

Universidad de las Ciencias Informáticas

FACULTAD 6



Título: Mercado de datos Cooperación internacional para la Sala
Situacional de la Universidad de las Ciencias Informáticas.

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

Autoras: Maria Teresa Rosales González

Yisel de los Ángeles González Pompa

Tutores: Ing. Irina Collazo Ávila

Ing. Leonel Pérez Nieblas

La Habana, junio de 2012

“Año 54 de la Revolución”

*La posibilidad de realizar un sueño es lo que hace que la
vida sea interesante*

Paulo Coelho

DECLARACIÓN DE AUTORÍA

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

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

Maria Teresa Rosales González

Yisel de los Ángeles González Pompa

Firma del Autor

Firma del Autor

Ing. Irina Collazo Ávila

Ing. Leonel Pérez Nieblas

Firma del Tutor

Firma del Tutor

Datos de contacto

Tutores:

Tutor: Ing. Irina Collazo Ávila

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: Instructor

Años de experiencia en el tema: 5

Años de graduado: 5

Correo Electrónico: icollazo@uci.cu

Tutor: Ing. Leonel Pérez Nieblas

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: Instructor Recién Graduado

Años de experiencia en el tema: 1

Años de graduado: 1

Correo Electrónico: lnieblas@uci.cu

AGRADECIMIENTOS

De Teresa:

No sé por cuánto tiempo he buscado las palabras perfectas para este momento, pero no hay palabras perfectas, ni justas, ni imperfectas para decir lo que siento.

Agradecerle a mi mamita linda del alma por ser mi protección, mi sostén, frente a todo mi mejor opción. Porque siempre has tenido la razón, por tu inmenso amor, porque todo lo que soy y seré te lo debo a ti. Te adoro.

A mi abuelita Geo, mi ángel, por no querer estar en el cielo que es tu lugar, sino al lado mío.

A mi virgencita por no fallarme nunca.

A mi papito por pedir una niña que se pareciera a él y quererme por encima de todo, por ser un hombre mejor todos los días para tus hijos y tu familia.

A mi mimita por llorar conmigo y vivir para mí, por malcriarme tanto que se me hace imposible no ser tu niña toda la vida.

A mi abuela Maruza y mi abuelo Hugo por todo su amor, por estar siempre ahí para mí.

A mi mami y mi tío Oscarito por criarme, consentirme y ser imprescindibles para mí.

A madrina por ser la mejor madrina del mundo, a Gy por ser como una madre y una amiga, a mi tía Nana y mi tía Vita por quererme tanto, a la Giselle por ser la tía más dulce y buena que alguien pudiera tener, a mi tía Lázara por siempre estar al tanto de mí y ser la alegría de la casa, a mi tío Huguito por quererme como su hija. A todos mis primitos por dejarme ser su ejemplo.

A Nelson mi otro papá, por quererme, por todo lo que siempre ha hecho por mí, por nunca olvidarme.

A Libis, Susi y mi hermanito bello, por ser mi apoyo durante todos estos años.

A Carmen por ser para mi mamá y para mi más de lo que podemos agradecer.

A Aleidys, por ser mi hermanita sin importar el momento y a su familia, por apoyarme tanto.

A Haylín por ser mi hermana mayor en la UCI, mi consejera.

A Sara, Leidys, Rusma, Fanny y Yadira, por ser mi familia donde esperé no tenerla y continuar siéndolo para toda la vida.

A Gladys por ser tan buena, por dejarme entrar y convertirse en mi hermana.

A Pompa por ser mi cómplice, mi hermana, por aguantarme en las buenas y malas que hemos pasado con la tesis y fuera de ella.

A Irina por su sentido del humor, por ser más que una tutora una amiga y a Leonel por apoyarnos siempre.

A todos los profesores que he tenido por ayudarme a formarme como persona y profesional.

A Silvio, Héctor, Rey, Yulí, Elenis, Luis Carlos, Dasley, Carlos Arce, los muchachos de CDI, a todas mis amistades, por todo su cariño, por todo lo que hemos pasado juntos siempre los llevaré en el corazón.

De Yisel:

A la persona más importante en mi vida, mi mamá, a quien le debo todo lo que soy, por ser mi luz en la oscuridad, por darme todo sin pedir nada a cambio, por creer en mí y apoyarme en todo momento, las palabras no bastan para expresar lo agradecida que estoy de la vida por darme una madre como tú.

A mi hermana Delita, que la quiero con la vida, por cuidar de mi mamá todo este tiempo que he estado lejos, por preocuparse siempre por mí, espero ser un buen ejemplo en su vida personal y profesional.

A mi papá que aunque no está físicamente conmigo sé que donde se encuentre está muy orgulloso de lo que he logrado, por preocuparse por mí y guiarme hasta donde la vida se lo permitió, ojalá estuvieras conmigo en este momento tan especial.

A mi padrastro por ser mi segundo padre y quererme siempre como su propia hija.

A mi novio Luis Carlos por todos estos años que hemos compartido, por quererme, comprenderme, ayudarme y mucho más de lo que se pueda expresar con palabras.

A toda mi familia, que es bastante larga, por aconsejarme y preocuparse por mí, especialmente a mi abuela a quien adoro y a mis primas Liset y Yole.

A mi hermana Marlenis y mi sobrina Yilian, por apoyarme incondicionalmente en todo que han podido y mucho más durante estos 5 años de carrera, por tapar ese vacío que deja el estar lejos de la familia durante tanto tiempo.

A mi amiga Leidys por estar ahí cuando la he necesitado todos estos años de universidad, por apoyarme en todo momento, eres una hermana para mí.

A Teresa mi otra gran amiga, más que una amiga una hermana, por todo lo que hemos compartido durante este tiempo, nos merecemos ser ingenieras por todo el empeño que le hemos puesto. Nunca las olvidaré ni a ti ni a Leidys y las voy a extrañar muchísimo.

A Maiquel por preocuparse por mí en todo momento, por apoyarme cuando más lo necesité.

A Irina y Leonel, por estar siempre disponibles cuando los necesitamos, por sus enseñanzas, a Irina en especial por tener siempre una sonrisa en la cara incluso en los peores momentos, por sus comentarios siempre para mejor, por ser ambos un ejemplo de profesionales y seres humanos.

A mis amistades, en especial a Bertica y Yaira por confiar en mí y estar a mi lado en las buenas y en las malas, a Dasley, Liset, Aniuska, Yaimet y Yisel Rodríguez a ustedes nunca las olvidaré.

A mis compañeros del 6507 por todos los momentos que compartimos, en especial a las Maylén, Yenei, Yisel Cobas, Anabel, Yuri y Karla.

DEDICATORIA

A nuestras madres por su entrega incondicional.

A toda la familia por su apoyo en todo momento.

A todos los profesores que han contribuido con nuestra formación.

RESUMEN

La Universidad de las Ciencias Informáticas es un centro de estudios universitarios que cuenta con varias direcciones para garantizar su correcto funcionamiento. Una de estas direcciones es Cooperación internacional, que es la encargada de gestionar todos los procesos de intercambio entre intelectuales de todas partes del mundo con los profesionales de la institución. La información de esta área es almacenada en diversos formatos y en períodos variables, convirtiéndose en una tarea compleja su unificación, manejo e interacción. Esta es una de las razones que hacen que la realización de los reportes del área sea de forma manual y en tiempos prolongados de solicitud, lo que trae consigo una demora en el análisis y estudio de la totalidad de la información. Este conjunto de factores trae como consecuencia, que los directivos de la universidad no cuenten con las herramientas requeridas para ofrecer análisis de los datos que respalden las decisiones tomadas. En la presente investigación se realiza una propuesta de solución basada en la implementación de un mercado de datos que contribuya al proceso de toma de decisiones. Para ello se desarrolló el análisis, diseño, implementación y prueba de la aplicación que constituye el resultado principal de la investigación desarrollada. Conjuntamente se emplearon varias herramientas y una metodología para guiar el desarrollo de la solución, obteniéndose un mercado de datos poblado y funcional, con una capa de inteligencia de negocio que brinda vistas de análisis actualizadas, facilitándole a los especialistas de la Universidad de las Ciencias Informáticas un mejor estudio de la información.

Palabras claves: Cooperación internacional, inteligencia de negocio, Mercado de datos.

TABLA DE CONTENIDO

AGRADECIMIENTOS	IV
DEDICATORIA.....	VI
RESUMEN.....	VII
INTRODUCCIÓN.....	1
CAPÍTULO 1: FUNDAMENTOS TEÓRICOS SOBRE EL DESARROLLO DE LOS ALMACENES DE DATOS.....	5
1.1 Cooperación internacional.....	5
1.2 Almacenes de datos	5
1.3 Mercado de datos.....	6
1.4 Modelo multidimensional	7
1.5 Procesamiento Analítico en Línea (OLAP).....	8
1.5.1 Implementaciones de OLAP	9
1.6 Procesos básicos de un almacén de datos.....	10
1.6.1 Integración de datos	10
1.6.2 Inteligencia de negocio	11
1.7 Metodología de desarrollo	11
1.8 Herramientas seleccionadas para la solución.....	13
1.8.1 Herramienta de modelado	13
1.8.2 Sistema gestor de bases de datos.....	14
1.8.3 Herramientas para la integración de datos.....	14
1.8.4 Herramientas para inteligencia de negocio	15
1.8.5 Herramienta para soporte de aplicaciones web	16
1.8.6 Herramienta para el control de versiones.....	17
1.9 Conclusiones del capítulo.....	17
CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS COOPERACIÓN INTERNACIONAL PARA LA SALA SITUACIONAL DE LA UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS.	18
2.1 Análisis del MD Cooperación internacional.....	18
2.2 Definición del negocio.....	18
2.3 Reglas del negocio	18
2.4 Necesidades de los usuarios	19
2.3.1 Requisitos de información.....	20
2.3.2 Requisitos funcionales	20
2.3.3 Requisitos no funcionales	21

2.3.4 Actores del sistema	23
2.3.5 Diagrama de Casos de uso del sistema.....	23
2.3.6 Especificación de casos de uso del sistema	24
2.4 Arquitectura del mercado de datos	26
2.4.1 Diseño del subsistema de almacenamiento	27
2.4.2 Diseño del subsistema de integración de datos	34
2.4.3 Diseño del subsistema de visualización	37
2.5 Política de respaldo y recuperación	39
2.6 Esquema de seguridad.....	39
2.7 Conclusiones del capítulo.....	41
CAPÍTULO 3: IMPLEMENTACIÓN DEL MERCADO DE DATOS COOPERACIÓN INTERNACIONAL PARA LA SALA SITUACIONAL DE LA UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS	42
3.1 Implementación del subsistema de almacenamiento.....	42
3.1.1 Modelo físico	42
3.1.2 Estándares de codificación	44
3.2 Implementación del subsistema de integración de datos	45
3.2.1 Gestión del cambio en las dimensiones.....	45
3.2.2 Gestión de los metadatos del proceso de integración de datos	47
3.2.3 Captura de los cambios en los datos	47
3.2.4 Implementación de los hechos.....	48
3.2.5 Implementación de los trabajos	48
3.3 Implementación del subsistema de visualización.....	49
3.3.1 Implementación de los cubos OLAP	49
3.3.2 Implementación de los reportes	51
3.4 Conclusiones del capítulo.....	52
CAPÍTULO 4: REALIZAR PRUEBAS AL MERCADO DE DATOS COOPERACIÓN INTERNACIONAL PARA LA SALA SITUACIONAL DE LA UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS	53
4.1 Pruebas de software.....	53
4.2 Modelo V	53
4.3 Herramientas para la aplicación de las pruebas	54
4.3.1 Aplicación de los casos de prueba por caso de uso.....	54
4.3.2 Aplicación de los casos de pruebas por las reglas de transformación.....	55
4.3.3 Aplicación de las listas de chequeo	56
4.3.4 Calidad de datos.....	58

4.4 Evaluación de los resultados	59
4.5 Conclusiones del capítulo.....	60
CONCLUSIONES GENERALES	61
RECOMENDACIONES.....	62
REFERENCIAS BIBLIOGRÁFICAS.....	63
BIBLIOGRAFÍA.....	65
GLOSARIO DE TÉRMINOS	67

Índice de Tablas

Tabla 1 Actores del sistema	23
Tabla 2 Descripción CU “Presentar información de colaboración internacional”	25
Tabla 3 Descripción de las dimensiones	28
Tabla 4 Descripción de los hechos	30
Tabla 5 Descripción de las medidas.....	31
Tabla 6 Reporte candidato “Total de acuerdos de colaboración por directivo UCI firmante”	38
Tabla 7 Usuarios en la BD	40
Tabla 8 Usuarios en la aplicación	41
Tabla 9 Esquemas y tablas del MD.....	42
Tabla 10 Estándares de codificación del MD Cooperación internacional	44
Tabla 11 Caso de prueba “Presentar información de colaboración internacional”	55
Tabla 12 Caso de prueba para regla de transformación	55
Tabla 13 Aplicación de las listas de chequeo a los artefactos de ETL.....	56
Tabla 14 Lista de chequeo realizada al artefacto “Registro del sistema fuente”	57

Índice de Figuras

Figura 1 Arquitectura descendente	7
Figura 2 Arquitectura ascendente	7
Figura 3 DCUS	24
Figura 4 Arquitectura del MD	27
Figura 5 Matriz dimensional	32
Figura 6 Ejemplo del esquema en estrella	33
Figura 7 Ejemplo del esquema copo de nieve.....	33
Figura 8 Modelo de datos	34
Figura 9 Arquitectura de integración	35
Figura 10 Distribución de la fuente por tipos de datos.....	35
Figura 11 Calidad de los datos.....	35
Figura 12 Diseño del proceso de integración de datos.....	36
Figura 13 Arquitectura de información	37
Figura 14 Seguridad en el proceso de ETL.....	40
Figura 15 Estructura de la BD.....	44
Figura 16 Componente Insertar/ Actualizar	46
Figura 17 Transformación hecho colaboración internacional	48
Figura 18 Trabajo general.....	49
Figura 19 Jerarquía de la dimensión firma uci.....	49
Figura 20 Nivel de la jerarquía firma uci.....	50
Figura 21 Cubo hecho colaboración internacional.....	50
Figura 22 Medida hecho colaboración internacional	50
Figura 23 Diseño de los cubos OLAP	51
Figura 24 Vista del reporte “Total de acuerdos de colaboración internacional por país”	52
Figura 25 Modelo V.....	54
Figura 26 Calidad de los datos.....	59
Figura 27 Comportamiento de los indicadores medidos en las pruebas unitarias	59
Figura 28 Comportamiento de los indicadores medidos en la lista de chequeo	60

INTRODUCCIÓN

El avance alcanzado por las Tecnologías de la Informática y las Comunicaciones (TIC) ha influido en el desarrollo del ser humano. Rompiendo paradigmas, jugando un papel significativo para las economías de los países destacados en este campo. Un ejemplo del impacto de la informática, es el hecho, de que la mayor parte de las empresas que producen alta tecnología surgieron hace menos de dos décadas, dando paso a la aparición de la denominada economía de la información, que abarca la producción de software, la prestación de servicios y el soporte para los canales de comunicación, creando a su vez nuevos problemas y oportunidades.

Continuamente el desarrollo de software amplía su alcance y se gestiona con mayor profundidad. Siendo hoy en día un proceso de simple y rápida ejecución para los usuarios de sistemas informáticos procesar operaciones muy complejas, donde se manejen enormes volúmenes de información, costosas en tiempo y recursos. Relacionado a estos sistemas y dentro de su entorno surgen las bases de datos (BD), respondiendo a necesidades marcadas en el almacenamiento y acceso a los datos.

Paralelo a ello, la necesidad de extraer información útil a partir de datos históricos se ve reflejada de forma más tangible y valiosa en el mundo de los negocios. Debido a que constituye uno de los puntos de partida para la toma de decisiones que garanticen el éxito futuro de las empresas. Con el objetivo de brindarle soporte a este proceso surgen los sistemas de apoyo a la toma de decisiones, creados para medir y controlar el desarrollo de las variables importantes del negocio, buscando identificar, proyectar y predecir tendencias a partir de los datos acumulados.

Los sistemas de apoyo a la toma de decisiones no se han visto excluidos del proceso de informatización de la sociedad que hace ya varios años viene desarrollándose en Cuba. Un eslabón fundamental de este proceso fue la creación de la Universidad de las Ciencias Informáticas (UCI), la cual cuenta con el Centro de Tecnologías de Gestión de Datos (DATEC) que se especializa en el desarrollo de soluciones, productos y servicios relacionados con las tecnologías de gestión de datos.

DATEC se encuentra colaborando con el proceso de informatización que se está llevando a cabo en la universidad, entre sus proyectos se encuentra el denominado Sala Situacional UCI. Este abarca varias áreas de la UCI, una de ellas es Cooperación internacional la cual para su funcionamiento se divide internamente en cuatro departamentos: Trámites, Colaboración internacional, Visitas y Economía, dentro de los cuales se manejan una serie de datos referentes a las entradas y salidas al país de trabajadores y estudiantes, teniendo en cuenta motivos de viajes, fechas, verificaciones entre otros elementos de vital importancia en este contexto. Además se gestiona lo referente a la atención de las personalidades que acuden a la UCI por diversos motivos. Se administran los asuntos relacionados con los acuerdos de colaboración internacional de la universidad y las becas en el extranjero de los

profesores. La información en detalle es almacenada en diversos formatos y en períodos variables, convirtiéndose en una tarea compleja su unificación, manejo e interacción. Debido a que tres de sus departamentos almacenan sus datos en ficheros excel y el departamento de Trámites cuenta con un sistema para gestión diaria de los datos. Por lo cual la realización de los reportes del área es de forma manual y con un gran costo de tiempo, lo que trae consigo una demora en el análisis y estudio de la totalidad de la información. Este conjunto de factores trae como consecuencia que los directivos de la UCI no cuenten con las herramientas requeridas para ofrecer datos que respalden las decisiones tomadas, lo que puede incidir de manera negativa en la ejecución de sus procesos regulares y en la prevención de comportamientos futuros que pueden afectar el adecuado funcionamiento de la misma.

Por todo lo anteriormente expuesto surge como **problema de la investigación**: ¿Cómo contribuir al proceso de toma de decisiones en el área de Cooperación Internacional de la Sala Situacional de la Universidad de las Ciencias Informáticas? Definiéndose como **objeto de estudio**: los almacenes de datos, enmarcándose en el **campo de acción**: Mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas.

Para darle solución al problema antes descrito se define como **objetivo general**: Desarrollar el mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas, que contribuya a la toma de decisiones.

Del cual se desglosan los siguientes **objetivos específicos**:

1. Fundamentar la selección de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.
2. Realizar el análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas.
3. Implementar el mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas.
4. Validar el mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas.

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

1. Fundamentación de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.
2. Levantamiento de requisitos.

3. Análisis de los sistemas fuentes.
4. Documentación de requisitos y casos de uso del mercado de datos.
5. Diseño de los casos de pruebas.
6. Definición de la arquitectura del mercado de datos.
7. Definición de los hechos, las medidas y las dimensiones del mercado de datos.
8. Diseño del subsistema de integración.
9. Diseño del subsistema de visualización.
10. Implementación del modelo de datos.
11. Documentación del modelo de datos.
12. Implementación del subsistema de integración.
13. Documentación del subsistema de integración.
14. Implementación del subsistema de visualización.
15. Documentación del subsistema de visualización.
16. Aplicación de las listas de chequeo.
17. Aplicación de los casos de prueba.

La **estructura del documento** se encuentra definida en cuatro capítulos:

Capítulo 1: Fundamentos teóricos sobre el desarrollo de los almacenes de datos.

En el transcurso del presente capítulo se abordan múltiples elementos teóricos pertenecientes a los almacenes de datos (AD) y los mercados de datos (MD), además de los procesos fundamentales de los mismos. Se efectúa una revisión bibliográfica con relación a las metodologías de desarrollo de los AD existentes a nivel mundial, seleccionando la metodología a utilizar en este caso. Además se especifican las herramientas a utilizar para solucionar el problema en cuestión.

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas.

En el presente capítulo se realizará el análisis del MD mediante el levantamiento de los requisitos y su agrupamiento en casos de uso. El diseño del subsistema de almacenamiento del mismo a través de la definición de los hechos, las medidas y las dimensiones identificadas para construir la matriz multidimensional y posteriormente el modelo de datos. Se realiza además el diseño del subsistema de integración de datos mediante el perfilado de los datos y el diseño general de las transformaciones. También se diseña el subsistema de visualización mediante el diseño de los reportes candidatos y la definición de la arquitectura de información.

Capítulo 3: Implementación del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas.

Este capítulo comprende la implementación del subsistema de almacenamiento, mediante la construcción del modelo físico y la definición de los estándares de codificación. Además incluye la implementación del subsistema de integración de datos a través de la gestión del cambio en las dimensiones, de los metadatos, la captura de los cambios en los datos y la implementación de los hechos y los trabajos. También se realizó la implementación del subsistema de visualización, el cual comprende la implementación de los cubos OLAP y los reportes con los que se le da solución a las necesidades del negocio.

Capítulo 4: Realizar pruebas al mercado de datos Cooperación Internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas.

En este capítulo se plasman los resultados arrojados por las pruebas aplicadas a la solución, mediante las listas de chequeo a los artefactos de ETL, los casos de pruebas (CP) por CU y por reglas de transformación y el perfilado de los datos para comprobar la calidad de la información cargada al MD. Estas pruebas son ejecutadas con el objetivo de avalar el cumplimiento de las exigencias del cliente y garantizar la calidad del sistema.

Capítulo 1: Fundamentos teóricos sobre el desarrollo de los almacenes de datos

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

En el transcurso del presente capítulo se abordan múltiples elementos teóricos pertenecientes a los AD y MD, además se describen los procesos fundamentales de los mismos. También se efectúa una revisión bibliográfica con relación a las metodologías existentes a nivel mundial para el desarrollo de estos sistemas, seleccionando la metodología a utilizar en la investigación. Asimismo se especifican las herramientas a utilizar para solucionar el problema en cuestión.

1.1 Cooperación internacional

El área de Cooperación internacional de la UCI maneja un flujo de información importante mediante los cuatro departamentos por los que se compone: **Trámites**, como su nombre expresa se encarga de tramitar las entradas y salidas al país de trabajadores y estudiantes, **Visitas**, gestiona lo referente a la atención de las personalidades que acuden a la UCI por diversos motivos, **Cooperación internacional**, administra los asuntos relacionados con los acuerdos de colaboración internacional de la universidad y de la becas en el extranjero de sus profesores, **Economía**, departamento responsable de controlar los gastos e ingresos de esta área de la universidad. En cada uno de estos departamentos se manipulan una serie de indicadores históricos significativos como: total de misiones realizadas, total de acuerdos de colaboración, presupuesto anual y total de visitas.

1.2 Almacenes de datos

Los AD constituyen un escalón superior en la evolución de las BD hacia mayor inteligencia y mejores funcionalidades. Brindando de esta manera mecanismos sencillos a los directivos de la empresa u organización para reflejar tendencias del mercado, evoluciones, hechos históricos en el negocio y posibilidades futuras.

Un AD es un gran repositorio lógico de datos que permite el acceso y la manipulación flexible de grandes volúmenes de información derivados tanto de transacciones pormenorizadas como datos añadidos de fuentes de distinta naturaleza (1).

Características, ventajas y desventajas

Los fundamentos o características de los AD según W. H. Inmon, padre de la disciplina, se basan en que estos sistemas consisten en una colección de datos **orientado a materias, integrado, no volátil, que varía con el tiempo** y que se diseñan para dar apoyo a los procesos de toma de decisiones (2).

Se puede decir que un AD es **orientado a materias** porque sus funcionalidades se concentran en un tema en específico o en base a los aspectos que son de interés para la empresa. Organizando la

Capítulo 1: Fundamentos teóricos sobre el desarrollo de los almacenes de datos

visualización de los datos de manera que sea más apropiada y simple para el usuario final, buscando en todo momento rapidez en el acceso a la información. Además es **integrado** debido a que se construye mediante la unificación de fuentes de datos múltiples y heterogéneas, eliminando las inconsistencias en el nombrado, estructuras codificadas y medidas de los atributos que puedan surgir entre ellas. Una vez que se encuentra poblado el almacén, los datos no cambian, sus actualizaciones no ocurren en su propio entorno, sino en las fuentes de datos de las que se nutre, siendo esta una de las razones por las que un AD es **no volátil**. La **variación en el tiempo** se ve evidenciada en un AD primeramente porque cada clave contiene una referencia a la fecha explícita o implícitamente. Igualmente el espacio de tiempo en él, es significativamente más largo que el de los sistemas operacionales debido a que la información es mostrada desde una perspectiva histórica. Una característica agregada y distintiva de estos sistemas es que contienen metadatos, es decir, datos sobre los datos (2; 3).

Ventajas

A continuación se refieren algunas ventajas que trae la implementación de un AD (2):

- Proporciona una herramienta para el proceso de toma de decisiones estratégicas y tácticas en cualquier área funcional, teniendo en cuenta información integrada y global del negocio.
- Simplifica la aplicación de técnicas estadísticas de análisis y modelación para encontrar relaciones ocultas entre los datos del almacén; obteniendo metadatos como valor añadido para el negocio sobre dicha información.
- Aporta la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.

Desventajas

Algunas de las desventajas de los AD son las que se presentan posteriormente:

- Subestimación de los recursos imprescindibles para la captura, carga y almacenamiento de los datos (4).
- Incremento continuo de los requerimientos de información del AD (4).
- Subestimación de las funcionalidades que ofrece la correcta utilización del AD (2).

1.3 Mercado de datos

Cuando se abarca el tema de los AD es imprescindible mencionar los MD. Estos son una solución que, comparten tecnología con los AD pero comprenden contenidos específicos, sus volúmenes de datos son más limitados y tienen un alcance histórico menor (5).

Capítulo 1: Fundamentos teóricos sobre el desarrollo de los almacenes de datos

Un MD es la implementación de un almacén con alcance limitado a un área funcional, problema en específico, departamento, tema o grupo de necesidades. Este se puede integrar a un AD de dos formas, dependiendo del enfoque de la arquitectura que se decida implementar:

Enfoque de Inmon

Define previamente el AD y posteriormente delimita sobre él los MD. Esta implementación tiene como principal inconveniente que se requiere una inversión muy costosa en tiempo y recursos (2).

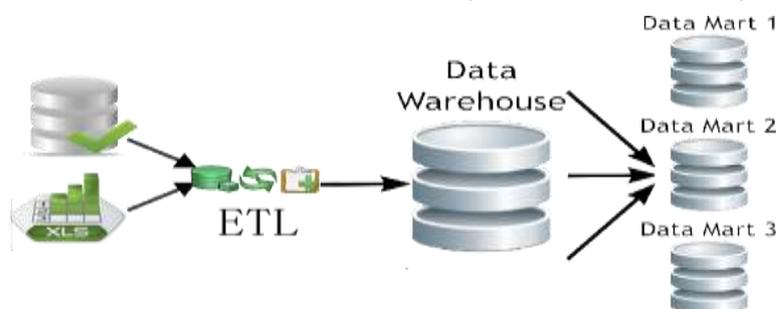


Figura 1 Arquitectura descendente

Enfoque de Kimball

Plantea crear previamente los MD departamentales y posteriormente integrarlos en un AD para la organización, la fundamental ventaja de esta implementación es que se desarrolla y despliega en un período de tiempo menor que una realizada con el enfoque de Inmon y con menos recursos. El principal problema que se presenta al aplicar esta solución está dado en tener que sincronizar los hechos cuando se realice la consolidación del AD (2).

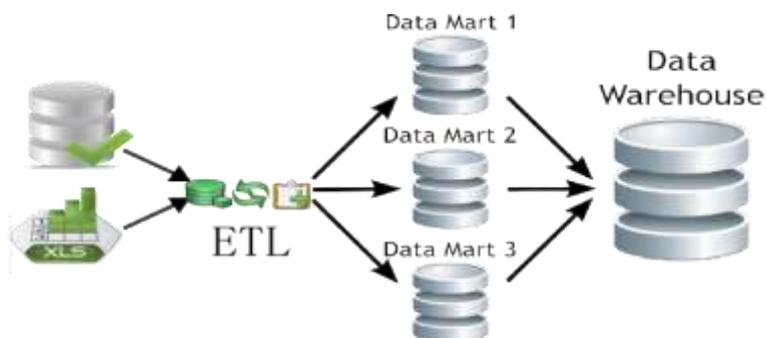


Figura 2 Arquitectura ascendente

1.4 Modelo multidimensional

En un AD se representan los datos usando modelos multidimensionales para facilitar su análisis. Un modelo multidimensional proporciona dos conceptos fundamentales: hecho y dimensión. Constituyendo el cubo el objeto central en el modelado de datos multidimensional.

Capítulo 1: Fundamentos teóricos sobre el desarrollo de los almacenes de datos

Un cubo consiste en un conjunto de celdas, de tal manera que cada una está identificada por la combinación de los miembros de las diferentes dimensiones y contiene el valor de la medida analizada para dicha combinación de dimensiones (6).

Las tablas de dimensiones especifican la organización lógica de los datos y proporcionan el medio para analizar el contexto del negocio. Representan los ejes del cubo, y los elementos del análisis, proporcionándole al usuario el filtrado y manipulación de la información almacenada en la tabla de hechos (6).

Contenida en estas tablas, en muchos casos aparece la relación lógica entre dos o más atributos, es decir las jerarquías de datos, conjuntamente, de acuerdo a las dimensiones del negocio, se refleja la granularidad, que representa el nivel de detalle en que se desea almacenar la información en el modelo (6).

Un hecho, es un valor en un entorno multidimensional determinado por dimensiones asociadas a él. Al insertar un hecho se define su llave primaria como la composición de las llaves primarias de las tablas de dimensiones relacionadas a este y poseen medidas de procesos del negocio.

1.5 Procesamiento Analítico en Línea (OLAP)

Un análisis correcto de los datos soporta a una acertada toma de decisiones, lo que conlleva a un aumento de los beneficios que se reportan al negocio. En esta área, una de las técnicas o herramientas más empleadas es OLAP. Esta solución es utilizada en el campo de la inteligencia de negocio (BI por sus siglas en inglés), para ello se manejan los datos extraídos y resumidos de BD o sistemas transaccionales en cubos OLAP, proporcionando la rapidez y flexibilidad requeridas para brindar apoyo al analista en tiempo real.

Las herramientas OLAP, son una tecnología de software creada para analizar, administrar y ejecutar consultas sobre los datos en línea, que permiten deducir información del comportamiento del negocio. Su principal objetivo es brindar respuestas rápidas a complicadas preguntas, para descifrar la situación del negocio y tomar decisiones (2).

De manera general OLAP presenta muchos beneficios y utilidades entre las que se destacan (7):

- Se tiene acceso a grandes cantidades de datos y se presentan en diversas perspectivas.
- Involucra datos agregados y los compara a través de períodos jerárquicos.
- Es de fácil uso, acceso flexible y responde con rapidez a las peticiones de los usuarios.

1.5.1 Implementaciones de OLAP

La tecnología OLAP posee múltiples implementaciones, las que se diferencian teniendo en cuenta sus características. Dichas implementaciones son:

Procesamiento Analítico Relacional en Línea (ROLAP por sus siglas en inglés): almacena los datos en un motor relacional de forma detallada en tablas normalizadas. Comúnmente se utilizan esquemas en estrella o copo de nieve para esta implementación, pero esta además permite trabajar sobre cualquier BD relacional. Siendo la principal ventaja de esta arquitectura que permite el análisis de gran cantidad de datos (8).

Procesamiento Analítico Multidimensional en Línea (MOLAP por sus siglas en inglés): almacena físicamente los datos en estructuras multidimensionales de modo que las representaciones externas e internas concuerden. Entre sus principales ventajas se encuentran que ostenta una tecnología perfeccionada para las consultas y el análisis basado en el modelo multidimensional. Aunque también presenta algunas desventajas, partiendo de que esta implementación no es adecuada cuando se necesita trabajar con gran cantidad de tablas de dimensiones y también que las herramientas diseñadas para esta son restringidas y propietarias (8).

Procesamiento Analítico Híbrido en Línea (HOLAP por sus siglas en inglés): sistema híbrido entre las arquitecturas ROLAP y MOLAP, que combina estas dos implementaciones para almacenar algunos datos, como los registros de detalle, en un motor relacional y otros, como las agregaciones, en un almacén multidimensional. Brinda una solución con las mejores características de ambas: desempeño superior y gran escalabilidad (8).

El tipo de procesamiento a utilizar es ROLAP debido a que este es el más apropiado para analizar grandes volúmenes de información y además el sistema gestor de base de datos (SGBD) a utilizar sólo soporta el procesamiento relacional de la información.

Operaciones sobre OLAP

- Rotar (*Swap*): sustituye las filas por columnas (permuta dos dimensiones de análisis) (7).
- Bajar (*Down*): desciende el nivel de visualización en las filas a una jerarquía inferior (7).
- Detallar (*Drill down*): permite visualizar los datos con mayor detalle, bajando por la jerarquía de una dimensión (2).
- Expandir (*Expand*): ídem al anterior, pero no pierde la información a nivel superior para éste y el resto de los valores (7).

1.6 Procesos básicos de un almacén de datos

Para el desarrollo exitoso de un AD se requiere seguir una serie de procesos en cada una de las etapas del mismo. Estas etapas van desde el análisis de la organización hasta el despliegue del AD y básicamente estos procesos se resumen en dos: la integración de los datos y el BI.

1.6.1 Integración de datos

Ralph Kimball define como integración de datos al proceso de unificación de los datos provenientes de múltiples fuentes. Este proceso se le aplica a los datos heterogéneos o “sucios”, procedentes de las distintas fuentes de datos (9).

La limpieza de los datos “sucios” es un proceso complicado y de muchas facetas que se fundamenta en el análisis los datos corporativos para revelar inexactitudes, anomalías y transformar los datos para certificar que sean precisos y coherentes (9).

Este proceso se puede realizar de distintas formas como por ejemplo: Replicación de datos, Integración de Aplicaciones Empresariales (EAI), Integración de Información Empresarial (EII) y el proceso de Extracción, Transformación y Carga (ETL), que será el aplicado en esta solución por lo que se menciona seguidamente.

Extracción, Transformación y Carga

El ETL constituye la base del AD y si se diseña apropiadamente, este extrae los datos desde los sistemas fuentes y hace cumplir la calidad y reglas de negocio (RN) definidas para ellos. Luego conforma los datos de manera que fuentes separadas pueden ser aprovechadas conjuntamente. Finalmente, hace su entrega en un formato listo para la presentación de forma tal, que los desarrolladores puedan construir la aplicación y los usuarios finales puedan tomar decisiones (9).

Aunque la creación del ETL constituye una actividad que pasa desapercibida en muchas ocasiones al usuario final, fácilmente consume el 70 por ciento de los recursos requeridos para la implementación y el mantenimiento de un típico AD (9). En el proceso ETL se añaden valores significativos a los datos. Este es mucho más que profundizar para obtener datos de los sistemas fuentes y del AD.

Fundamentalmente el ETL permite realizar funcionalidades como:

- Eliminar errores y corregir datos faltantes.
- Capturar el flujo de datos transaccionales para su protección.
- Concordar los datos de múltiples fuentes para ser usadas conjuntamente.
- Estructurar los datos para ser usables por las herramientas del usuario final (9).

1.6.2 Inteligencia de negocio

En la actualidad la información se ha convertido en fundamento para conseguir superioridad competitiva en los negocios. Para lograr este objetivo en una empresa los gerentes y funcionarios que tienen que tomar decisiones, necesitan acceder rápida y fácilmente a la información útil y valiosa de su organización. Para solucionar este problema una de las soluciones más aplicadas es el uso de BI (11).

El BI constituye la habilidad corporativa para tomar decisiones. Esto se logra mediante el uso de metodologías, herramientas y tecnologías que permiten reunir, depurar, transformar datos, y aplicar en ellos técnicas analíticas de extracción de información (11).

Las funcionalidades de las herramientas de BI están encaminadas a generar reportes, gráficos en diferentes perspectivas, vistas de análisis y todo lo que se necesite para que el usuario final pueda analizar y razonar los datos con mayor facilidad, con el objetivo de arribar a conclusiones correctas con mayor prontitud (12).

Como principales ventajas de la utilización de BI se tienen (12):

- Mayor rapidez en la obtención y lectura de la información, y en consecuencia para la toma de decisiones.
- Visibilidad inmediata y automática de los nuevos datos incorporados.
- Se genera un flujo constante de información válida, organizada y visualizada de tal forma que contribuye a minimizar la incertidumbre.

1.7 Metodología de desarrollo

Una metodología es un conjunto de pasos o acciones definidas para guiar un proceso determinado. En el caso del desarrollo del software, su definición no se aleja de este marco, a esto se le añade la definición de las etapas, los roles y las tareas para la generación de los artefactos finales de productos informáticos (13).

Con el transcurso de la madurez de los AD, se han convertido en referente en cuanto a sistemas para el apoyo en la toma de decisiones partiendo de información histórica, a su vez ha ido aparejada a dicho proceso la evolución de las metodologías por las que son regidos. Llegando dos de ellas a estar abundantemente documentadas y definidas, estas son la metodología de Kimball y la metodología de Inmon, nombradas así por su alusión a los creadores de cada una respectivamente.

Para las soluciones de esta rama DATEC propone una metodología que abarca todas las fases del desarrollo de un AD, la misma se nombra: Propuesta de metodología para el desarrollo de almacenes

Capítulo 1: Fundamentos teóricos sobre el desarrollo de los almacenes de datos

de datos en DATEC (14). La cual toma como fundamento la metodología de Kimball en los siguientes elementos:

1. Propone los conceptos de hechos y dimensiones, lo cual apoya la toma de decisiones y también en el proceso de desarrollo.
2. Plantea que la construcción del AD sea a medida que se construyan las BD departamentales, lo que da la impresión de organización en la empresa, entidad u organización.
3. Abundante documentación sobre la misma, posibilitando así la aclaración de las dudas que vayan surgiendo relacionadas a ese tema.
4. Amplio reconocimiento por los especialistas que se desempeñan en esta disciplina, pues constituye una metodología madura y que tiene bien definidas las etapas, actividades, artefactos y roles (1).

Además esta metodología se ajusta a las condiciones y características de la producción en DATEC y la UCI y está constituida por ocho fases las cuales se explican a continuación:

Estudio preliminar y planeación: se lleva a cabo un estudio abarcador de la entidad cliente. Donde se realiza el diagnóstico de información, de datos y de infraestructura tecnológica. Además se ejecutan las tareas de planeación del proyecto y otras series de actividades (14).

Levantamiento de requisitos: se realizan actividades para lograr definir las metas y objetivos de la organización, identificar las necesidades de información de los clientes, las RN y los requisitos (14).

Arquitectura: aquí se define la arquitectura de la solución según los requisitos no funcionales (RNF) obtenidos. Importante aclarar que esta fase puede desarrollarse paralelamente con la fase de levantamiento de requisitos (14).

Diseño e implantación: se realiza el diseño de los subsistemas partiendo de las estructuras de almacenamiento. Además se realizan las actividades de implementación de cada uno de los subsistemas. Esta fase puede llevarse a cabo de manera paralela para el caso del subsistema de integración de datos y visualización, pero no sucede así para el de almacenamiento ya que ambos dependen de este (14).

Prueba: aquí se aplican varias pruebas a la solución, comenzando por las pruebas unitarias hasta llegar a las pruebas de aceptación con el cliente final. Es válido aclarar que en esta fase no es la única en la que se realizan pruebas durante el desarrollo del proyecto, pues en todas las fases hay actividades de aseguramiento de la calidad (14).

Despliegue: comprende dos etapas, la primera es un despliegue piloto, donde se preparan los servidores necesarios según la arquitectura definida, se cargan una muestra de los datos en un ambiente controlado, con el fin de demostrar al cliente final que la solución funciona. Es aquí el momento ideal para realizar la capacitación y transferencia tecnológica a los clientes. Como resultado fundamental de esta fase se presenta la solución desplegada en el entorno real y en correcto funcionamiento (14).

Soporte y mantenimiento: esta fase se inicia cuando la solución está implantada y en explotación, y se ejecuta según el contrato firmado y las condiciones de soporte establecidas (14).

Gestión y administración del proyecto: Esta fase se aplica a lo largo de todo la duración del ciclo de vida del proyecto. Aquí es donde se controla, gestiona y chequea todo el desarrollo, es la columna vertebral del proyecto (14).

Una vez definida la metodología se procedió a la selección de las herramientas a utilizar las cuáles son abordadas en el epígrafe siguiente.

1.8 Herramientas seleccionadas para la solución

Seguidamente se abordan elementos sobre las herramientas seleccionadas teniendo en cuenta la arquitectura definida por DATEC.

1.8.1 Herramienta de modelado

Con la evolución de la industria del software se ha generado la necesidad de desarrollar herramientas para el modelado de sistemas, que faciliten de esta forma el trabajo de analistas y diseñadores de software. Como herramienta para el modelado de la solución se empleará *Visual Paradigm for UML 8.0 Enterprise Edition* por las cualidades que este presentan a continuación.

Visual Paradigm for UML 8.0 Enterprise Edition

Herramienta profesional que brinda soporte al modelado visual mediante la notación UML¹ para la totalidad de las fases que comprende el ciclo vital del desarrollo de software. Constituye una aplicación muy completa y con abundantes facilidades, su utilización contribuye a la creación rápida de aplicaciones de calidad y con un costo muy económico. Tiene disponibilidad en múltiples plataformas. Además Brinda la posibilidad de generar código a partir de los diagramas, así como obtener los diagramas a partir del código. Posibilita la exportación a diferentes formatos y suministra abundantes tutoriales (15).

¹ UML: Lenguaje gráfico para visualizar, especificar, construir y documentar un sistema.

1.8.2 Sistema gestor de bases de datos

Un SGBD es una colección de programas cuyo principal objetivo es servir de interfaz entre la BD, el usuario y las aplicaciones. Estos permiten definir los datos en distintos niveles de abstracción y manipularlos, garantizando la seguridad e integridad de los mismos (16). Para la elaboración de la solución se utilizará PostgreSQL 9.1 por las razones que se exponen seguidamente.

PostgreSQL 9.1

Constituye el SGBD objeto-relacional de código abierto más desarrollado del mundo y opera bajo licencia BSD². Alcanza un conjunto de características tales como: herencia, tipos de datos, funciones, restricciones, disparadores, reglas e integridad transaccional (17).

Algunas de las ventajas que presenta son (17):

- Multiplataforma.
- Documentación muy bien organizada, pública y libre.
- Soporte nativo para los lenguajes más populares del medio.
- Conectores de datos foráneos: permite añadir y consultar fuentes de datos externas desde PostgreSQL (18).

PgAdminIII

Constituye la más completa y popular herramienta de código abierto para la administración de BD PostgreSQL. Hace más sencilla la administración y está diseñado para satisfacer las necesidades de todos los usuarios, teniendo en cuenta la realización de consultas SQL desde la más sencilla, hasta el desarrollo de BD de alta complejidad (19).

Incluye entre sus características (19):

- Interfaz administrativa gráfica.
- Herramienta de consulta SQL.
- Editor de código procedural.

1.8.3 Herramientas para la integración de datos

Las herramientas de integración de datos brindan soporte a las actividades de extracción de los datos desde las diversas fuentes, transformación de ellos mediante un conjunto de procesos que devienen en la obtención de datos optimizados para el análisis y la presentación de reportes, y carga de dichos datos en el almacén (20). Para la realización de la solución se emplearán el *Pentaho Data Integration (PDI) 4.2.1* y el *DataCleaner 1.5.3* por los elementos que se mencionan a continuación.

²Berkeley Software Distribution (BSD), sólo requiere que el código fuente licenciado mantenga la información de derechos de autor y licenciamiento.

Pentaho Data Integration 4.2.1

Herramienta que incluye componentes que facilitan el modelado y la ejecución de transformaciones sobre flujos de datos. Constituye una de las soluciones ETL de código abierto más antiguas, desarrolladas y mejor valoradas del mercado. Entre las principales ventajas del PDI sobresale que es multiplataforma, además soporta un entorno de ejecución de Java 1.4 o superior, utiliza *JavaScript* en la realización de transformaciones avanzadas consistentes en cálculos y fórmula, mediante una interfaz empotrada (21).

Su diseño es orientado a metadatos, ofrece la posibilidad de copiar y leer del mismo fichero en paralelo, permitiendo maximizar la capacidad de entrada/salida en el entorno ETL. La herramienta *Spoon* también conocida como *Kettle* se encarga del diseño gráfico de trabajos y transformaciones utilizando componentes que se entrelazan entre sí mediante saltos, estos últimos establecen el flujo de datos entre los diferentes componentes. Otras herramientas son *Kitchen* y *Pan*, responsables de ejecutar los trabajos y transformaciones, diseñados anteriormente con *Spoon* ambos archivos XML³ y se guardan en ficheros con extensiones *.kjb* y *.ktr* en ese orden (21).

DataCleaner 1.5.3

Herramienta que permite la valoración del nivel de calidad de los datos comprendidos en el sistema de información preparando estos para futuras transformaciones en el proceso ETL. La misma requiere un entorno de ejecución de Java 5.0 o una versión superior y de drivers JDBC⁴. Es de fácil utilización, genera sofisticados gráficos e informes que permiten a los usuarios identificar y analizar la estructura del origen de datos y combinar resultados y gráficos, estableciendo vistas fáciles de interpretar para evaluar la calidad de los datos (22).

1.8.4 Herramientas para inteligencia de negocio

Los sistemas BI son un tipo de software en crecimiento en el mercado mundial, ellos brindan una amplia gama de capacidades funcionales para realizar todos los procesos de esta etapa. En general estas herramientas procesan gran cantidad de datos con el objetivo de generar información relevante y disminuir la incertidumbre en la toma de decisiones de negocio (23). Las herramientas seleccionadas para de la solución son *Mondrian OLAP server 3.0.4*, *Pentaho Schema Workbench 3.2.1* y *Pentaho Business Intelligence 3.10* por las razones que se explican seguidamente.

³XML: Lenguaje de Marcas Extensible, acrónimo de sus siglas en inglés.

⁴ JDBC: *Java Database Connectivity* es una aplicación para trabajar con BD desde Java.

Mondrian OLAP server 3.0.4

Motor OLAP desarrollado en Java que gestiona la comunicación entre una aplicación OLAP y las fuentes de datos. La facilidad para el análisis de grandes volúmenes de información que se encuentran en la BD es una de sus características más significativas.

Esta herramienta permite la realización de consultas al AD y brinda la posibilidad de mostrar los resultados a través de un navegador de forma tal que el usuario pueda realizar actividades típicas de navegación sobre cubos OLAP. Se ha convertido en una de las aplicaciones más importantes de la Suite de Pentaho BI.

Pentaho BI Server 3.10

La plataforma Pentaho BI proporciona la arquitectura y la infraestructura necesaria para crear soluciones de BI. Dentro de esta plataforma la aplicación más conocida es el BI server, diseñado para integrarse fácilmente en cualquier proceso de negocio, cuyo funcionamiento web se basa en un sistema de gestión de reportes, servidores de integración de aplicaciones y en el motor de flujo de trabajo ligero (24). Este servidor cuenta con una interfaz de usuario donde se encuentran disponibles todos los reportes, vistas OLAP y cuadros de mando; además tiene una consola de administración que permite gestionar y supervisar la aplicación, los usuarios de la misma, los reportes que revisa cada usuario, cuándo se accedió a cada reporte, entre otras funciones (25).

Pentaho Schema Workbench 3.2.1

Es una herramienta para el desarrollo y prueba de cubos OLAP de forma visual. La definición del XML no es extremadamente compleja, pero en la práctica resulta engorroso recordar cada uno de los elementos junto a sus atributos y sub-elementos tal y como se encuentran en el almacén. Con esta aplicación, se puede configurar una conexión con el modelo físico, para luego elaborar el esquema lógico de manera simple y efectiva. Para ello la herramienta ofrece un editor de esquemas con la fuente de datos subyacente para su validación (25).

1.8.5 Herramienta para soporte de aplicaciones web

Las herramientas para soporte web se emplean para brindar contenido estático y cargar archivos en un navegador. Este intercambio se realiza mediante protocolo HTTP (*Hypertext Transfer Protocol*).

Apache Tomcat 6.0

Apache Tomcat fue implementado en Java, por lo que trabaja en cualquier sistema operativo que tenga instalado *Java Development Kit* (JDK). Es un servidor web de software libre bajo licencia Apache en su versión 2.0, gestionado por la fundación Apache Tomcat, es el más utilizado a la hora de trabajar con Java en entornos web. Puede funcionar como servidor HTTP facilitando de esta forma la gestión

de solicitudes y respuestas, o conectado a otro servidor HTTP como Apache HTTP Server (26). Esta herramienta fue escogida debido a que la plataforma de Pentaho BI utilizada la instala por defecto.

1.8.6 Herramienta para el control de versiones

Para el desarrollo de un sistema es de mucha trascendencia emplear un correcto control de versiones. Estos son los métodos y herramientas disponibles para controlar los aspectos relacionados con los cambios en el tiempo de un archivo. La herramienta que se seleccionó para esta actividad en la realización de la solución es *Subversion 1.5.4* por el conjunto de razones que se muestran a continuación.

Subversion (SVN 1.5.4)

Sistema de control de versiones libre y de código fuente abierto. Posee como elemento fundamental un repositorio, el cual constituye un almacén central de datos. Este guarda información en forma de árbol de archivos. Soporta cualquier cantidad de conexiones simultáneas al repositorio y pudiendo cada uno de los usuarios leer o escribir en esos archivos. Al escribir datos, un cliente pone a disposición de otros la información; así mismo al leer datos, el cliente recibe información de otros (27).

1.9 Conclusiones del capítulo

Luego de finalizado este capítulo, se concluye que fueron abordados los principales conceptos relacionados con los MD, lo que permitió deducir la necesidad de crear un MD para el área Cooperación internacional. La metodología seleccionada: Propuesta de metodología para el desarrollo de almacenes de datos en DATEC, comprende todas las fases de la construcción de un MD y ofrece varias ventajas que facilitan su desarrollo. El *Visual Paradigm for UML 8.0* permitirá elaborar los diagramas fundamentales de la solución. El uso de PostgreSQL 9.1 como SGBD y el PgAdmin III 1.14.0 como herramienta de administración de los datos, permiten el soporte de las estructuras físicas para brindar un correcto almacenamiento de la información. Las herramientas seleccionadas para los procesos de ETL, DataCleaner 1.5.3 y *Pentaho Data Integration 4.2.1*, posibilitarán que los datos cuenten con la calidad necesaria para ser cargados al MD. Las herramientas de BI elegidas (*Schema Workbench 3.2.1*, *Pentaho BI server 3.10*, *Mondrian OLAP server 3.0.4* y *Apache Tomcat 6.0*) permitirán la implementación de la capa de visualización del MD que mostrarán los reportes para el soporte a la toma de decisiones.

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS COOPERACIÓN INTERNACIONAL PARA LA SALA SITUACIONAL DE LA UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS.

En el presente capítulo se realizará el análisis del MD mediante el levantamiento de los requisitos y su agrupamiento en casos de uso. El diseño del subsistema de almacenamiento del mismo a través de la definición de los hechos, las medidas y las dimensiones identificadas para construir la matriz multidimensional y posteriormente el modelo de datos. Se realiza además el diseño del subsistema de integración de datos mediante el perfilado de los datos y el diseño general de las transformaciones. También se diseña el subsistema de visualización mediante el diseño de los reportes candidatos y la definición de la arquitectura de información.

2.1 Análisis del MD Cooperación internacional

El análisis comprende una serie de tareas, como por ejemplo: estudio preliminar del negocio, investigación, negociación, identificación de las necesidades de información, clasificación y posterior especificación de los requisitos, que acarrearán con la elaboración del diagrama de casos de uso del sistema (DCUS) y la especificación de cada caso de uso (CU) (28).

2.2 Definición del negocio

El proyecto Sala Situacional UCI constituye un entorno de análisis, monitoreo y control de la información de las principales direcciones de esta universidad, con el objetivo de ofrecer a sus directivos información verídica y actualizada de cada una de ellas. Cooperación internacional es una de estas direcciones, en la cual se maneja una serie de datos relacionados con la entrada y salida al país de sus trabajadores y estudiantes, los acuerdos de colaboración con otros países o universidades, las personalidades que la visitan y los ingresos y gastos de esta área. Los datos de las áreas visitas, economía y colaboración internacional se encuentran almacenados en diversos documentos de formato excel y el departamento de trámites cuenta para la manipulación de su información con un sistema de gestión. Los datos de esta área son acumulados de forma mensual en el departamento de trámites y anual en los restantes departamentos. Partiendo de esto se definen para la presente investigación como áreas de análisis: trámites, visitas, colaboración internacional y economía.

2.3 Reglas del negocio

Las RN especifican en un nivel adecuado de detalle lo que debe hacer o no una organización, para lo cual se es necesario expresarlas en lenguaje natural y deben estar individualmente centradas en ese negocio (28).

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

A continuación se exponen las RN que serán aplicadas para el MD Cooperación internacional con el objetivo de homogenizar y estructurar la información para de esa forma facilitar su análisis:

- RN1.** Una vez que los datos estén cargados en el almacén, no pueden existir campos nulos.
- RN2.** El código de los atributos en cada una de las dimensiones no puede tomar valores repetidos.
- RN3.** No pueden existir campos con valores negativos.
- RN4.** Los acuerdos de colaboración que tengan el campo ENTIDAD vacío se les debe renombrar por Desconocido.
- RN5.** En el campo CIUDADANÍA de los extranjeros se renombraran los valores ARGENTINA, COLOMBIANA, ECUATORIANA y BOLIVIANA respectivamente por Argentina, Colombia, Ecuador y Bolivia.
- RN6.** Las visitas, extranjeros y acuerdos de colaboración que tengan el campo País vacío se reemplazará por Desconocido.
- RN7.** El campo nombre_pais_iso en la tabla sq_cooperacion.tb_npais se renombrará a País.
- RN8.** En los acuerdos de colaboración que contengan el campo Firma UCI vacío se reemplazará por Desconocido.
- RN9.** Las visitas que contengan el campo Organismo solicitante vacío se reemplazará por Desconocido.
- RN10.** Las visitas que contengan el campo Institución/Cargo vacío se reemplazará por Desconocido.
- RN11.** Los campos periodo_CUC, mes_cuc, periodo_MN y mes_mn del presupuesto se renombrarán a *periodo* y *mes* respectivamente.
- RN12.** Los tipos de material que contengan el campo descripción vacío se reemplazará por Desconocido.
- RN13.** Los tipos de moneda que contengan el campo descripción vacío se reemplazará por Desconocido.

2.4 Necesidades de los usuarios

En el proceso de análisis se lleva a cabo la traducción de las necesidades de los usuarios, es decir lo que precisan que el sistema les facilite, en requisitos. De la calidad con la cual se realiza la extracción de estas necesidades depende en gran medida el éxito o fracaso del software. Para el caso de los AD o MD, a los requisitos funcionales (RF) y a los RNF se les adicionan los requisitos de información (RI).

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

2.3.1 Requisitos de información

Se definen como RI al conjunto de funcionalidades que se deben implementar para analizar los datos en cuanto a parámetros previamente definidos. Dichos requisitos constituyen la principal entrada para el proceso de BI y la posterior visualización de los reportes. Como resultado del proceso de análisis efectuado para esta investigación se identificaron 16 RI pertenecientes a los cuatro temas de análisis mencionados anteriormente.

A continuación se declaran los RI identificados para esta investigación:

- RI1** Obtener el total de misiones realizadas por país, tipo de misión y/o período.
- RI2** Obtener el total de prórrogas por estado de prórroga, nombre de misión, tipo de misión y/o período.
- RI3** Obtener el total de pasaportes UCI por tipo de pasaporte, período y/o estado del pasaporte.
- RI4** Obtener el total de trámites realizados por nombre de misión, tipo de misión, país y/o período.
- RI5** Obtener el total de becas intergubernamentales por país y/o período.
- RI6** Obtener el total de misioneros en el exterior por país, nombre de misión, tipo de misión y/o período.
- RI7** Obtener el total de misioneros con misión cumplida por país, tipo de misión, nombre de misión y/o período.
- RI8** Obtener el total de extranjeros con residencia temporal en Cuba por país y/o período.
- RI9** Obtener el total de acuerdos de colaboración por país, entidad, directivo de la UCI firmante y/o período de firma.
- RI10** Obtener el total de profesores invitados por país, actividad y/o período.
- RI11** Obtener el total de visitas por país, clasificación y/o período.
- RI12** Obtener el total de personal que ha visitado la UCI por país y/o período.
- RI13** Obtener el total de regalos entregados por tipo de artículo y/o período.
- RI14** Obtener el presupuesto del área por epígrafe, tipo de moneda y/o período.
- RI15** Obtener el total de gastos autorizados por el CAD por período.
- RI16** Obtener cantidad de materiales gastados por tipo de epígrafe y/o período.

2.3.2 Requisitos funcionales

De acuerdo con la Ayuda del Proceso Unificado de Software (RUP) un RF es una capacidad de software necesaria para que el usuario solucione un problema al alcanzar un objetivo (29). A continuación se exponen los RF identificados para esta investigación:

- RF1.** Validar datos y contraseña

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

- RF2.** Adicionar rol
- RF3.** Eliminar rol
- RF4.** Visualizar rol
- RF5.** Modificar rol
- RF6.** Adicionar usuario
- RF7.** Eliminar usuario
- RF8.** Visualizar usuario
- RF9.** Modificar usuario
- RF10.** Adicionar reporte
- RF11.** Visualizar reporte
- RF12.** Eliminar reporte
- RF13.** Modificar reporte
- RF14.** Extraer datos de la fuente
- RF15.** Realizar transformación de los datos
- RF16.** Cargar los datos al MD Cooperación internacional
- RF17.** Personalizar reporte
- RF18.** Exportar a otro formato

2.3.3 Requisitos no funcionales

Los RNF describen atributos del sistema o atributos del entorno del sistema (28). Es decir no tienen que ver directamente con las funcionalidades de sistema. En la presente investigación fueron identificados 14 RNF que comprenden temas como la usabilidad, confiabilidad, soporte y restricciones de diseño entre otros, los cuales se muestran a continuación:

RNF 1 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 2 Establecer tiempo de entrenamiento requerido para que usuarios normales sean productivos operando el sistema: el tiempo de entrenamiento deberá ser entre 7 y 14 días. Para usuarios con un nivel avanzado será 7 días. Para lograr el cumplimiento de los tiempos establecidos por parte de los usuarios es necesario un dominio del funcionamiento del negocio en correspondencia con el rol que ocupen.

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

RNF 3 Garantizar la persistencia de la información: se realizará un respaldo total de los datos del MD con una frecuencia anual. Toda esta información se almacenará en el área de la dirección de informática del departamento de investigaciones de la UCI. Este banco estará ubicado en el edificio correspondiente al departamento de investigaciones y será responsabilidad de su grupo de administración de redes.

RNF 4 Garantizar el cumplimiento de actualización de los datos en el almacén: la información contenida en el AD tendrá una precisión y exactitud anual.

RNF 5 Lograr la homogeneidad de la estructura de los elementos definidos en el almacén. Las estructuras del AD serán nombradas de una manera estándar teniendo en cuenta el tipo de estructura que sea. En la siguiente tabla se definen convenciones de nombrado con el objetivo de manejar un vocabulario común en todo el AD, permitiendo un entendimiento claro y conciso por parte de los desarrolladores.

RNF 6 Utilizar los lenguajes de programación definidos durante la investigación: para la realización de MD se utilizará como SGBD el PostgreSQL, la herramienta para la administración de este es el pgAdmin, el lenguaje utilizado por este es el PL/pgSQL En la implementación de los procesos de integración de datos se el lenguaje a utilizar es JavaScript y para la realización de las consultas se utilizará el lenguaje MDX.

RNF 7 Confección de un manual de usuario.

RNF 8 Proporcionar características mínimas de hardware a los servidores: los servidores deben contar con los siguientes requerimientos de hardware para lograr una explotación aceptable del sistema:

- 1 GB RAM o superior.
- 1 Microprocesador Core2Duo.
- Pentium 4.
- Almacenamiento en disco 1GB.

RNF 9 Acceder al sistema: el usuario deberá acceder a la aplicación usando el navegador Firefox 2.0 o una versión superior a través del protocolo HTTP.

RNF 10 Para un correcto funcionamiento del sistema la aplicación debe contar con :

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

- Java Virtual Machine versión 6.6 o superior y *Mondrian Schema Workbench* versión 3.2.1 o superior, en caso de que un usuario capacitado requiera la construcción de esquemas multidimensionales para el diseño de nuevos reportes.
- *Pentaho Data Integration* versión 4.2.0 o superior, en caso de que un usuario capacitado requiera realizar la carga de los datos nuevamente.

2.5 Casos de uso del sistema

Los CU constituyen un conjunto de funcionalidades agrupadas por un patrón determinado que benefician a uno o varios actores (29). La creación de estos para el desarrollo del MD Cooperación internacional fue en correspondencia con lo planteado en la metodología seleccionada.

2.3.4 Actores del sistema

Los actores del sistema son personas que interactúan con la aplicación y se benefician de ella (29). Los actores identificados para el MD Cooperación internacional son los siguientes:

Tabla 1 Actores del sistema

Actor	Responsabilidad en el sistema
Administrador	Realiza todas las actividades de administración del MD. Además se puede comportar como el actor analista debido a la relación de generalización/especialización que ellos tienen, evidenciándose de esta manera el patrón múltiples actores rol común.
Administrador ETL	Realiza las actividades de extracción de los datos de las fuentes, así como la transformación y la carga de estos hacia el MD.
Analista	Consulta la información de los reportes.

2.3.5 Diagrama de Casos de uso del sistema

La interacción entre los CU y los actores del sistema se modela en un DCUS. Para la elaboración de este se tomó como base los RI y RF anteriormente identificados agrupados en 11 CU, de ellos cuatro son CU de información agrupados por las áreas de análisis definidas anteriormente:

- CU Presentar información de trámites.
- CU Presentar información de visitas
- CU Presentar información de colaboración internacional.
- CU Presentar información de economía.

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

Los siete restantes son CU funcionales:

- CU Gestionar reporte
- CU Gestionar rol
- CU Gestionar usuario
- CU Autenticar usuario
- CU Realizar extracción de los datos
- CU Realizar transformación y carga de los datos
- CU Realizar operaciones sobre reportes.

De ellos tres son agrupados por el patrón CRUD, que plantea la agrupación de las funcionalidades crear, leer, modificar y eliminar. Este se ve reflejado en los CU Gestionar reporte, Gestionar usuario y Gestionar rol. Otro patrón de CU reflejado en este diagrama es concordancia por adición, el cual se pone de manifiesto en la relación de extensión que existe entre los CU de información y el CU Realizar operaciones sobre reportes.

Después de todo lo previamente expuesto se presenta el DCUS del MD Cooperación internacional:

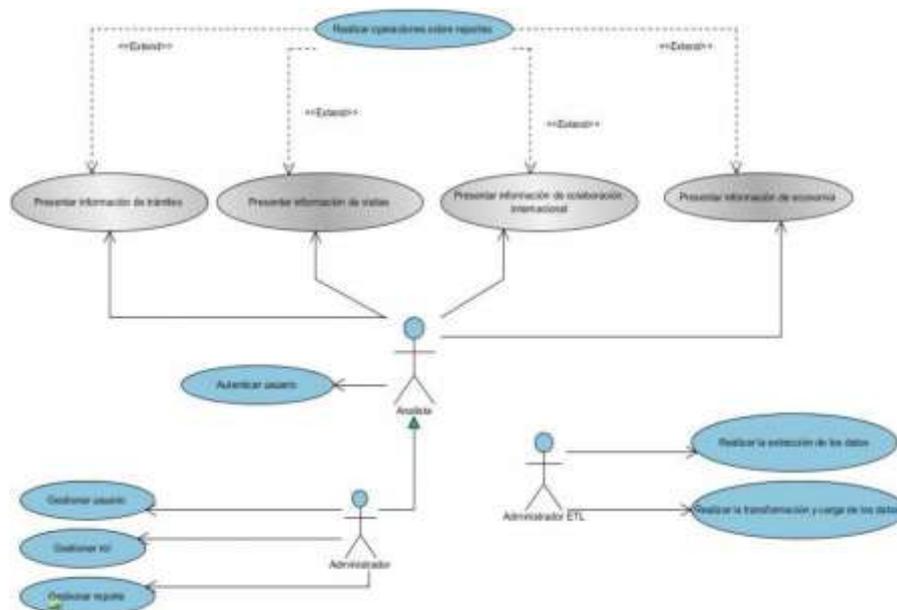


Figura 3 DCUS

2.3.6 Especificación de casos de uso del sistema

Seguidamente se muestra como ejemplo la descripción del CU “Presentar información de colaboración internacional” (ver tabla 2), el resto de las descripciones se encuentran en los Anexos.

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

Tabla 2 Descripción CU “Presentar información de colaboración internacional”

Caso de Uso:	Presentar información de colaboración internacional	
Objetivo	Presentar la información referente a los acuerdos de colaboración	
Actores	Analista y Administrador	
Resumen	El caso de uso inicia cuando el actor solicita realizar un análisis de la información referente a los acuerdos de colaboración desde diferentes perspectivas. El actor selecciona el reporte que desea visualizar, el sistema muestra la información contenida en él y las opciones de los posibles cambios que le puede efectuar al reporte. El caso de uso concluye cuando el actor finaliza el análisis de la información referente a los acuerdos de colaboración.	
Complejidad	Media	
Prioridad	Media	
Precondiciones	El usuario debe estar autenticado en el sistema. Mercado de datos poblado y funcional.	
Postcondiciones	Los reportes pertenecientes al caso de uso fueron consultados.	
Flujo de eventos		
Flujo básico Presentar información de colaboración internacional		
	Actor	Sistema
1.	Selecciona el área de análisis general A.A.G_Sala_Situacional_UCI	
2.		Muestra las áreas de análisis contenidas en el A.A.G_Sala_Situacional_UCI
3.	Selecciona el área de análisis A.A_Cooperación_internacional	
4.		Muestra los libros de trabajo contenidos en el A.A_Cooperación_internacional
5.	Selecciona el libro de trabajo LT_Colaboración_internacional	
6.		Muestra los reportes pertenecientes a al LT_Colaboración_internacional
7.	Selecciona el reporte que desea analizar	
8.		Presenta la información contenida en el

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

		reporte escogido y ofrece opciones al actor para realizar cambios al reporte durante su análisis. Ir al CU Realizar operaciones sobre reportes. Se finaliza el caso de uso															
Opciones de reportes de Presentar información de colaboración internacional																	
Perspectivas de análisis	Posibles resultados																
	Medidas	Periodicidad															
Variables de entradas disponibles relacionadas con el CU Presentar información de colaboración internacional: ✓Entidad del acuerdo ✓Tiempo ✓Funcionario UCI ✓País	Variables de salidas relacionadas con el hecho hech_colaboracion_internacional: ✓Total de acuerdos de colaboración	Rango de tiempo en que se solicitan las variables de salida: ✓Anual															
Prototipo de interfaz de usuario	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th colspan="4"></th> <th style="text-align: right;">Medidas</th> </tr> </thead> <tbody> <tr> <td style="text-align: center;">Entidad del acuerdo</td> <td style="text-align: center;">País</td> <td style="text-align: center;">Tiempo</td> <td style="text-align: center;">Funcionario UCI</td> <td style="text-align: right;">• Total de acuerdos de colaboración</td> </tr> <tr> <td style="text-align: center;">+ Todas</td> <td style="text-align: center;">+ Todos</td> <td style="text-align: center;">+ Todos</td> <td style="text-align: center;">+ Todos</td> <td></td> </tr> </tbody> </table>						Medidas	Entidad del acuerdo	País	Tiempo	Funcionario UCI	• Total de acuerdos de colaboración	+ Todas	+ Todos	+ Todos	+ Todos	
				Medidas													
Entidad del acuerdo	País	Tiempo	Funcionario UCI	• Total de acuerdos de colaboración													
+ Todas	+ Todos	+ Todos	+ Todos														
Relaciones	CU Incluidos	No aplica															
	CU Extendidos	CU Realizar operaciones sobre reportes															
Requisitos funcionales no	RNF 1, RNF 3																
Asuntos pendientes																	

2.4 Arquitectura del mercado de datos

La arquitectura de la solución se define a partir de los RNF obtenidos en el levantamiento de requisitos. Esta comprende aspectos como, la comunicación entre los subsistemas y la tecnología a utilizar, entre otros aspectos de gran importancia (14). Para el caso del MD Cooperación internacional la arquitectura definida consta de tres subsistemas: integración de datos, almacenamiento y visualización.

El subsistema de integración de datos extrae los datos desde las distintas fuentes, en este caso excel y BD, comunicándose con la última mediante el protocolo TCP/IP. Luego se encarga de realizar la

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

limpieza, transformación e integración de los mismos mediante las herramientas Kettle, del PDI 4.2.1 y DataCleaner 1.5.3, dejándolos listos para la carga en el MD.

El subsistema de almacenamiento toma como elementos de partida los datos manipulados por el subsistema de integración utilizando el protocolo de comunicación TCP/IP. En él se encuentra el modelo físico propuesto soportado en el SGBD PostgreSQL en su versión 9.1 y mediante la utilización de la herramienta PgAdminIII 1.14.0 se garantizará la interacción de los clientes de administración con este.

Los datos que están en el subsistema de almacenamiento son consultados por el subsistema de visualización mediante protocolo TCP/IP con la herramienta Pentaho BI server 3.10. A este subsistema es que acceden los distintos clientes para obtener los reportes deseados. Para un mejor entendimiento de lo antes expuesto ver figura 4.

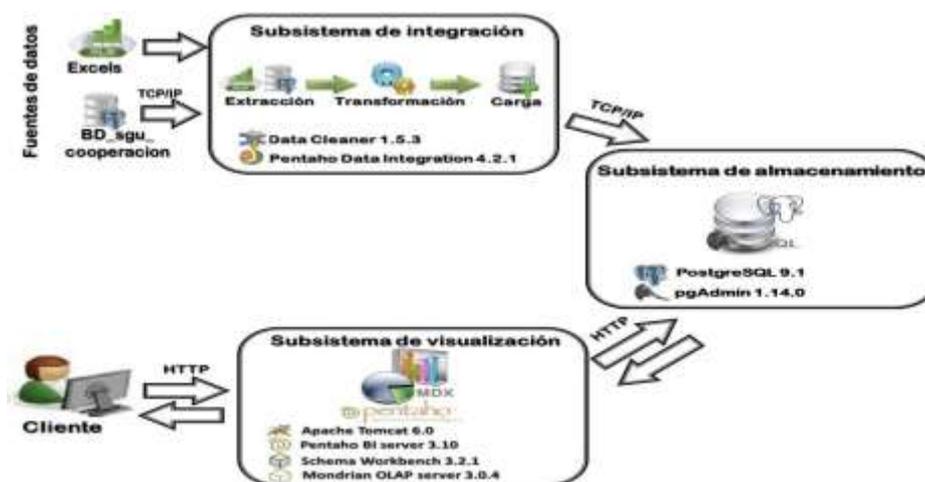


Figura 4 Arquitectura del MD

2.4.1 Diseño del subsistema de almacenamiento

El proceso de diseño del subsistema de almacenamiento de un MD se basa en la identificación de los datos de interés para el negocio y su transformación al modelo lógico como medidas, dimensiones, hechos. Dicho modelo se traducirá posteriormente al modelo físico, en el cual se almacenará el MD.

Dimensiones, hechos y medidas

Los términos hecho, dimensión y medida son muy comunes en la disciplina de los AD, debido a que estos constituyen uno de los principales aspectos que la caracterizan y diferencian sobre las BD. Un hecho representa la ocurrencia de un proceso específico en el interior de la organización; por su parte una dimensión constituye una perspectiva mediante la cual se puede llevar a cabo el análisis sobre el hecho y una medida constituye un valor o indicador de análisis del hecho (6).

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

Partiendo de lo identificado anteriormente en el análisis y en conjunto con el estudio de las fuentes de datos se elaboró el modelo de datos en el cual se identificaron 20 tablas de dimensiones, las cuales se presentan a continuación:

Tabla 3 Descripción de las dimensiones

Dimensión	Jerarquía	Descripción
dim_pais	<i>país</i>	Indica el nombre de cada país y su código, con los cuales tienen relación con el trabajo de los diferentes departamentos del área de Cooperación internacional.
dim_tiempo_dia	<i>tiempo</i>	Registra un conjunto de atributos relacionados con la fecha en la cual se realiza el trabajo de los diferentes departamentos del área de Cooperación internacional.
dim_tipo_moneda	<i>tipo_moneda</i>	Contiene el tipo de moneda con que se realizan los gastos y se define el presupuesto del área de Cooperación internacional. Además de contar con su respectivo código.
dim_tipo_importe	<i>tipo_importe</i>	Refleja si es un gasto o un plan el importe realizado por los especialistas del área de Cooperación internacional. Además de contar con su respectivo código.
dim_visitas	<i>visitas</i>	Se compone de los datos referentes a las personalidades que han visitado la UCI.
dim_regalos	<i>regalos</i>	Registra los regalos que se le han entregado a las personalidades que han visitado la UCI.
dim_acuerdos_colaboracion	<i>acuerdos_colaboración</i>	Indica las entidades con las cuales la UCI ha establecido acuerdos de colaboración. Además del código perteneciente al acuerdo.
dim_firma_uci	<i>firma_uci</i>	Indica el nombre del funcionario que por parte de la UCI firmó algún acuerdo de colaboración. Además del código correspondiente a dicho funcionario.
dim_partida	<i>partida</i>	Contiene la información relacionada con los

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

		gastos por concepto de partidas y subpartidas que se llevan a cabo en el área de Cooperación internacional.
dim_cartas_cad	<i>cartas_cad</i>	Contiene la información asociada a los gastos por concepto de cartas a la Comisión de Aprobación de Divisas (CAD) del MIC del área de Cooperación internacional.
dim_material	<i>material</i>	Comprende los tipos de materiales y los artículos gastados por el área de Cooperación internacional.
dim_clasificacion	<i>clasificación</i>	Comprende las diferentes clasificaciones de los visitantes que acuden a la UCI.
dim_extranjero	<i>extranjero</i>	Contiene los datos de los extranjeros que se les ha gestionado el carnet de residencia temporal en Cuba por parte de la UCI.
dim_mision	<i>misión</i>	Contiene los datos referentes a las misiones con las que cuenta la UCI.
dim_actividad	<i>actividad</i>	Contiene las actividades que realizan los visitantes a la UCI.
dim_misionero	<i>misionero</i>	Comprende los datos de los profesores, trabajadores o estudiantes de la UCI que son misioneros.
dim_tramite	<i>tramite</i>	Contiene los tipos de trámites que se les realizan a los misioneros de la UCI.
dim_estado_pasaporte	<i>estado_pasaporte</i>	Contiene los datos del estado de los pasaportes que son gestionados por la UCI para sus misioneros.
dim_tipo_pasaporte	<i>tipo_pasaporte</i>	Contiene los tipos de pasaportes que son gestionados por la UCI para sus misioneros.
dim_estado_prorroga	<i>estado_prorroga</i>	Registran los datos sobre el estado de las prórrogas que se gestionan a los misioneros por parte del área de Cooperación internacional de la UCI.

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

Después de descritas las dimensiones, se presenta en la siguiente tabla la descripción de los 11 hechos definidos para el MD Cooperación internacional:

Tabla 4 Descripción de los hechos

Hecho	Descripción
hech_visitas	Es objeto de análisis del negocio. Comprende todo lo relacionado con las personalidades que han visitado la UCI.
hech_colaboracion_internacional	Es objeto de análisis del negocio. Comprende todo lo relacionado con los acuerdos de colaboración que se han firmado en la UCI.
hech_presupuesto	Es objeto de análisis del negocio. Comprende todo lo relacionado con el presupuesto asignado al área de Cooperación internacional.
hech_cartas_cad	Es objeto de análisis del negocio. Comprende todo lo relacionado con los gastos por concepto de cartas al CAD del área de Cooperación internacional.
hech_materiales	Es objeto de análisis del negocio. Comprende todo lo relacionado con los materiales gastados por el área de Cooperación internacional.
hech_carnet_extranjeros	Es objeto de análisis del negocio. Comprende todo lo relacionado con los carnets de residencia temporal en Cuba gestionados por parte de la UCI.
hech_mision	Es objeto de análisis del negocio. Comprende todo lo relacionado con las misiones con las que cuenta la UCI.
hech_pasaportes	Es objeto de análisis del negocio. Comprende todo lo relacionado con los pasaportes entregados por los especialistas de Cooperación internacional de la UCI.
hech_prorrogas	Es objeto de análisis del negocio. Comprende todo lo relacionado con las prórrogas a los misioneros, gestionadas por los especialistas de Cooperación internacional de la UCI.
hech_regalos_entregados	Es objeto de análisis del negocio. Comprende todo lo relacionado con los regalos que se les han entregado a las personalidades que han visitado la UCI.
hech_tramites	Es objeto de análisis del negocio. Comprende todo lo

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

	relacionado con los trámites a los misioneros que son realizados por los especialistas de Cooperación internacional de la UCI.
--	--

Las medidas se clasifican dependiendo de su aditividad en:

- Aditivas: cuando pueden ser agregadas y no pierde sentido su valor
- Semi aditivas: cuando pueden ser agregadas en algunas dimensiones pero en otras no
- No aditivas: cuando no pueden ser agregadas.

Además dependiendo del tipo de la variable que puede ser:

Física cuando la variable es algo físico que debe estar en el almacén

Calculable: es cuando la variable se calcula de alguna manera teniendo en cuenta medidas físicas del almacén, y se calculan a través de BI.

En la tabla 5 se detallan las 12 medidas contenidas en cada uno de los hechos anteriormente descritos. Los tipos de datos de ellas son *integer* y *float*.

Tabla 5 Descripción de las medidas

Medidas	Clasificación	Descripción
cantidad_carnet_extranjeros	No aditiva Física	Refleja la cantidad de extranjeros con carnet de residencia temporal gestionados por la UCI.
total_importes_cartas_cad	No aditiva Física	Muestra el total del importe por cartas al CAD.
total_acuerdos	No aditiva Física	Refleja el total de acuerdos de colaboración internacional con los que cuenta la UCI.
total_materiales	No aditiva Física	Muestra el total de materiales gastados por los especialistas del área de Cooperación internacional.
cantidad_misiones	No aditiva Física	Comprende el total de misiones de la UCI.
cantidad_misioneros	No aditiva Física	Comprende el total de misioneros de la UCI.
cantidad_prorrogas	No aditiva Física	Comprende el total de las prórrogas gestionadas a los misioneros de la UCI.

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

cantidad_pasaportes	No aditiva Física	Comprende el total de pasaportes gestionados por la UCI.
importe	No aditiva Física	Comprende el importe del presupuesto del área de Cooperación internacional de la UCI.
cantidad_visitas	No aditiva Física	Refleja el total de visitas que ha recibido la UCI.
cantidad_regalos_entregados	No aditiva Física	Muestra el total de regalos que han sido entregados a los profesores invitados y personalidades visitantes a la UCI.
cantidad_tramites	No aditiva Física	Comprende el total de trámites gestionados por la UCI.

Matriz dimensional

Para validar el correcto diseño del modelo de datos se realiza la matriz dimensional. Esta se emplea para mostrar la relación existente entre los hechos y las dimensiones y de esta manera garantizar que diferentes hechos no compartan exactamente las mismas dimensiones. A continuación se muestra la matriz dimensional correspondiente al MD Cooperación internacional (ver figura 5).

Las celdas marcadas revelan que la fila del hecho está relacionada con la columna de la dimensión. Al analizar el resultado de esta matriz estos arrojaron que las dimensiones dim_tiempo_dia, dim_pais, dim_mision y dim_misionero son compartidas por múltiples hechos. Además en ella se evidencia la inexistencia de solapamiento entre los hechos. Por lo que se puede afirmar que el diseño del modelo de datos del MD Cooperación internacional es correcto. Seguidamente se presenta la matriz bus del MD Cooperación internacional:

hech/dim	dim_1	dim_2	dim_3	dim_4	dim_5	dim_6	dim_7	dim_8	dim_9	dim_10	dim_11	dim_12	dim_13	dim_14	dim_15	dim_16	dim_17	dim_18	dim_19	dim_20
hech_1	x	x			x	x						x								
hech_2	x	x					x													x
hech_3		x	x	x				x			x									
hech_4		x							x											
hech_5		x									x									
hech_6	x	x											x							
hech_7	x	x												x	x					
hech_8		x														x	x			
hech_9		x																	x	
hech_10		x					x													
hech_11		x												x	x					x

hech_visitas: hech_1	hech_materiales: hech_5	hech_pasaportes: hech_5
hech_colaboracion_internacional: hech_2	hech_carnet_extranjeros: hech_6	hech_prorrogas: hech_9
hech_presupuesto: hech_3	hech_mision: hech_7	hech_regalos_entregados: hech_10
hech_cartas_cad: hech_4		hech_tramites: hech_11

dim_pais: dim_1	dim_regalos: dim_7	dim_extranjero: dim_13
dim_tiempo_dia: dim_2	dim_acuerdos_colaboracion: dim_8	dim_mision: dim_14
dim_tipo_moneda: dim_3	dim_partida: dim_9	dim_misionero: dim_15
dim_tipo_importe: dim_4	dim_cartas_cad: dim_10	dim_estado_pasaporte: dim_16
dim_visitas: dim_5	dim_tipo_material: dim_11	dim_tipo_pasaporte: dim_17
dim_actividad: dim_6	dim_clasificacion: dim_12	dim_estado_prorroga: dim_18
dim_tramite: dim_19		
dim_firma_uci: dim_20		

Figura 5 Matriz dimensional

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

Esquemas multidimensionales

Esquema en estrella

Consta de una tabla de hechos central y varias tablas de dimensiones relacionadas a este, a través de sus respectivas llaves. Su diseño es fácilmente modificable y mejora los tiempos de respuesta ante las consultas de los usuarios, debido a que existe paralelismo entre su diseño y la forma en que ellos visualizan y analizan los datos (2).

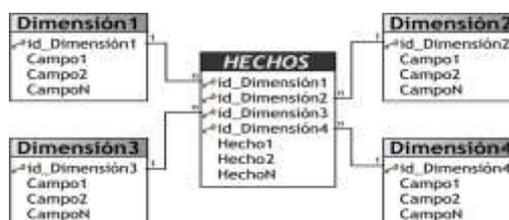


Figura 6 Ejemplo del esquema en estrella

Esquema copo de nieve

Constituye una extensión del modelo en estrella, cuando las dimensiones se ordenan teniendo en cuenta las jerarquías de las dimensiones. La diferencia del esquema copo de nieve en comparación con el esquema en estrella, radica en la estructura de las tablas de las dimensiones, que en este esquema se encuentran normalizadas (2).

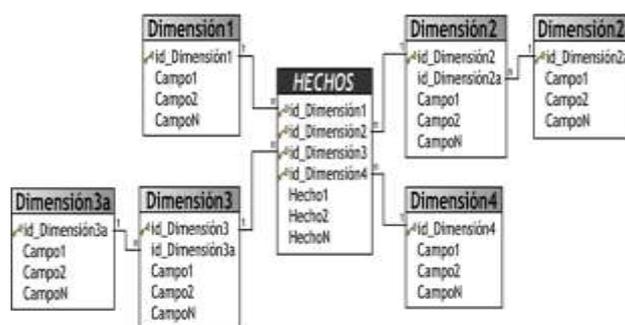


Figura 7 Ejemplo del esquema copo de nieve

Esquema constelación de hechos

Está formado por una sucesión de esquemas en estrella, se conforma por una tabla de hecho principal y por una o más tablas de hechos auxiliares, las cuales pueden ser agregaciones de la principal. Estas tablas se encuentran centradas en el modelo y se relacionan con sus correspondientes tablas de dimensiones.

Para esta investigación se seleccionó el esquema constelación de hechos, debido a que constituye el más apropiado teniendo en cuenta que su diseño parte de la unión de varios esquemas en estrella, por lo que es fácilmente modificable de acuerdo con las nuevas características o necesidades que surjan en el área.

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

Modelo de datos

Como salida principal del diseño se presenta el modelo de datos del MD Cooperación internacional, en el cual se relacionan las dimensiones, los hechos y las medidas identificadas previamente mediante la utilización del esquema constelación de hechos.

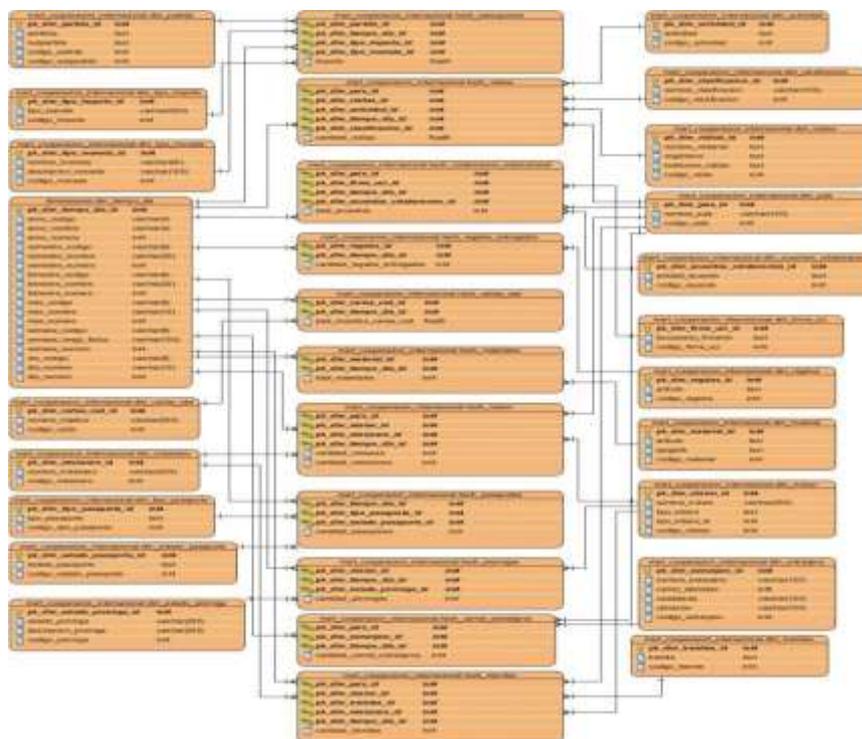


Figura 8 Modelo de datos

2.4.2 Diseño del subsistema de integración de datos

Para garantizar que la implementación del subsistema de integración de datos se realice de forma exitosa primero se debe concebir un diseño que lo soporte, teniendo en cuenta la mayor cantidad de aspectos y con el mayor nivel de detalle posible.

Arquitectura de integración

La arquitectura de integración de datos presentada a continuación (ver figura 9) forma parte de la arquitectura del MD, esta refleja cómo se realizará el proceso de ETL y se compone de los siguientes elementos:

- **Fuente de datos:** son los archivos pertenecientes a los sistemas fuente, los cuales son excel y BD, a esta última la conexión se realiza mediante protocolo TCP/IP.
- **Entorno de trabajo:** constituye el entorno donde se efectúa la preparación de los datos en el proceso de integración mediante las herramientas que se mencionaron en la definición de la arquitectura del MD.

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

- **Metadatos:** son los datos que se capturan para controlar la ejecución de los procesos de integración de datos.
- **Mercado de datos:** destino final de la carga de los datos integrados.



Figura 9 Arquitectura de integración

Perfilado de los datos

El perfilado de los datos constituye una de las primeras tareas en el proceso de calidad de datos, esta consiste en ejecutar un primer análisis sobre los datos fuentes, usualmente sobre las tablas, con el objetivo de comenzar a conocer su estructura, formato y nivel de calidad (30). Al realizar el perfilado de los datos a las fuentes del MD Cooperación internacional se identificó que los tipos de datos de las fuentes son en su mayoría *varchar* e *integer*, pero además existe la presencia de fechas desde el año 2002 hasta la actualidad con formato *timestamp* y en la fuente *Presupuesto.xls* los datos son *float*, lo que permitió definir en el modelo de datos mostrado anteriormente los tipos de datos (ver figura 10). Además se encontraron 612 valores nulos, lo que permitió crear nuevas RN que serán solucionadas en la implementación de las transformaciones. Igualmente no se hallaron valores vacíos ni negativos. Asimismo 363 fue el mayor valor de un campo de texto encontrado en la columna Clasificación o Cargo de la fuente *Visitas.xls*, lo que trajo como consecuencia que el tipo de dato para esta columna fuera definido como *text* (ver figura 11).

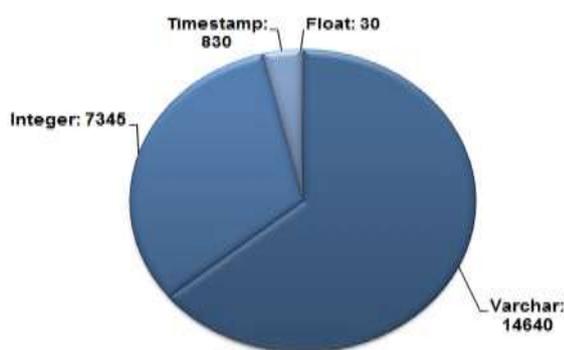


Figura 10 Distribución de la fuente por tipos de datos



Figura 11 Calidad de los datos

Diseño general del proceso de integración de datos

El diseño del proceso de ETL se realiza a partir los resultados arrojados en el análisis de los datos y teniendo en cuenta la estructura de la BD. En este es importante tener en cuenta todos los factores

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

que intervienen en este proceso, para el caso del MD Cooperación internacional se comienza conectándose a la BD fuente nombrada BD_sgu_cooperacion mediante protocolo TCP/IP, luego se extraen los datos desde las fuentes y se seleccionan los que transformarán, posteriormente se realiza la limpieza y estandarización de los mismos. Luego se obtienen los identificadores de las dimensiones y a continuación se validan los datos, si existe algún error se envía a la Salida excel creada para el tratamiento de los mismos. Seguidamente se cargan los datos en la tabla de salida de la transformación, se actualizan las fechas de inicio y fin y se obtienen la información del sistema para generar los metadatos. Finalmente se generan los metadatos históricos. Este diseño queda conformado de la siguiente forma:

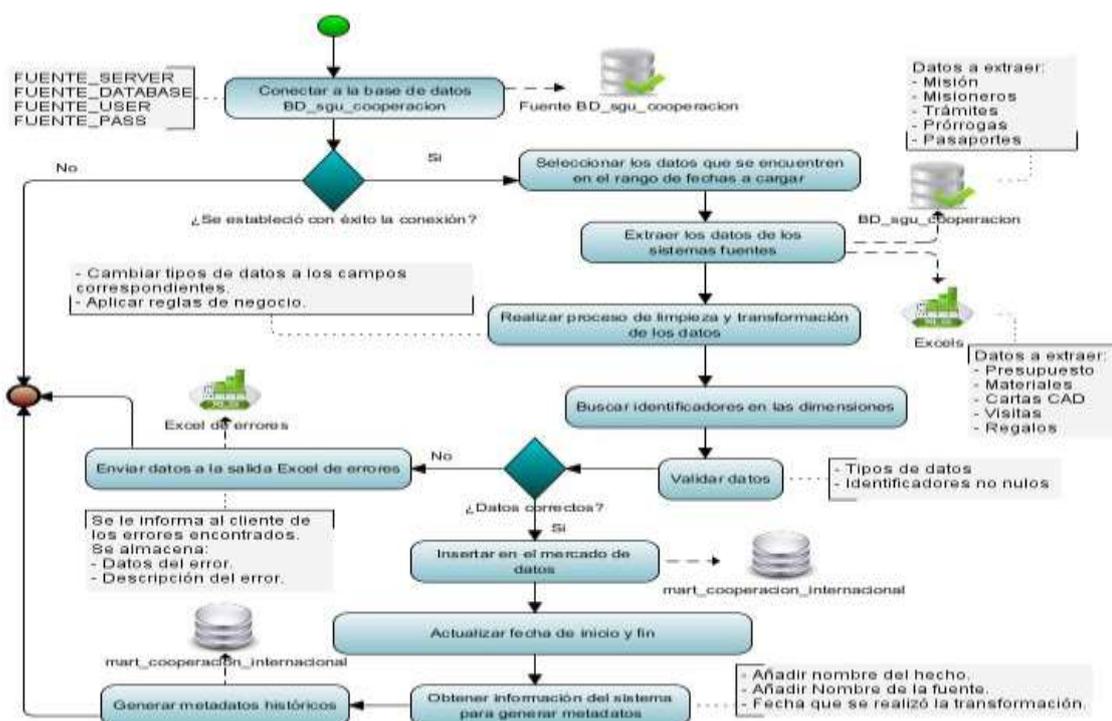


Figura 12 Diseño del proceso de integración de datos

Registros de sistemas fuentes

Los datos se capturan desde dos tipos de fuentes distintas desde los excel de forma anual y la información de la BD perteneciente al módulo de Cooperación internacional del sistema de gestión universitaria, se obtendrá mensualmente mediante una conexión por protocolo TCP/IP al servidor donde se encuentra la misma montada sobre el SGBD PostgreSQL. Mediante el artefacto “Registros de sistemas fuentes” se describe detalladamente las fuentes de datos correspondientes al área Cooperación internacional, este se realiza para lograr una mayor comprensión de la información contenida en las mismas. Con esto se logró dejar claramente establecido el estado físico de los datos,

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

quiénes son los responsables de la gestión de los mismos, entre otras características del sistema fuente.

Diccionario de datos

Para describir la información necesaria para la correcta comprensión del sistema fuente se emplea el artefacto “Diccionario de datos”, este está enfocado fundamentalmente en las variables que se relacionan con la fuente de datos, también se utiliza para establecer una correspondencia entre los valores la fuente y los que se manipulan el MD. Siendo este de mucha importancia para la implementación de las transformaciones, debido a que detalla la totalidad de las variables y el significado que tienen en el negocio.

Mapa lógico de datos

El “Mapa lógico de datos” es otro de los artefactos definidos por la metodología para el diseño del ETL, este se utiliza para establecer una correspondencia entre los datos de las fuentes y los propios del MD. Mediante el uso de este artefacto se logró identificar que las tuplas y los nombres de las columnas recogían información de interés para el mismo. Pudiéndose de esta forma definir los datos a normalizar y establecer una trazabilidad de los mismos desde la fuente hasta su destino final en el MD.

2.4.3 Diseño del subsistema de visualización

El subsistema de visualización constituye un elemento fundamental para el MD, debido a que el usuario final interactúa con él. Por ello realizar un correcto y detallado diseño del mismo brinda muchas facilidades en la implementación.

Arquitectura de información

La arquitectura de información posibilita organizar los contenidos para que el acceso a ellos se pueda realizar de forma más simple y directa. Para el caso del MD Cooperación internacional (ver figura 13) esta define la estructura de navegación del mismo y cuenta con un área análisis general (A.A.G), un área de análisis (A.A) y cuatro libros de trabajo (L.T). Cada L.T contiene las vistas de reportes, sumando entre todos un total de 44.

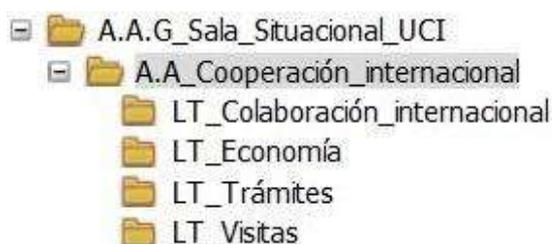


Figura 13 Arquitectura de información

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

A.A.G_Sala_Situacional_UCI: área análisis general Sala Situacional UCI, contiene la información de todos los MD que componen el almacén.

A.A_Cooperación_internacional: área de análisis Cooperación internacional, controla la información referente al trabajo de esta dirección de la UCI que están contenida en los LT.

LT_Colaboración_internacional: agrupa la información referente a los acuerdos de colaboración que mantiene la UCI con otras entidades del mundo. Contiene un total de cuatro reportes, que permiten realizar un análisis de los datos de ese departamento.

LT_Economía: agrupa la información relativa al departamento de economía de Cooperación internacional. Comprende siete reportes, con los que se pueden analizar los datos de ese departamento.

L.T_Trámites: agrupa la información de todo lo relacionado con las entradas y salidas al país de trabajadores y estudiantes. Contiene 23 reportes para analizar la totalidad de la información de esa área.

LT_Visitas: agrupa la información de las visitas que recibe la UCI. Contiene diez reportes, que posibilitan el análisis de los datos de este departamento.

Diseño de los reportes candidatos

En el diseño del subsistema de visualización es importante detallar los reportes candidatos, estos dan respuesta a las necesidades de los usuarios, permitiendo identificar las salidas que debe producir el sistema y los principales elementos del reporte. A continuación se muestra un ejemplo del reporte candidato “Total de acuerdos de colaboración por directivo UCI firmante”.

Tabla 6 Reporte candidato “Total de acuerdos de colaboración por directivo UCI firmante”

Área de análisis (AA)	Colaboración internacional
Libro de Trabajo (LT)	LT1 –Colaboración_internacional
Reporte (Tabla de Salida – TS)	TS1 – Colaboración internacional
Descripción	Reporte que muestra el total de acuerdos de colaboración firmados por algún directivo de la UCI
Elementos del reporte	<ul style="list-style-type: none">• Perspectivas de análisis dim_firma_uci

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

	<ul style="list-style-type: none">Indicadores a medir total_acuerdos
Frecuencia de emisión	Anual
Funciones	Suma de los acuerdos de colaboración firmados por cada directivo
Gráfico	Pastel por filas

2.5 Política de respaldo y recuperación

Contar con una política de respaldo y recuperación para cualquier sistema informático es de imprescindible orden, debido a que esto garantiza la disponibilidad y protección de los datos que se almacenan en ellos y además mantiene la capacidad de recuperar los datos perdidos tras un incidente de seguridad, para el MD Cooperación internacional la política de respaldo y recuperación cuenta con tres aspectos esenciales:

- Periodicidad de salvallas del sistema: las salvallas se llevarán a cabo durante todo el año, salvallas parciales con periodicidad mensual y salvallas totales en los meses de junio y diciembre. Para el primer caso se utiliza un tipo de copia incremental, es decir, sólo se copiarán los archivos creados o modificados desde la última copia de seguridad normal o incremental. Teniendo en cuenta siempre que se debe verificar previamente la existencia de una copia de toda la información almacenada. Para las salvallas a realizar en los meses de junio y diciembre se contará con una copia de seguridad de copia, dígame que todos los archivos seleccionados son copiados pero no se marcan particularmente como copiados.
- Tablas involucradas: las tablas que normalmente se involucran en la realización son los 11 hechos con los que cuenta el MD. Además de las tablas de dimensiones que hayan sufrido algún tipo de cambio.
- Salvallas existentes: actualmente en el área Cooperación internacional se realizan salvallas de los datos pero sin periodicidad definida. La información salvada se almacena en la oficina del informático del área en formato de Disco Compacto (CD).

2.6 Esquema de seguridad

El desarrollo de un completo esquema de seguridad para una aplicación informática, garantiza en gran medida que los datos no se dañen o alteren por circunstancias o factores externos, brindando de esta forma protección, resguardo y recuperación a ellos. La seguridad para el MD Cooperación internacional se maneja en gran medida por los niveles de acceso al sistema, teniendo como apoyo los roles creados para la interacción de los usuarios con el sistema.

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

Seguridad en la base de datos

Para acceder a la BD se definieron los roles que se describen a continuación en la tabla 7. Esto tiene como objetivo que cada rol maneje el acceso a las funcionalidades de acceso a los datos que le corresponden específicamente.

Tabla 7 Usuarios en la BD

Usuario	Descripción
Administrador ETL	Cuenta con permiso sólo para la realización de las funciones del proceso de ETL.
Administrador	Cuenta con acceso total a la BD. Ejecuta las funciones de administración en la misma.
Analista	Cuenta con acceso de sólo lectura a la BD. Accede a la misma solo mediante la aplicación.

Seguridad en el proceso de ETL

En el proceso de ETL se garantiza que los datos de las fuentes, las transformaciones y los trabajos no sean manipulados mediante la seguridad que ofrece el sistema operativo, al marcar el atributo de sólo lectura en las propiedades de la carpeta donde se almacenan todos ellos (ver figura 14). De esta forma sólo puede modificar estos ficheros el usuario que tenga permisos de escritura para ello.

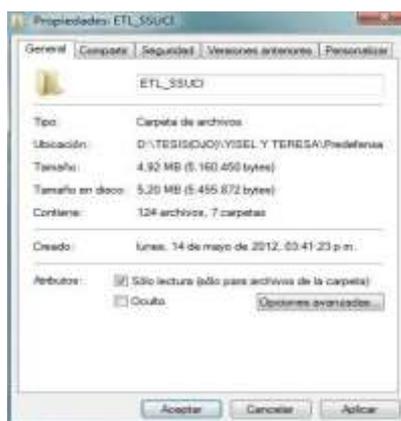


Figura 14 Seguridad en el proceso de ETL

Seguridad en la aplicación

Con el objetivo de garantizar que accedan al sistema sólo los usuarios definidos por el cliente, es importante la creación de un esquema de seguridad que permita la sostenibilidad en la gestión de la misma en este servidor. A continuación se muestran los roles delimitados con el objetivo de que el acceso a la aplicación se haga de forma organizada y con los permisos correspondientes.

Capítulo 2: Análisis y diseño del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

Tabla 8 Usuarios en la aplicación

Actores	Descripción
Administrador	Tiene acceso total al sistema Sala Situacional UCI. Realiza las operaciones de administración sobre el mismo.
Analista	Tiene acceso de sólo lectura al MD Cooperación internacional. Analiza la información contenida en los reportes.

Para alcanzar mejoras en cuanto a organización y rendimiento en la seguridad del sistema, la herramienta Pentaho BI server 3.10 ofrece tres aspectos fundamentales, estos son (25):

- Seguridad de acceso a datos de objetos: contiene los usuarios, contraseñas, autorizaciones permitidas, recursos web y protección a datos.
- Autenticación: se compone de todo lo relacionado con el procesamiento de información interactiva de inicio de sesión.
- Autorización de recursos web (URL): brinda protección a las URL para gestionar el acceso a las páginas por parte de un usuario determinado.

2.7 Conclusiones del capítulo

Después de realizar los procesos de análisis y diseño del MD Cooperación internacional se concluye que con el levantamiento de requisitos se obtuvo como principal resultado la identificación de 16 RI, 18 RF y 10 RFN, permitiendo a partir de ellos la elaboración del DCUS. Asimismo, se identificaron 13 RN, posibilitando de esta forma la definición de las reglas de transformación que serán utilizadas en los procesos de integración de datos. Además, se efectuó la descripción de los CU del MD permitiendo describir las funcionalidades del sistema. Con la definición de la arquitectura del MD, se posibilitó identificar los elementos y subsistemas involucrados en el desarrollo de la solución. También se diseñó el modelo de datos, que permite representar las relaciones entre las 11 tablas de hechos y las 20 dimensiones identificadas para el MD. Con el diseño de los subsistemas de integración de datos y visualización, se logró definir el diseño general del proceso de ETL, la arquitectura de información del MD y el diseño de los reportes candidatos, todo esto permitió guiar la implementación de dichos subsistemas. Mediante las políticas de respaldo y recuperación creadas, se definió la periodicidad de las salvadas de forma semestral y las tablas de la BD implicadas en esto. Además los roles y permisos definidos en el MD, se permite colaborar con la seguridad del sistema.

Capítulo 3: Implementación del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

CAPÍTULO 3: IMPLEMENTACIÓN DEL MERCADO DE DATOS COOPERACIÓN INTERNACIONAL PARA LA SALA SITUACIONAL DE LA UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

Este capítulo comprende la implementación del subsistema de almacenamiento, mediante la construcción del modelo físico y la definición de los estándares de codificación. Además incluye la implementación del subsistema de integración de datos a través de la gestión del cambio en las dimensiones, de los metadatos, la captura de los cambios en los datos y la implementación de los hechos y los trabajos. También se realizó la implementación del subsistema de visualización, el cual comprende la implementación de los cubos OLAP y los reportes con los que se le da solución a las necesidades del negocio.

3.1 Implementación del subsistema de almacenamiento

3.1.1 Modelo físico

El modelo físico constituye una colección integrada de entidades que describe las estructuras de los datos, las restricciones de integridad y las operaciones de manipulación de los mismos. Dicho modelo se genera a partir del modelo lógico dimensional; conteniendo las relaciones entre las tablas de hechos y dimensiones que conforman el MD. En la solución este modelo se compone por tres esquemas, 72 secuencias, 55 índices y 35 tablas.

Para garantizar un acceso más rápido de las consultas realizadas a los datos del MD Cooperación internacional se implementaron los índices, elaborándose 55 de tipo *clustered* en los hechos para llaves de las dimensiones relacionadas con ellos y las medidas propias de cada uno. Estos índices se implementaron mediante la estrategia *btree* (árboles B), debido a que proporcionan acceso rápido para consultas de gran complejidad y constituye una de las estrategias que brinda mejores resultados cuando se realizan comparaciones igual-igual (31).

Para proveer una mejor organización de las tablas en el MD se crearon tres esquemas, “dimensiones”, “metadatos” y “mart_cooperacion_internacional” la función de cada uno de ellos se describe en los estándares de codificación y las tablas contenidas en cada uno de ellos se mencionan a continuación:

Tabla 9 Esquemas y tablas del MD

Esquemas	Tablas
 dimensiones	 dim_tiempo_dia
	 hech_visitas

Capítulo 3: Implementación del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

 mart_cooperacion_internacional	 hech_colaboracion_internacional
	 hech_presupuesto
	 hech_cartas_cad
	 hech_materiales
	 hech_carnet_extranjeros
	 hech_mision
	 hech_pasaportes
	 hech_prorrogas
	 hech_regalos_entregados
	 dim_pais
	 dim_tipo_moneda
	 dim_tipo_importe
	 dim_visitas
	 dim_regalos
	 dim_acuerdos_colaboracion
	 dim_firma_uci
	 dim_partida
	 dim_cartas_cad
	 dim_tipo_material
	 dim_clasificacion
	 dim_actividad
	 dim_extranjero
	 dim_mision
	 dim_misionero
	 dim_tipo_pasaporte
	 dim_estado_pasaporte
 dim_estado_prorroga	
 metadatos	 transformación
	 trabajo
	 gest_cambio_datos_fuentes
	 gestión_histórico_datos fuente

Capítulo 3: Implementación del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

A continuación se presenta como queda estructurada la BD en la herramienta pgAdmin:



Figura 15 Estructura de la BD

3.1.2 Estándares de codificación

Los estándares de codificación constituyen un conjunto de normas para el personal implicado en el desarrollo de un proyecto entienda la codificación de cada atributo del mismo. Los utilizados en el MD Cooperación internacional se muestran a continuación:

Tabla 10 Estándares de codificación del MD Cooperación internacional

Tipo de objeto	Función	Nomenclatura	Descripción
Esquema	Dimensiones compartidas	dimensiones	Esquema donde se encuentran las dimensiones compartidas por varios MD.
	Metadatos	Metadatos	Esquema donde se encuentran las tablas que almacenan los metadatos.
	Tablas específicas del MD	mart_[nombre MD]	Esquema donde se almacenan las tablas de hechos y dimensiones definidas para recoger los datos asociados a cada área.
Tablas	Tablas de dimensiones	dim_[nombre dimensión]	Todas las tablas de dimensiones tendrán una cadena (dim_) y el concepto que describen.

Capítulo 3: Implementación del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

	Tablas de hechos	hech_[nombre hecho]	Todas las tablas de hechos tendrán una cadena (hech_) y el concepto que describen.
	Tablas de metadatos	[nombre metadatos]	Todas las tablas de metadatos tendrán el concepto que describen.
Contrains	Llaves primarias	pk_[nombre tabla]_id	Todas las llaves primarias tendrán una cadena que lo identifica (pk_) y el nombre de la tabla a la que pertenecen, y además tendrán el identificador (_id), serán de tipo integer y autoincremental.
	Llaves foráneas	fk_[nombre tabla]_id	Todas las llaves foráneas tendrán una cadena que las identifica (fk_) y el nombre de la tabla a la que pertenecen, y además tendrán el identificador (_id), serán de tipo integer y autoincremental.
Índices	Índices	ind_[nombre atributo]	Los índices se nombrarán con una cadena que los identifique (ind_) y a continuación el nombre del atributo de la tabla al que hace referencia.

3.2 Implementación del subsistema de integración de datos

La implementación del subsistema de integración de datos requiere gran parte de esfuerzo en el desarrollo del MD. En este proceso se tienen en cuenta una serie de elementos que van desde el perfilado de los datos fuentes para conocer el estado en que se encuentran, hasta la carga de estos en el MD.

3.2.1 Gestión del cambio en las dimensiones

Durante la implementación del subsistema de integración, se procedió primeramente a realizar un total de 20 transformaciones correspondientes a la carga de las dimensiones. La totalidad de estas dimensiones son lentamente cambiantes.

Las dimensiones lentamente cambiantes o SCD (*Slowly Changing Dimensions*) son dimensiones en las cuales sus datos tienden a ser modificados en el transcurso del tiempo. Estos cambios se pueden realizar de forma eventual o constante, y puede implicar a un solo registro o la tabla completa. Cuando ocurren estos cambios, se puede optar por seguir alguna de estas dos grandes opciones (9):

Capítulo 3: Implementación del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

- Registrar el historial de cambios.
- Reemplazar los valores que sean necesarios.

La estrategia a utilizar para la solución será registrar el historial de cambios. Para esto existen diversas estrategias a seguir las cuales se muestran inmediatamente:

- SCD Tipo 0: enfoque pasivo, no se realiza nada para ello.
- SCD Tipo 1 (sobrescribir): empleado cuando no es valiosa la información histórica.
- SCD Tipo 2 (añadir fila): enfoque que se plantea insertar un nuevo registro cada vez que se realiza un cambio a la dimensión.
- SCD Tipo 3 (añadir columna): controla el cambio agregando nuevas columnas (9).
- SCD Tipo 4 (tabla de historia separada): almacena en otra tabla los detalles de los cambios históricos.
- SCD Tipo 6 (híbrido): estrategia que combina los tipos 1, 2 y 3.

Las dimensiones pertenecientes al MD Cooperación internacional se clasifican como SCD tipo 2. Para garantizar esto se utilizó el componente Insertar/Actualizar marcando la opción de no realizar actualizaciones (ver figura 16), mediante esto se asegura que no se realicen actualizaciones en el MD sino que se inserta una nueva fila en las tablas.

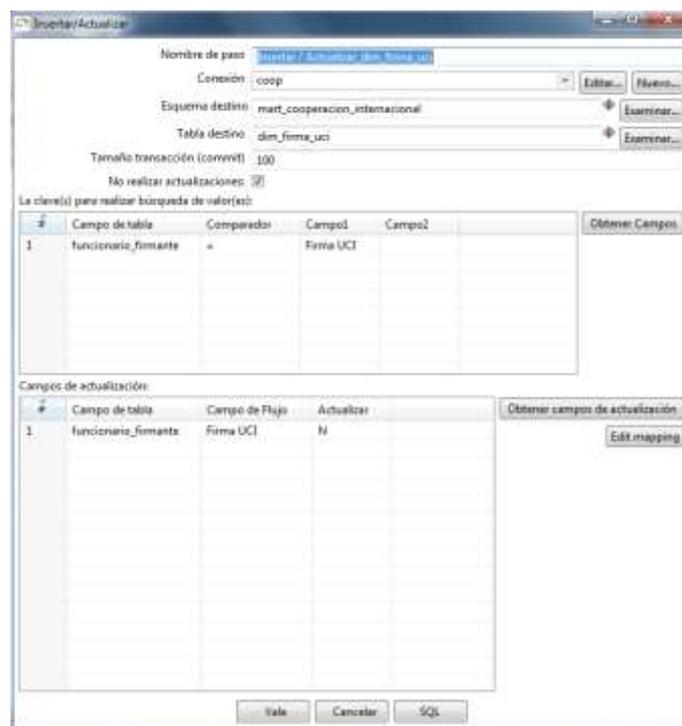


Figura 16 Componente Insertar/ Actualizar

Capítulo 3: Implementación del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

3.2.2 Gestión de los metadatos del proceso de integración de datos

Los metadatos son datos que detallan y/o posibilitan encontrar, gestionar, fiscalizar, comprender y resguardar otros datos; es decir, son datos sobre los propios datos (9). Ellos se clasifican en varios tipos, como los siguientes:

- Metadatos administrativos: empleados para el control y administración de los recursos de información.
- Metadatos descriptivos y de descubrimiento: se usan para detallar y revelar los recursos de información.
- Metadatos técnicos o modelos: se aplican para describir el funcionamiento de un sistema o el modo en que se relacionan sus componentes.
- Metadatos de proceso: utilizados para capturar la información de los procesos en que se ejecutan.
- Metadatos de negocio: permite obtener los datos y la información que describen el negocio, usualmente son los datos fuente.

Los metadatos de proceso fueron los utilizados para esta investigación con el objetivo de obtener la información de las transformaciones y los trabajos pertinentes a los subprocesos de ETL.

3.2.3 Captura de los cambios en los datos

El conocimiento del flujo de información dentro de los diferentes departamentos del área Cooperación internacional es de vital importancia para determinar la forma en que se realizará la captura de los cambios. La captura de la información perteneciente a trámites se lleva a cabo desde los proyectos, centros productivos y facultades hasta llegar a esta área, la información de los restantes departamentos es gestionada en la propia área, donde se realizan las operaciones pertinentes por los especialistas de cada departamento.

Teniendo en cuenta lo descrito previamente se define la obtención de los cambios en los datos en el mercado anualmente, este cambio se gestiona mediante la tabla de metadatos *gestion_cambio_datos_fuente*. Esta tabla almacena las fechas que corresponde cargar de cada uno de los hechos. Al realizar esta carga, se actualizan las fechas de inicio y fin; posibilitando así conocer la fecha en que se realizó la última carga al MD. Además se gestiona el cambio histórico de los datos mediante la tabla *gestión_histórico_datos fuente*.

Capítulo 3: Implementación del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

3.2.4 Implementación de los hechos

Luego de cargar las dimensiones se realizaron 11 transformaciones para la carga de los hechos. En las cuales se utilizaron un conjunto de subsistemas que Kimball propone para la implementación del proceso de ETL, a continuación se explica la transformación del hecho colaboración internacional:

A través de la utilización del subsistema de extracción se obtuvieron los datos desde los ficheros *Acuerdos de Colaboración.xls*. Luego de realizarse la extracción de los datos se procedió a la limpieza y transformación de los mismos, proceso en el cual se estandarizó los datos y se aplicaron las reglas de transformación identificadas con anterioridad; también se capturaron los datos que no cumplen con ellas, a través del subsistema rastreo de eventos de errores. A estos errores se les dará tratamiento mediante una estrategia que propone informarle al cliente los campos que tuvieron irregularidades para que en conjunto con el equipo de soporte del MD decidir la solución que sea más factible para ambos. Además se utiliza el subsistema repositorio de metadatos para implementar el registro histórico de las cargas y de la fuente.

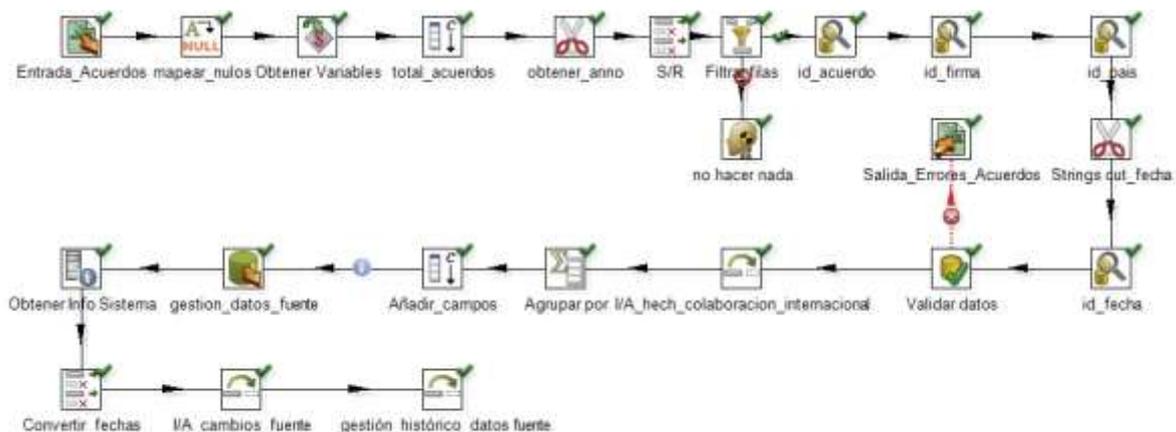


Figura 17 Transformación hecho colaboración internacional

3.2.5 Implementación de los trabajos

Luego de finalizada la totalidad de las transformaciones requeridas para la carga de los datos, se llevó a cabo la implementación de los trabajos, mediante la aplicación del subsistema programador de trabajos, estos son los encargados de ejecutar todas las transformaciones en un orden lógico predeterminado. Para el caso del MD Cooperación internacional se creó un trabajo para llenar las dimensiones compartidas por varios hechos, 10 para cargar los hechos, los cuales ejecutan las transformaciones propias de cada hecho y posteriormente el hecho y un trabajo general que es el encargado de ejecutar el resto de los trabajos (ver figura 18). Con lo anteriormente se garantiza que si

Capítulo 3: Implementación del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

no se cargan las dimensiones compartidas, el trabajo se aborte y además que se puedan cargar algunos hechos y otros no, dependiendo si en alguna de sus dimensiones propias surja algún error.

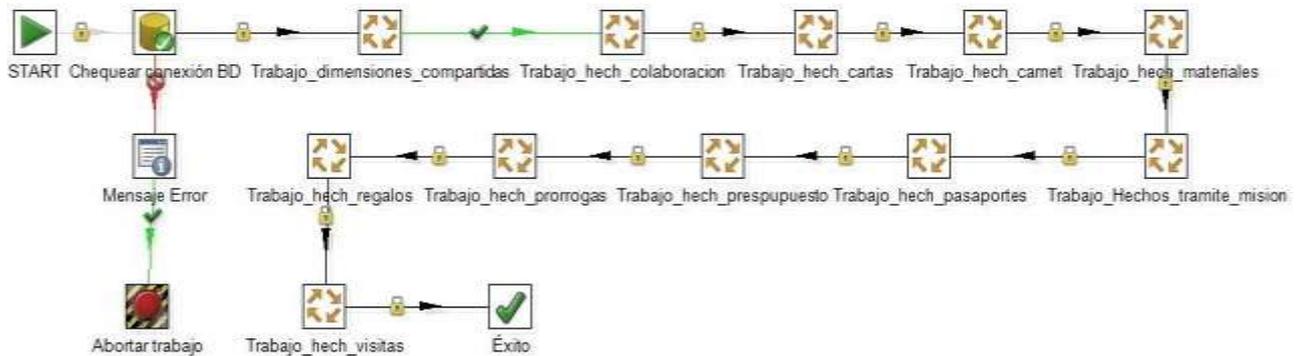


Figura 18 Trabajo general

3.3 Implementación del subsistema de visualización

La implementación del subsistema de visualización es un proceso muy importante en el desarrollo de un MD, debido a que define qué información se presenta para los usuarios y cómo se muestra esta para el análisis.

3.3.1 Implementación de los cubos OLAP

Como parte de la implementación del subsistema de visualización se desarrollaron un total 11 cubos OLAP, uno por cada hecho. Estos permiten estructurar los datos para que concuerden con el modo que tienen los usuarios de analizarlos.

En este proceso inicialmente se definieron las dimensiones compartidas entre ellos, dentro de estas fueron primeramente creadas 20 jerarquías, en cada una de ellas se indicaron una serie de parámetros como el *hasAll* para agrupar todos los valores de la jerarquía y su descripción *allMemberName* en el caso de que esté marcado, se definió como *Todos* dependiendo del género de la dimensión. Se selecciona además la llave primaria de la jerarquía y por último el *caption* que define el nombre que va a representar los valores de dicha dimensión. Esta configuración se muestra en la siguiente figura:

Attribute	Value
name	
description	
hasAll	<input checked="" type="checkbox"/>
allMemberName	Todos
allMemberCaption	
allLevelName	
defaultMember	
memberReaderClass	
primaryKeyTable	
primaryKey	pk_dim_firma_uci_id
caption	Funcionario UCI

Figura 19 Jerarquía de la dimensión firma uci

Capítulo 3: Implementación del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

También se crearon 24 niveles que componen las jerarquías, para ello se indica en los atributos la tabla a la que pertenece, la columna de la BD que la distingue, el tipo de datos, el tipo de nivel, el atributo *hideMemberIf* y por último el *caption* que define el nombre que va a representar los valores de este nivel. Esto se realizó para cada uno de los niveles de las jerarquías definidas.

Attribute	Value
name	firma_uci
description	
table	dim_firma_uci
column	codigo_firma_uci
nameColumn	funcionario_firmante
parentColumn	
nullParentValue	
ordinalColumn	
type	String
uniqueMembers	<input checked="" type="checkbox"/>
levelType	Regular
hideMemberIf	Never
approxRowCount	
caption	Funcionario UCI
captionColumn	
formatter	

Figura 20 Nivel de la jerarquía firma uci

Posteriormente se crearon los 11 cubos OLAP, para los cuales se definieron el nombre y el *caption*, las opciones cache y la opción *enabled* para que el cubo sea visible.

Attribute	Value
name	hech_colaboracion_internacional
description	
caption	Colaboración internacional
cache	<input checked="" type="checkbox"/>
enabled	<input checked="" type="checkbox"/>

Figura 21 Cubo hecho colaboración internacional

Además se definieron 14 medidas, todas calculadas directamente con campos de la BD. Indicándose su nombre, función de agregación, la columna que la genera, se marca la opción visible en todos los casos, tipo de dato del atributo y el *caption*.

Attribute	Value
name	total_acuerdos
description	
aggregator	sum
column	total_acuerdos
formatString	
visible	<input checked="" type="checkbox"/>
datatype	Numeric
formatter	
caption	Total de acuerdos de colaboración

Figura 22 Medida hecho colaboración internacional

A continuación se muestran la totalidad de los cubos diseñados (ver figura 23):

Capítulo 3: Implementación del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas



Figura 23 Diseño de los cubos OLAP

3.3.2 Implementación de los reportes

Para la realización de los reportes se utilizó la herramienta Pentaho BI server 3.10, en la cual se realizaron varias configuraciones para el uso propio del MD. Como parte de esto se definió una conexión única para el AD nombrada ssuci_dw en la cual se establecieron los parámetros necesarios para acceder a la BD. Además se crearon dos roles: administrador y analista y varios usuarios con lo que se garantiza la seguridad del sistema.

Para la implementación de los reportes se emplearon consultas MDX. A continuación se muestra la consulta del reporte “Total de acuerdos de colaboración internacional por directivo de la UCI firmante”. La cual retorna como resultado los directivos de la UCI que han firmado algún acuerdo de colaboración y el total de acuerdos de colaboración que ha firmado cada uno.

```
SELECT NON EMPTY {[Measures].[total_acuerdos]} ON COLUMNS,  
NON EMPTY [dim_firma_uci].[Todos].CHILDREN ON ROWS  
FROM [hech_colaboracion_internacional]
```

Estos reportes serán empleados como punto de partida del proceso de toma de decisiones. A continuación se muestra la imagen de la vista de análisis “Total de acuerdos de colaboración internacional por directivo de la UCI firmante” perteneciente al “LT Colaboración_internacional” (ver figura 24).

Capítulo 3: Implementación del mercado de datos Cooperación internacional para la Sala Situacional de la Universidad de las Ciencias Informáticas

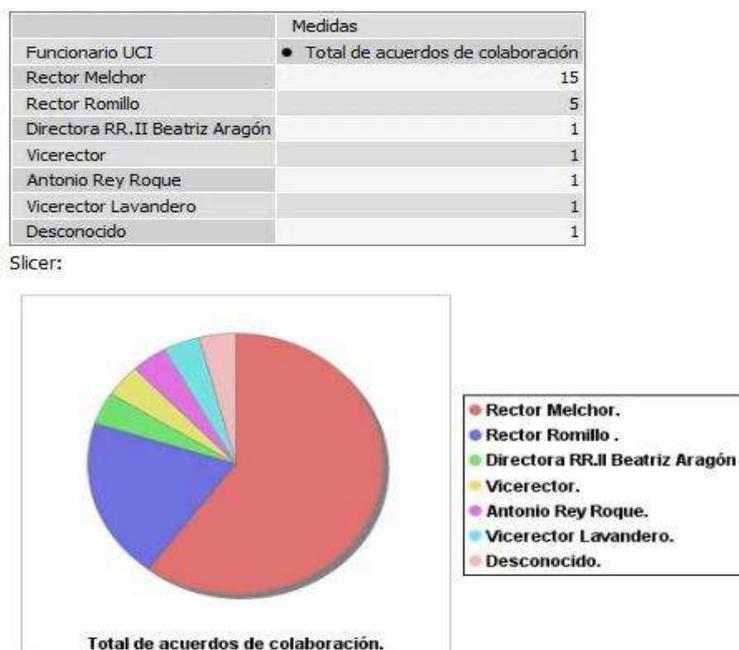


Figura 24 Vista del reporte “Total de acuerdos de colaboración internacional por país”.

3.4 Conclusiones del capítulo

En el presente capítulo se detallaron los elementos correspondientes a la implementación del MD Cooperación internacional y como conclusiones se plantean que con la implementación del subsistema de almacenamiento quedaron definidas las estructuras físicas del MD con tres esquemas, 35 tablas, N índices y 72 secuencias. También que con la implementación del subsistema de integración de datos, se lograron ejecutar 12 trabajos y 31 transformaciones: 20 para la carga de las dimensiones y 11 para la carga de los hechos, pudiendo de esta forma poblar el MD favorablemente. En este proceso se definió además que el tipo de SCD empleado es Tipo 2 lo que permitió la gestión del cambio en las dimensiones. Además los metadatos definidos fueron de proceso lo que permitió gestionar y capturar los datos al ejecutarse las transformaciones y los trabajos, garantizando que los mismos funcionen correctamente. Mediante la implementación del subsistema de visualización quedó establecida un AA, cuatro LT y 44 reportes, los cuales fueron implementados, permitiendo la visualización de la información. Además se crearon roles y permisos de los usuarios que accederán a la información, para brindar seguridad al sistema.

Capítulo 4: Realizar pruebas al mercado de datos Cooperación internacional para la Sala Situacional para la Universidad de las Ciencias Informáticas

CAPÍTULO 4: REALIZAR PRUEBAS AL MERCADO DE DATOS COOPERACIÓN INTERNACIONAL PARA LA SALA SITUACIONAL DE LA UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

En este capítulo se plasman los resultados arrojados por las pruebas aplicadas a la solución, mediante las listas de chequeo a los artefactos de ETL, los casos de pruebas (CP) por CU y por reglas de transformación y el perfilado de los datos para comprobar la calidad de la información cargada al MD. Estas pruebas son ejecutadas con el objetivo de avalar el cumplimiento de las exigencias del cliente y garantizar la calidad del sistema.

4.1 Pruebas de software

Para poder afirmar que se ha realizado un producto correctamente, primero se debe asegurar que este haya pasado por un adecuado proceso de pruebas. El software no queda exento de este proceso.

De acuerdo a la IEEE una prueba se define como: “Actividad en la cual un sistema o componente es ejecutado bajo condiciones específicas, se observan o almacenan los resultados y se realiza una evaluación de algún aspecto del sistema o componente” (32). Una prueba se considera exitosa si encuentra alguna deficiencia en el software. Para obtener diferentes tipos de errores en el sistema se hace necesario aplicar un amplio conjunto de pruebas (28).

A continuación se explican diversos tipos de pruebas que se le pueden realizar a un software:

Pruebas unitarias: la prueba unitaria o de unidad se centra en el módulo, estas constituyen comprobaciones que se le realizan a las unidades lógicas del sistema (28).

Pruebas de integración: tienen como objetivo captar los módulos probados en las pruebas de unidad y detectar errores relacionados con la interacción entre dichos módulos (28).

Pruebas de validación: se prueba el sistema como un todo para verificar si cumple con los RF y RNF (28).

Pruebas de aceptación: son realizadas con los clientes y define su aceptación del sistema (28).

4.2 Modelo V

Para validar el correcto funcionamiento de un software es importante ir probándolo desde el inicio de su construcción hasta después de aceptado por el cliente. Las actividades de control de la calidad del mismo deben tener un orden de ejecución lógico y funcional, por lo cual el centro DATEC define el uso del modelo V con estos fines.

Capítulo 4: Realizar pruebas al mercado de datos Cooperación internacional para la Sala Situacional para la Universidad de las Ciencias Informáticas

Este modelo tiene mucha similitud con el Modelo en cascada pues constituye una evolución del mismo. Lo que diferencia uno del otro es que el modelo V toma las actividades de prueba como un proceso que corre en paralelo con las actividades de análisis y diseño, en lugar de establecer una fase independiente al concluir el proyecto como define el Modelo en cascada (33).

A continuación se muestra una imagen del Modelo V (ver figura 25). Donde se refleja el ciclo de vida del software propuesta en el Modelo V, donde se muestran las diferentes etapas de desarrollo en la izquierda, mientras que las pruebas pertenecientes a cada una de ellas se exponen a la derecha.

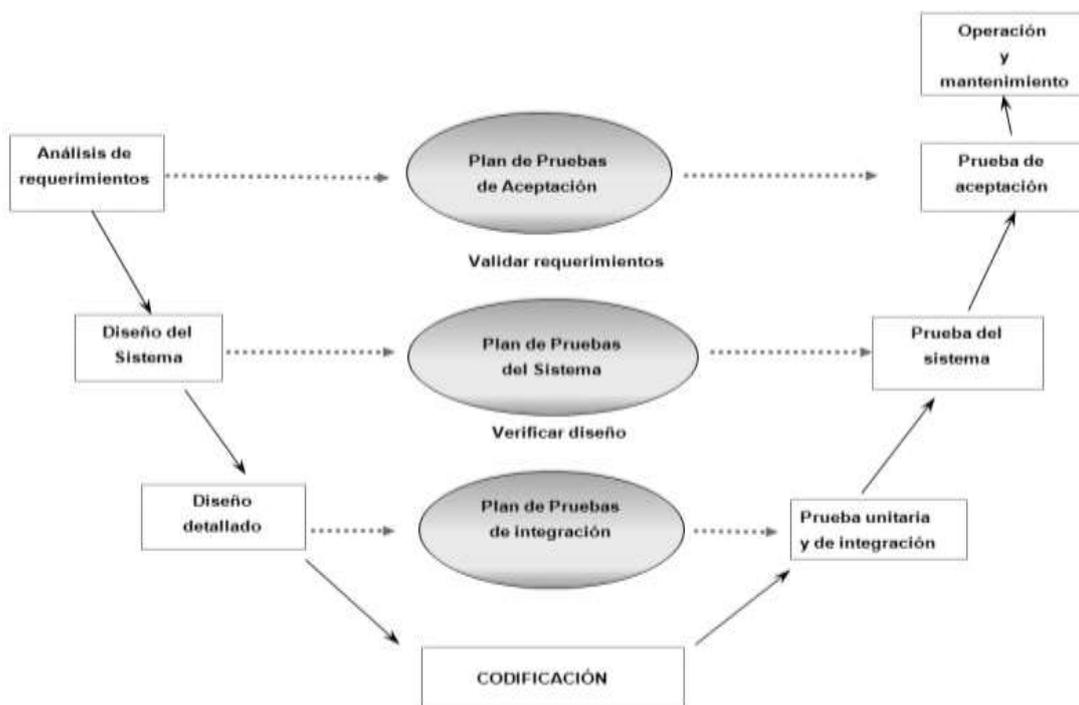


Figura 25 Modelo V

4.3 Herramientas para la aplicación de las pruebas

Como parte de la estrategia de prueba para aplicar los tipos de pruebas mencionados anteriormente al MD Cooperación internacional, se emplearon como herramientas los CP para CU y reglas de transformación, las listas de chequeo a los artefactos de ETL y el perfilado de los datos destino para probar la calidad de los mismos.

4.3.1 Aplicación de los casos de prueba por caso de uso

Los CP por CU se diseñan en correspondencia con las funcionalidades detalladas en los CU de información. Este diseño se confecciona antes de realizar las pruebas funcionales de la aplicación. Para el MD Cooperación internacional fueron diseñados cuatro CP.

Capítulo 4: Realizar pruebas al mercado de datos Cooperación internacional para la Sala Situacional para la Universidad de las Ciencias Informáticas

Después de realizado cada CP de validación, se cumplirá una de las dos condiciones posibles:

- La funcionalidad cumple con la especificación y se le acepta.
- Se descubre un error en la funcionalidad y se crea una lista con las no conformidades encontradas (28).

A continuación se muestra el CP “Presentar información de colaboración internacional” (ver tabla 11), para analizar los restantes CP realizados a la aplicación ir a los anexos.

Tabla 11 Caso de prueba “Presentar información de colaboración internacional”

Escenario	Descripción	VARIABLES DE ENTRADA	VARIABLES DE SALIDA	Respuesta del sistema	Flujo central
EC 1.1 Cubo de acuerdos de colaboración.	Permite visualizar los acuerdos de colaboración realizados por directivo de la Universidad de las Ciencias Informáticas que firma el acuerdo, por entidad, país y período.	V Funcionario UCI Entidad del acuerdo País Tiempo	V Total de acuerdos de colaboración	Se muestra la tabla con los valores correspondientes a cada escenario.	Se abre la aplicación. Se autentica. Entra al sistema. En la parte superior izquierda se selecciona el AA Cooperación internacional . Seguidamente se selecciona el LT Colaboración internacional . En la parte inferior izquierda se selecciona el reporte deseado. En el área de trabajo se visualiza la tabla correspondiente al reporte.

4.3.2 Aplicación de los casos de pruebas por las reglas de transformación

En el proceso de ETL también se aplicaron CP, en este caso para las reglas de transformación con la finalidad de comprobar si luego de ejecutada cada transformación, los datos cargados sean los esperados. Como ejemplo de este proceso se presenta el CP para la regla: Donde el campo firma_uci se encuentre vacío se pondrá el valor Desconocido (ver tabla 12).

Tabla 12 Caso de prueba para regla de transformación

Caso de prueba						
Nombre variable	firma_uci					
Escenario	hech_colaboracion_internacional					
Regla de transformación	Donde el campo firma_uci se encuentre vacío se pondrá el valor Desconocido					
Valor de entrada	Estado del dato	Resultado esperado	Respuesta del Sistema	Resultado Real	Comentario	Resultado de la Prueba
	No válido	Desconocido	El sistema transforma el dato aplicando	Desconocido	La prueba se realizó sin ocurrir	Satisfactorio

Capítulo 4: Realizar pruebas al mercado de datos Cooperación internacional para la Sala Situacional para la Universidad de las Ciencias Informáticas

			la regla de transformación.		errores	
Rector Melchor	Válido	Rector Melchor	El sistema agrega el dato satisfactoriamente.	Rector Melchor	La prueba se realizó sin ocurrir errores	Satisfactorio

4.3.3 Aplicación de las listas de chequeo

Las listas de chequeo son un mecanismo para controlar los riesgos, su función básica consiste en detectar condiciones peligrosas que puedan acarrear problemas al software. Constituyen un documento con un conjunto de parámetros a medir de un aspecto específico, sea documentación o aplicación. Estos parámetros son básicamente una serie de preguntas referentes al atributo de calidad que se está probando y a las individualidades del documento en el caso de la documentación. A cada pregunta se le asocia una evaluación en una escala que da como resultado una medida del grado de cumplimiento y disponibilidad de lo evaluado, de esta manera se determina la evaluación del elemento probado. La lista de chequeo contiene diferentes indicadores a evaluar los cuales se encuentran distribuidos 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.
- Semántica del documento: contempla todos los indicadores a evaluar respecto a la ortografía, redacción y demás.

Las listas de chequeo se le aplicaron a los artefactos de ETL “Registro del sistema fuente (RSF)”, “Perfilado de datos (PD)”, “Diccionario de datos (DD)” y “Mapa lógico de datos (MLD)”. Seguidamente se muestra una tabla en la que se encuentran los principales aspectos que fueron evaluados en las listas de chequeo y resultados que arrojaron la aplicación de los mismos:

Tabla 13 Aplicación de las listas de chequeo a los artefactos de ETL

Secciones	RSF	PD	DD	MLD
Estructura	9	5	9	5
Indicadores	1	1	1	1
Semántica	3	3	3	3
TI	13	9	13	9

Capítulo 4: Realizar pruebas al mercado de datos Cooperación internacional para la Sala Situacional para la Universidad de las Ciencias Informáticas

IC	5	5	5	5
NC	10	6	5	7
NC-IC	1	2	2	1

TI: total de indicadores

IC: indicadores críticos

NC: no conformidades detectadas en la lista de chequeo

NC-IC: no conformidades detectadas que pertenecen a indicadores críticos

A continuación se muestra como ejemplo la lista de chequeo realizada en la segunda iteración al artefacto “Registro del sistema fuente”:

Tabla 14 Lista de chequeo realizada al artefacto “Registro del sistema fuente”

Estructura del documento					
Peso	Indicadores a evaluar	Eval	(NP)	Cantidad de elementos afectados	Comentarios
crítico	1. ¿El entregable contiene las secciones obligatorias de la plantilla estándar definida para el expediente de proyecto?	0		0	
crítico	2. ¿El alcance del proyecto describe correctamente los datos de las dimensiones y hechos del mercado de datos?	0		0	
crítico	3. ¿El objetivo expresa correctamente el propósito del documento?	0		0	
	4. ¿Se hace un uso adecuado del control del documento?	0		0	
	5. ¿En la sección de acrónimos se definen todos los acrónimos utilizados en el documento?	0		0	
	6. ¿Queda establecida en el entregable el sistema fuente de los datos?	0		0	

Capítulo 4: Realizar pruebas al mercado de datos Cooperación internacional para la Sala Situacional para la Universidad de las Ciencias Informáticas

	7. ¿Queda definido en el entregable los datos de contacto dentro de la empresa con la que se trabaja?	0		0	
	8. ¿Se especifica el sistema gestor en el que se encuentra disponible la fuente de datos?	0		0	
	9. ¿Se especifica en el entregable las propiedades de la fuente de datos?	0		0	
Indicadores definidos en el desarrollo					
Peso	Indicadores a evaluar	Eval	(NP)	Cantidad de elementos afectados	Comentarios
	1. ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?	0			
Semántica del documento					
Peso	Indicadores a evaluar	Eval	(NP)	Cantidad de elementos afectados	Comentarios
Crítico	1. ¿Se han identificado errores ortográficos en los entregables?	0		0	
Crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?	0			
	3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?	0			

4.3.4 Calidad de datos

Garantizar que los datos cargados tengan calidad es de vital importancia en el desarrollo de los MD, debido a que de esta forma se puede afirmar que ellos no contienen errores. Este proceso de validación se realiza mediante el perfilado de ellos luego de concluir con el proceso de ETL.

Capítulo 4: Realizar pruebas al mercado de datos Cooperación internacional para la Sala Situacional para la Universidad de las Ciencias Informáticas

Con el perfilado se obtiene información y estadísticas sobre los datos, lo cual permite solucionar las irregularidades que puede arrojar este proceso. Para ello se utilizó la herramienta *DataCleaner 1.5.3* en la cual se generaron reportes que sustentan con resultados positivos la calidad de lo cargado en el mercado, además se puede afirmar que la información contenida en el MD es la que proviene de la fuente y que no existen valores nulos ni vacíos. En la gráfica que se presenta a continuación se evidencia el resultado captado del perfilado de datos aplicado al hecho colaboración internacional:



Figura 26 Calidad de los datos

4.4 Evaluación de los resultados

Pruebas unitarias y de integración

Al concluir la implementación del sistema se le aplicaron pruebas unitarias a los flujos de integración de datos y a la capa de visualización, arrojando como resultado tres no conformidades en cada caso, para un total de seis, las cuales fueron resueltas satisfactoriamente. Seguidamente se muestra una gráfica con los resultados de estas pruebas (ver figura 27):

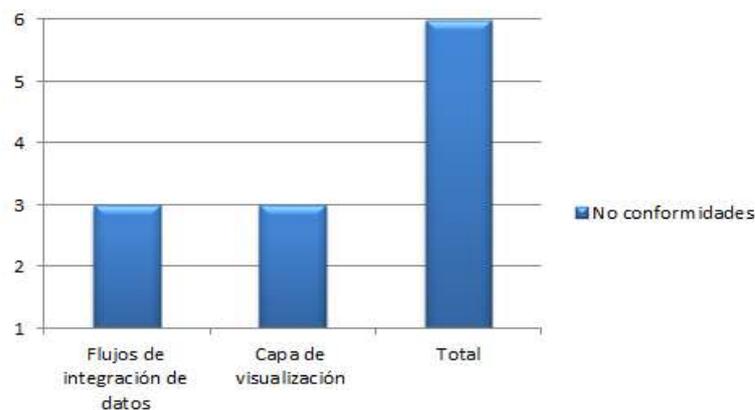


Figura 27 Comportamiento de los indicadores medidos en las pruebas unitarias

Pruebas de sistema

Las pruebas del sistema al MD Cooperación internacional se aplicaron por parte del grupo de calidad del departamento de AD, encontrándose tres no conformidades de complejidad alta, todas resueltas satisfactoriamente. Quedando el sistema liberado para su despliegue en el área.

Capítulo 4: Realizar pruebas al mercado de datos Cooperación internacional para la Sala Situacional para la Universidad de las Ciencias Informáticas

Listas de chequeo

La lista de chequeo se aplicaron en dos iteraciones, en la primera encontrando un total de 28 no conformidades, quedando resueltas satisfactoriamente. En la segunda iteración no se encontraron no conformidades. En la gráfica siguiente se refleja el comportamiento de los indicadores medidos en la lista de chequeo y las no conformidades encontradas en cada uno de ellos.



Figura 28 Comportamiento de los indicadores medidos en las lista de chequeo

Pruebas de aceptación

En conjunto con el cliente se realizaron las pruebas de aceptación de la solución, las cuales arrojaron resultados satisfactorios, quedando comprobado que el sistema cumple con sus necesidades y que están satisfechos con el producto elaborado. Ver anexos.

4.5 Conclusiones del capítulo

Después de finalizada la realización de las pruebas al MD Cooperación internacional se puede concluir que con la aplicación de las listas de chequeo, creadas con el objetivo de probar que los artefactos del proceso de ETL tuviesen la estructura correcta, los CP por reglas de transformación y por casos de uso definidos para validar los procesos de ETL y de BI y el perfilado de los datos del mercado para examinar su calidad y las pruebas de aceptación del cliente, se logró probar el MD, y de esta forma afirmar que el sistema es correcto y cumple con los requisitos del cliente.

CONCLUSIONES GENERALES

Con la realización del trabajo de diploma se logró cumplir con los objetivos planteados, desarrollándose una solución que ofrece soporte a la toma de decisiones mediante un mercado de datos para el área de Cooperación internacional de la Universidad de las Ciencias Informáticas. Por lo que se concluye que:

1. El estudio de los fundamentos teóricos de la investigación, permitió seleccionar una metodología para organizar de manera estructurada el proceso de desarrollo de software, así como las herramientas que cuentan con las características necesarias para desarrollar la solución propuesta.
2. El análisis y diseño del mercado de datos, posibilitó un mejor entendimiento del proceso de negocio de la dirección de Cooperación internacional permitiendo la implementación de una solución que responde a las necesidades del cliente.
3. La implementación del mercado de datos contribuye al proceso de apoyo a la toma de decisiones mediante el desarrollo de los diferentes subsistemas, garantizado la correcta organización, carga y visualización de los datos.
4. La validación de la solución mediante la aplicación de los casos de pruebas, el perfilado de los datos y las listas de chequeo permitió obtener un mercado funcional que cuenta con una correcta calidad de los datos y que cumple con los requisitos especificados por el cliente.

RECOMENDACIONES

Una vez concluida la implementación del mercado de datos se recomienda para futuras investigaciones lo siguiente:

- Poblar los hechos pertenecientes al área de análisis Trámites del mercado de datos con información real y desplegarlo en el área Cooperación internacional de la UCI.
- Realizar pruebas de rendimiento a la BD del MD Cooperación internacional, teniendo en cuenta que se realizan cargas anuales por lo que el número de tuplas debe aumentar considerablemente.
- Aplicar posteriormente al mercado de datos otras técnicas de BI como la minería de datos y los tableros de mando digital.

REFERENCIAS BIBLIOGRÁFICAS

1. Kimball, Ralph and Ross, Margy. The Datawarehouse Toolkit. Second Edition. s.l. : John Wiley & Sons, Inc, 2002. Vol. The Complete Guide to Dimensional Modeling.
2. Bernabeu, Ricardo Dario. HEFESTO: Metodología propia para la Construcción de un Data Warehouse. Córdoba, Argentina : s.n., 2007.
3. Giménez, Matilde Celma. Almacenes de Datos. Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia.
4. Zorrilla, Marta. Data warehouse y OLAP. Cantabria, España : s.n., 2007 - 2008.
5. HUAMANTUMBA, R. Datamart paso a paso. 2007.
6. Sistemas de Información Orientados a la Toma de Decisiones:el enfoque multidimensional. Dapena Bosquet, Isabel. Muñoz San Roque, Antonio. Sánchez Miralles, Álvaro. Madrid, España : s.n., 2005.
7. Sistema de Apoyo Gerencial Universitario. Nader, Javier.
8. [Online] www.1keydata.com/datawarehousing/molap-rolap.html.
9. Kimball, Ralph and Caserta, Joe. The Data Warehouse ETL Toolkit. Canadá : Wiley Publishing, 2004.
10. Cano, Josep Lluís. Bussines Intelligence: Competir con información . 2011.
11. Ballard, C., Abdel-Hamid A., Frankus R., Hasegawa F., Larrechart J., Leo P. y Ramos J. [Online] 2006. <http://www.redbooks.ibm.com/redbooks/pdfs/sg247210.pdf>.
12. Ecured. [Online] 2011. http://www.ecured.cu/index.php/Metodolog%C3%ADas_de_desarrollo_de_software.
13. González Hernández, Ing.Yanisbel. PROPUESTA DE METODOLOGIA PARA EL DESARROLLO DE ALMACENES DE DATOS EN DATEC. 2012.
14. Visual Paradigm. [Online] Mayo 25, 2012. <http://www.visual-paradigm.com/aboutus/newsreleases/vpuml80.jsp>.
15. Ecured. [Online] [Cited: 1 10, 2012.] http://www.ecured.cu/index.php/Sistema_Gestor_de_Base_de_Datos.

16. Postgres SQL Cuba. [Online] UCI, 2011. <http://postgresql.uci.cu>.
17. PostgreSQL. [Online] Mayo 25, 2012. <http://www.postgresql.org/es/node/655>.
18. pgAdmin. [Online] [Cited: Enero 22, 2012.] <http://www.pgadmin.org/>.
19. Levin, Jonathan. [Online] marzo 20, 2008. <http://mysqlbarbeque.blogspot.com>.
20. Pulvirenti, Adrián Sergio. Roldán, María Carina. Pentaho Data Integration. s.l. : Packt Publishing, 2011.
21. freecode. [Online] 2012. [Cited: Enero 12, 2012.] <http://ep.freecode.com/projects/datacleaner/releases/306957>.
22. Campazzo, Eduardo Nicolás and Santos, Virginia Inés. Business Intelligence: Negocios Inteligentes para Empresas Inteligentes. Universidad Nacional de La Rioja. Rene Favalaro esq. Laprida, CP5300, La Rioja, Argentina : s.n.
23. [Online] <http://www.summan.com/pentaho/pentaho-bi-platform-server>.
24. Pentaho Open Source Bussiness Intelligence. [Online] 2011. <http://www.pentaho.com>.
25. Apache Tomcat. [Online] 2010. <http://tomcat.apache.org/>.
26. slax. [Online] [Cited: Febrero 12, 2012.] <http://www.slax.org/modules.php?action=detail&id=1374>.
27. Pressman, Roger S. Ingeniería del software. Un enfoque práctico. s.l. : McGrawHill. Vol. 6ta edición. ISBN-970-10-5473-3.
28. Corporation, IBM. Ayuda de Rational Unified Process. [Online] 7.0.1, 1987-2006.
29. Dataprix. [Online] [Cited: Junio 10, 2012.] <http://www.dataprix.com/category/integracion-datos/integracion-datos/calidad-datos/limpieza-datos/perfilado-datos>.
30. Pérez, Santiago. [Online] <http://www.edutecne.utn.edu.ar/sistemas-informacion/Data%20Mining-DataWarehouse.pdf>.
31. cdn.bitbucket.org. [Online] 2012. [Cited: Febrero 12, 2012.] <http://cdn.bitbucket.org/cuattrorios/calidad-de-software/downloads/2.%20Conceptos%20de%20Calidad%20del%20Software.pdf>.
32. Sommerville, Ian. Ingeniería del Software. s.l. : Prentice Hal, 2005. ISBN: 8478290745.

BIBLIOGRAFÍA

1. Bernabeu, Ricardo Dario. HEFESTO: Metodología propia para la Construcción de un Data Warehouse. Córdoba, Argentina: s.n., 2007.
2. Campazzo, Eduardo Nicolás y Santos, Virginia Inés. Business Intelligence: Negocios Inteligentes para Empresas Inteligentes. Universidad Nacional de La Rioja. Rene Favalaro esq. Laprida,CP5300, La Rioja, Argentina : s.n
3. Cano, Josep Lluís. Bussines Intelligence: Competir con información . 2011
4. Corporation, IBM. Ayuda de Rational Unified Process. [Online] 7.0.1, 1987-2006.
5. Dapena Bosquet, Isabel. Muñoz San Roque, Antonio. Sánchez Miralles, Álvaro. Sistemas de Información Orientados a la Toma de Decisiones:el enfoque multidimensional. Madrid, España : s.n., 2005.
6. Giménez, Matilde Celma. Almacenes de Datos. Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia.
7. González Hernández, Ing.Yanisbel. PROPUESTA DE METODOLOGIA PARA EL DESARROLLO DE ALMACENES DE DATOS EN DATEC. 2012.
8. HUAMANTUMBA, R. Datamart paso a paso. 2007.
9. Inmon, W.H. Building the Datawarehouse,Fourth Edition. Indianapolis, EE.UU : Wiley Publishing, Inc., 2005.
10. Kimball, Ralph and Caserta, Joe. The Data Warehouse ETL Toolkit. Canadá : Wiley Publishing, 2004.
11. Kimball, Ralph and Ross, Margy. The Datawarehouse Toolkit. Second Edition. s.l. : John Wiley & Sons, Inc, 2002. Vol. The Complete Guide to Dimensional Modeling.
12. Nader, Javier. Sistema de Apoyo Gerencial Universitario.
13. Pressman, Roger S. Ingeniería del software. Un enfoque práctico. s.l. : McGrawHill. Vol. 6ta edición. ISBN-970-10-5473-3.
14. Pulvirenti, Adrián Sergio. Roldán, María Carina. Pentaho Data Integration. s.l. : Packt Publishing, 2011.
15. Sommerville, Ian. Ingeniería del Software. s.l. : Prentice Hal, 2005. ISBN: 8478290745.
16. Zorrilla, Marta. Data warehouse y OLAP. Cantabria, España : s.n., 2007 - 2008.
17. [Online] [Cited: Febrero 12, 2012.] <http://www.slax.org/modules.php?action=detail&id=1374>.
18. [Online] [Cited: Junio 10, 2012.] <http://www.dataprix.com/category/integracion-datos/integracion-datos/calidad-datos/limpieza-datos/perfilado-datos>.

19. [Online] [Cited: Junio 11, 2012.] <http://bi.social.uoc.edu/smc/blog/34-subsistemas-etl-de-kimball>.
20. [Online] 2012. [Cited: Enero 12, 2012.] <http://ep.freecode.com/projects/datacleaner/releases/306957>.
21. [Online] 2012. [Cited: Febrero 12, 2012.] <http://cdn.bitbucket.org/cuattrorios/calidad-de-software/downloads/2.%20Conceptos%20de%20Calidad%20del%20Software.pdf>.
22. [Online] <http://todobi.blogspot.com/2005/05/molap-rolap-holap-no-es-solo-jerga.html>.
23. [Online] <http://www.summan.com/pentaho/pentaho-bi-platform-server>.
24. Apache Tomcat. [Online] 2010. <http://tomcat.apache.org/>.
25. Ballard, C., Abdel-Hamid A., Frankus R., Hasegawa F., Larrechart J., Leo P. y Ramos J. [Online] 2006. <http://www.redbooks.ibm.com/redbooks/pdfs/sg247210.pdf>.
26. Ecured. [Online] [Cited: 1 10, 2012.] http://www.ecured.cu/index.php/Sistema_Gestor_de_Base_de_Datos.
27. Ecured. [Online] 2011. http://www.ecured.cu/index.php/Metodolog%C3%ADas_de_desarrollo_de_software.
28. Levin, Jonathan. [Online] marzo 20, 2008. <http://mysqlbarbeque.blogspot.com>.
29. Pentaho Open Source Bussiness Intelligence. [Online] 2012. <http://www.pentaho.com>.
30. Pérez, Santiago. [Online] <http://www.edutecne.utn.edu.ar/sistemas-informacion/Data%20Mining-DataWarehouse.pdf>.
31. pgAdmin. [Online] [Cited: Enero 22, 2012.] <http://www.pgadmin.org/>.
32. Postgres SQL Cuba. [Online] UCI, 2011. <http://postgresql.uci.cu>.
33. PostgreSQL. [Online] Mayo 25, 2012. <http://www.postgresql.org/es/node/655>.
34. Techno Logic Consulting S.A. [Online] [Cited: octubre 12, 2011.] <http://www.tlc-sa.com.ar/warehouses.htm>.
35. Visual Paradigm. [Online] Mayo 25, 2012. <http://www.visual-paradigm.com/aboutus/newsreleases/vpuml80.jsp>.

GLOSARIO DE TÉRMINOS

Almacén de datos: gran repositorio lógico de datos que permite el acceso y la manipulación flexible de grandes volúmenes de información procedente de diferentes fuentes.

Mercado de datos: implementación de un almacén con alcance limitado a un departamento.

BI (Inteligencia de negocio): habilidad corporativa para tomar decisiones mediante el uso de metodologías, herramientas y tecnologías que permiten aplicar en ellos técnicas analíticas de extracción de información.

ETL (Extracción, Transformación y Carga): extrae los datos desde los sistemas fuentes, los transforma y luego los carga al almacén o mercado de datos.

OLAP: Procesamiento Analítico en Línea.

RN: Regla de negocio.

CU: Caso de uso.

SGBD (Sistema Gestor de Bases de Datos): colección de programas cuyo principal objetivo es servir de interfaz entre la base de datos, el usuario y las aplicaciones.

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

Granularidad: nivel de detalle de almacenamiento de los datos, definido particularmente para el negocio.

TCP/IP: siglas de Protocolo de Control de Transmisión/Protocolo de Internet (del inglés *Transmission Control Protocol/Internet Protocol*), sistema de protocolos que posibilita diversos servicios de red.

Cubo: conjunto de dimensiones y medidas asociadas a un tema en específico.

Hecho: representa la ocurrencia de un proceso específico en el interior de la organización.

Dimensión: perspectiva mediante la cual se puede llevar a cabo el análisis sobre el hecho.

Medidas: constituye un valor o indicador de análisis del hecho.