SGBD PostgreSQL que organizan las dimensiones y los hechos, lo que permitió separar las tablas de dimensiones comunes de las específicas del MD.
- Se implementaron 20 transformaciones que permitieron la carga de los datos hacia el MD, diez cubos OLAP y 20 reportes que corresponden con las necesidades del cliente.
- La aplicación de las pruebas permitió identificar 21 NC, que fueron corregidas satisfactoriamente.

RECOMENDACIONES

- Aplicar otras técnicas de análisis de información como minerías de datos para detectar patrones de comportamiento, que permitan realizar análisis más específicos de la información manejada en el MD.

REFERENCIAS BIBLIOGRÁFICAS

1. **Rodríguez Gómez, Roberto.** El significado de la extensión universitaria en el presente. *ASCUN*. [En línea] 2010. [Citado el: 19 de Enero de 2012.] <http://www.ascun.org.co/foro/iveeu/errodriguez.pdf> 31.
2. **Valadés, Diego.** Segunda Conferencia Latinoamericana de difusión cultural y extensión universitaria. Jurídica. [En línea] [Citado el: 19 de Enero de 2012.]. <http://biblio.juridicas.unam.mx/revista/pdf/DerechoComparado/13/inf/inf38.pdf>.
3. **Kimball, Ralph.** Data warehouse Lifecycle Toolkit. 2da. s.l. : John Wiley & Sons, 2010.
4. **Sevilla Berrios, Ericka Graciela.** Guía Metodológica para la definición y desarrollo de un Data Warehouse. Managua, Nicaragua : s.n., 2003.
5. **Inmon, Willian H.** Building the Data Warehouse, Technical Publish Group. 1992.
6. **Galingo González, Lic. Carlos y Pérez Vázquez, Dr. Ramiro.** Gestipolis.
7. **Giménez, Matilde Celma.** Almacenes de Datos (Data Warehouse). 5ta . Valencia : s.n., 2002, pág. 21
8. **Kimball, Ralph y Kimball, Margy Ross.** The Data Warehouse Tolkit: the Complete Guide to Dimensional Modelling. New York : John Wiley & Sons, 2002.
9. **Casales Cabrera, María Evelia.** *Data Warehouse*. Uruguay : s.n., 2009. 1263.
10. **Perez Niebla, Leonel.** Sistema de Información de Gobierno. Mercado de datos para el área de Construcción. La Habana : Universidad de las Ciencias Informáticas, 2011. Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas.
11. **Montenegro Campos, Elianys María y Noguera Libera, Maybel María.** Mercado de datos Transporte para el Sistema de Información de Gobierno. La Habana : s.n., 2011. Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas.
12. **Bernabeu, Dario.** Data Warehouse Manager. [En línea] 6 de Mayo de 2009. [Citado el: 30 de Marzo de 2012.] <http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse>.
13. **Bigatti, Ing. Cristian G.** Data Warehouse. [En línea] 1 de Diciembre de 2005. . [Citado el: 31 de Marzo de 2012.] <http://www.edutecne.utn.edu.ar/sist-gestion-II>.
14. **Alvarez Corral, Alma Lizet.** Inteligencia de Negocio. 2011, Unidad V, pág. 100.
15. **Limia Navarro, Alberto, y otros.** Metodología para el desarrollo de soluciones de almacenes de datos e inteligencia de negocios en DATEC. La Habana : Universidad de las Ciencias Informáticas, 2011.

16. **Zepeda Sánchez, Leopoldo Zenaido.** Metodología para el diseño conceptual de almacenes de datos. [En línea] 10 de Noviembre de 2010. [Citado el: 12 de Enero de 2012.] <http://www.dsic.upv.es/docs/bib-dig/tesis/etd-04032008-120043/tesisprimeraversionLeopoldo.pdf>.
17. **Luviano Nava, Jesus Cristhian.** Reporte de Investigación. 2011, 5.2 Herramientas Case. 0-807-0027. 2011.
18. Free Download Manager. [En línea] 5 de Marzo de 2007. [Citado el: 11 de Noviembre de 2011.] http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML_%28M%C%8D%29_14720_p.
19. **Mendoza Pacheco, Henry Jesus.** ¿Qué es un Sistema Gestor de Bases de Datos o SGBD? [En línea] 2010. [Citado el: 20 de Marzo de 2012.] <http://www.monografias.com/trabajos56/sistemas-bases-de-datos/sistemas-bases-de-datos.shtml>.
20. **Escobar Domínguez, René Rayde.** Tesis de Diploma: Análisis, Diseño e Implementación del Almacén de Datos para el Control del Consumo Energético en la Oficina Nacional de Estadística. La Habana : Universidad de las Ciencias Informáticas.
21. **Reingart, Mariano.** ArPUG, Grupo de usuarios PostgreSQL de Argentina. [En línea] [Citado el: 11 de Noviembre de 2011.] www.arpug.com.ar/trac/wiki/PgAdmin.
22. **Gravitar.** Características Pentaho. [En línea] [Citado el: 11 de Noviembre de 2011.] <http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho>.
23. **Vega Torres, Ing.Lilliam y Rojas Díaz, Ing.Luis.** La Inteligencia de Negocio.Su implemntación mediante la plataforma Pentaho. La Habana : s.n., 2008.
24. Portada sobre la plataforma Pentaho OpenSource Bussiness Intelligent. [En línea] 2006. [Citado el: 5 de Mayo de 2012.] <http://pentaho.almacen-datos.com>.
25. Pentaho open source bussines intelligence. [En línea] 2010. [Citado el: 10 de Mayo de 2012.] <http://mondrian.pentaho.com>.
26. **Foundation, The Apache Software.** Apache Tomcat. [En línea] 2011. [Citado el: 20 de Mayo de 2012.] <http://tomcat.apache.org>.
27. **Sherman Wood, JasperSoft.** Pentaho. [En línea] Abril de 2007. [Citado el: 15 de Noviembre de 2011.] <http://mondrian.pentaho.com/documentation/workbench.php>.
28. **Pressman, Roger S.** Ingeniería del software: un enfoque práctico MacGraw-Hill. 2005.
29. **Rivadera, Gustavo R.** La metodología de Kimball para el diseño de almacenes de datos. 2010.

30. **Kimball, Ralph y Caserta, Joe.** Data Ware house ETL Toolkit. Practical Techniques for Extracting, Cleaning Conforming and Deliverning Data. s.l. : Wiley Publishing, Inc, 2004. Vol. I. 0-764-57923-1.
31. **Romalde Ruiz, Wendy y Verdecia Ruiz, Yezenia.** Sistema de Información de Gobierno. Mercado de datos Control de energía. La Habana : s.n., 2011.
32. **EVA.** Conferencia 7. Ingeniería de Software II. Sobre la disciplina de Prueba. La Habana : Universidad de las Ciencias Informáticas, 2010-2011.

BIBLIOGRAFÍA

- **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. La Habana. 2010.
- **Inmon, Willian H.** Building the Data Warehouse, Technical Publish Group. 1992.
- **Kimball, Ralph y Kimball, Margy Ross.** The Data Warehouse Toolkit: the Complete Guide to Dimensional Modeling. New York : John Wiley & Sons, 2002.
- **Poole, John, y otros.** Common Warehouse Metamodel Developer's Guide. Indiana: Wiley Publishing, Inc, 2003. 0-471-20243-6.
- **Pulverenti, Adrián Sergio y Roldán, María Carina.** Pentaho data integration 4 Cookbook. Birmingham, UK : Packt, 2011. 978-1-849515-24-5.
- **Thomsen, Erik.** OLAP Solutions. Building Multidimensional Information Systems. s.l. : Second Edition, 2002. 0-471-40030-0.

ANEXOS

Anexo 1: Casos de prueba

Obtener información de Dir. Extensión cultural

Escenario	Descripción	Variables de Entrada	Variables de Salida	Respuesta del sistema	Flujo central
EC 1.1 Caso de prueba cultura	Muestra la cantidad de premios en la dirección de cultura por cada una de las variables de entrada.	- Premios artísticos - Manifestaciones - Áreas - Año - Eventos artístico	- Cantidad de premios en cultura	El sistema muestra todas las variables disponibles para el análisis, ubicados en las filas y las columnas que pueden ser visualizadas para cada reporte	- El usuario se autentica en el sistema - Se selecciona el A.A Extensión universitaria - Se selecciona el A.A_Vicer._Extensión universitaria - Se selecciona el L.T_Dir._Extensión_Cultural - En la parte inferior izquierda se selecciona el reporte deseado - En el área de trabajo se visualiza el reporte

Obtener información de Dir. Deporte

Escenario	Descripción	Variables de Entrada	Variables de Salida	Respuesta del sistema	Flujo central
EC 1.2 Caso de prueba deportista	Muestra la cantidad de deportistas y los premios por cada una de las variables de entrada.	- Premios Deportivos - Eventos deportivos - Deportes - Áreas - Año	- Cantidad de deportistas en la Dirección de deporte - Cantidad de premios en deporte	El sistema muestra todas las variables disponibles para el análisis, ubicados en las filas y las columnas que pueden ser visualizadas para cada reporte	- El usuario se autentica en el sistema - Se selecciona el A.A Extensión universitaria - Se selecciona el A.A_Vicer._Extensión universitaria - Se selecciona el L.T_Dir._Deporte - En la parte inferior izquierda se selecciona el reporte deseado - En el área de trabajo se visualiza el reporte

Anexo 2: Carta de aceptación

ACTA DE ACEPTACIÓN

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

De una parte, la Vicerrectoría de Extensión Universitaria, representado en este acto por Damaris Cruz Amarán, 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 **Yaneisy Pedraza González**, 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: **Extensión universitaria.**
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



Damaris Cruz Amarán

Por DATEC



Yaneisy Pedraza González

GLOSARIO

Área de Análisis: agrupación de información según su propósito, aunque el criterio depende de las necesidades de la institución o empresa donde se aplica el sistema. Permite restringir el número de usuarios que acceden a los datos.

Base de datos relacional: es una base de datos que cumple con el modelo relacional, el cual es el modelo más utilizado en la actualidad para implementar bases de datos ya planificadas. Permiten establecer relaciones entre los datos (que están guardados en tablas), y a través de ellas relacionar los datos de ambas tablas, de ahí proviene el nombre de: "Modelo Relacional".

Cubo: colección de dimensiones y medidas en un área temática particular.

Jerarquía: implica una organización de niveles dentro de una dimensión, donde cada nivel representa el total agregado de los datos del nivel inferior.

Libro de Trabajo: estructura organizativa que agrupa los reportes generados dentro de las Áreas de Análisis. Puede ser creado teniendo en cuenta criterios que permitan organizar la información: emisor de los reportes, receptor del reporte, contenido, entre otros.

TCP/IP: son las siglas de Protocolo de Control de Transmisión/Protocolo de Internet (por el inglés Transmission Control Protocol/Internet Protocol), un sistema de protocolos que hacen posibles servicios Telnet, FTP, E-mail, y otros, entre ordenadores que no pertenecen a la misma red. TCP garantiza la entrega de datos y que los paquetes sean entregados en el mismo orden en el cual fueron enviados. IP utiliza direcciones que son series de cuatro números octetos (byte) con un formato de punto decimal.