Universidad de las Ciencias Informáticas

Facultad 3

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

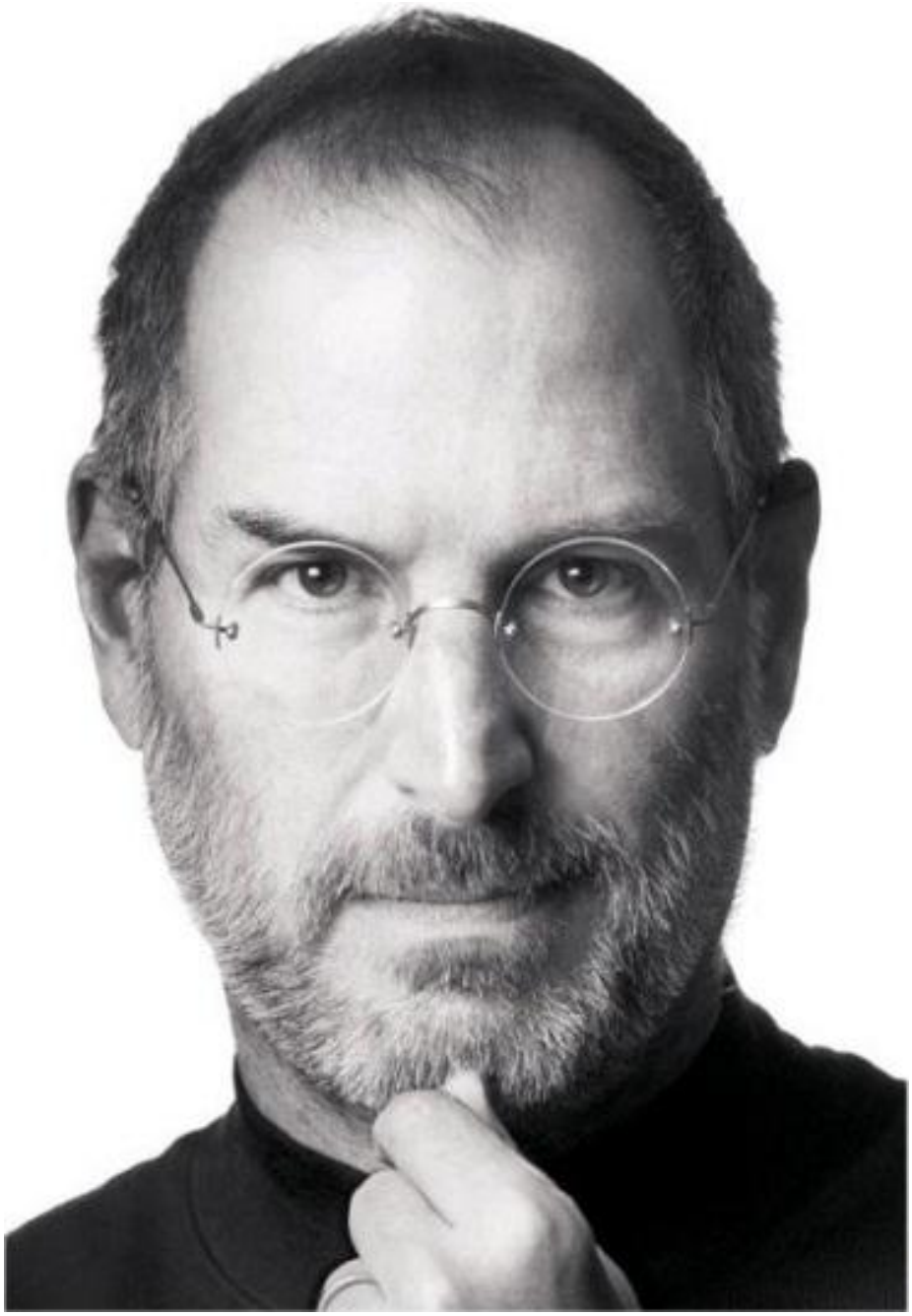
Mercado de Datos para la obtención de indicadores de Ciencia, Tecnología e Innovación del Centro de Informatización de la Gestión de Entidades.

Autor: Romel Trutié Quintana

Tutor: Ing. Yadini Pérez López

La Habana, junio de 2014

"Año 56 de la Revolución"



"No tenemos la oportunidad de hacer muchas cosas, por lo que cada cosa que hagamos debe ser excelente. Porque ésta es nuestra vida."

Steve Jobs

DECLARACIÓN DE AUTORÍA

Declaro ser autor de este trabajo y autorizo a la Universidad de las Ciencias Informáticas los derechos patrimoniales del mismo, con carácter exclusivo. Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

_____	_____
Romel Trutié Quintana	**Ing. Yadini Pérez López**
Firma del Autor	Firma del Tutor

DATOS DE CONTACTO

Autor

Romel Trutié Quintana

Correo electrónico: rtrutie@estudiantes.uci.cu

Universidad de las Ciencias Informáticas, La Habana, Cuba

Tutor

Ing. Yadini Pérez López

Correo electrónico: yadini@uci.cu

Universidad de las Ciencias Informáticas, La Habana, Cuba

AGRADECIMIENTOS

A mis padres y mi hermana, los cuales quiero mucho, por su apoyo incondicional, educación, dedicación y atención brindada a lo largo de mi vida y fundamentalmente en estos cinco años de carrera.

A mis abuelos, a mis tías, tíos, primas y primos que de una forma u otra siempre se preocupan de mí, me ayudan a continuar adelante y prepararme para la vida.

A todos mis compañeros de grupo que en algún momento tuvimos que enfrentar situaciones difíciles y salir juntos adelante.

A todos aquellas personas que conocí durante la carrera y compartimos buenos y malos momentos.

A mis amigos Dariam, Ronaldo, Yasmany, Elier, Erick, Osviel, Ornoldo, Michel, Carlos, Yendrie, Abraham, Yaniris, Lianet, Yanela, Yisel, Dailin y Neivi con quienes pasé gran parte de mi tiempo y están ahí siempre que los necesite.

A mi tutora que supo guiarme y ayudarme durante la realización del trabajo.

Al tribunal por su apoyo, guía, seguimiento y corrección del trabajo.

A todos aquellos que contribuyeron en la elaboración de este trabajo de diploma, muchas gracias.

DEDICATORIA

El presente trabajo se lo dedico a las personas más importantes en mi vida, gracias a ellos he llegado hasta aquí:

A mis queridos padres y hermana.

A mi hermano Rey Pedro a quien quiero mucho y espero que siga mis pasos.

A Nayeli Garciga Rodríguez quien me apoyó, cuidó, me dio su amor, cariño y dedicación durante la realización de este trabajo.

A Claudia e Ilianys las cuales adoro, quiero y espero se gradúen sin problemas.

RESUMEN

En un sistema de Ciencia, Tecnología e Innovación (CTI), una correcta gestión de la información es el elemento crítico para el funcionamiento adecuado. La posibilidad de acceder a información referente a los trabajos y publicaciones realizadas por cada uno de los recursos humanos (RRHH) asociados a cada departamento del Centro de Informatización de la Gestión de Entidades (CEIGE). Apreciar la evolución de los profesionales del centro, así como las proyecciones de maestrías de los mismos, se convierte en objetivo clave para el ámbito productivo – académico.

La presente investigación se basa en el análisis, diseño, implementación y validación de un mercado de datos que permita la obtención de los indicadores generados en el centro, con el fin de conocer el estado de cada una de las actividades realizadas respecto al desarrollo científico – tecnológico. Para esto se llevó a cabo un estudio sobre el almacenamiento de la información, se decidió llevar a cabo la construcción de un mercado de datos (MD), para lo cual se seleccionaron las herramientas adecuadas para el desarrollo del MD, se siguieron las fases que propone la metodología de la XETID, obteniendo de la misma, el modelo conceptual y modelo de datos del MD, entre otros artefactos, además se realizaron pruebas para verificar la efectividad de la solución.

La solución obtenida posibilitará a la dirección de investigación y postgrado del centro, realizar el análisis de los datos contenidos en el MD y apoyar a la toma de decisiones en dependencia de los mismos.

Palabras clave:

Mercado de datos, indicadores, toma de decisiones.

TABLA DE CONTENIDOS

INTRODUCCIÓN 3
CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA SOBRE MERCADO DE DATOS 7
1.1 Introducción del Capítulo 7
1.2 Superación de los RRHH 7
1.3 Almacenes de Datos 8
1.3.1 Definición 8
1.3.2 Características del AD 8
1.4 Mercados de Datos 9
1.4.1 Ventajas de los MD 10
1.4.2 Desventajas de los MD 10
1.5 Formas de modelar un MD 11
1.5.1 Modelo Multidimensional 11
1.6 Tipos de Esquemas para el modelamiento de un MD 11
1.6.1 Esquema en estrella 11
1.6.2 Esquema Copo de Nieve 12
1.6.3 Esquema Constelación 14
1.7 Procesamiento Analítico en Línea 15
1.7.1 Modos de Almacenamiento 16
1.7.2 Comparación entre sistemas OLAP 18
1.8 Sistemas para el Almacenamiento y Análisis de Información 19
1.8.1 Sistemas Internacionales 19
1.8.2 Sistemas Nacionales 22
1.8.3 Análisis de los sistemas de gestión descritos 24
1.9 Metodologías para el desarrollo de un MD 24
1.9.1 Ralph Kimball 24
1.9.2 Hefesto 25
1.9.3 Empresa de Tecnologías de la Información para la Defensa (XETID) 25
1.9.4 Selección de la metodología adecuada 26
1.10 Proceso ETL 27
1.10.1 Extracción 27
1.10.2 Transformación 28
1.10.3 Carga 28
1.11 Herramientas de desarrollo 29
1.11.1 Herramienta Case Visual Paradigm 29
1.11.2 Sistema Gestor de Base de Datos PostgreSQL 30

1.11.3 Suite Pentaho 31
1.12 Conclusiones del Capítulo 32
CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS 33
2.1 Introducción 33
2.2 Análisis 33
2.2.1 Identificación de Preguntas 33
2.2.2 Identificar Perspectivas e Indicadores 35
2.2.3 Agrupación de indicadores por perspectivas comunes. 38
2.2.4 Definición de las Tablas de Hechos y Dimensiones 39
2.3 Diseño 41
2.3.1 Diseño de los Modelos Conceptuales 41
2.3.2 Modelo de Datos del MD 45
2.4 Conclusiones del Capítulo 46
CAPÍTULO 3: IMPLEMENTACIÓN Y VALIDACIÓN DEL MERCADO DE DATOS 47
3.1 Introducción al Capítulo 47
3.2 Proceso ETL 47
3.2.1 Mapeo de los datos fuente al destino 47
3.2.2 Establecer condiciones adicionales y de restricciones 48
3.2.3 Cargas incrementales de datos 50
3.2.4 Automatización del proceso ETL 51
3.3 Validación del Mercado de Datos 52
3.3.1 Pruebas para validar la calidad del Mercado de Datos 53
3.3.2 Pruebas para validar el rendimiento del Mercado de Datos 58
3.4 Conclusiones del capítulo 58
CONCLUSIONES GENERALES 60
RECOMENDACIONES 61
REFERENCIAS BIBLIOGRÁFICAS 62

INTRODUCCIÓN

A nivel mundial, para las empresas, contar con personal altamente calificado se ha convertido en una prioridad, esto en ocasiones ha influido en el éxito competitivo entre una u otra. Las corporaciones emplean un tiempo considerable dentro de su ciclo de producción a la capacitación de sus recursos humanos (RRHH), a potenciar el desarrollo científico-tecnológico y a la generación de nuevos conocimientos, conllevando a un desarrollo productivo de la entidad. Los resultados obtenidos en esta superación de los RRHH, son expuestos en eventos de diversa índole o expresados como publicaciones.

El control de este proceso de capacitación de los RRHH trae consigo un gran cúmulo de información que en muchos casos es compleja de gestionar y analizar, principalmente si el proceso se realiza manualmente sin el empleo de alguna herramienta que permita su automatización.

En la actual sociedad del conocimiento, este ritmo creciente de información dentro de las empresas y organizaciones es inevitable. Toda información generada se ha convertido en uno de los ejes esenciales para el negocio de las entidades, tanto por su valor para conocer certeramente a sus actuales y potenciales clientes, como para gestionar sus propios procesos de manera eficiente (1).

En respuesta a la necesidad de contener estos volúmenes de datos que se generan, han surgido un grupo de teorías, metodologías, modelos y técnicas que, apoyados por la tecnología, se conoce como Inteligencia de Negocios (IN). La IN se centra en convertir los datos en información relevante, haciéndola accesible de manera ágil y sencilla a todos los empleados de la organización (2).

Los productos de software que responden a esta necesidad de conservar la información generada por las empresas para su posterior análisis, se han caracterizado por poseer los mismos principios en cuanto a arquitectura. De manera general, estas soluciones permiten reunir, depurar y transformar los datos del quehacer diario del negocio, en información estructurada para su posterior explotación (3).

Una de las soluciones más comunes empleadas en la actualidad son los almacenes de datos (AD) y los MD. Los MD son sistemas que permiten la recogida de datos de un entorno

transaccional de la compañía, los filtran y procesan para su almacenamiento, proporcionando una plataforma sólida de datos para su posterior análisis. Mientras que un AD es la unión de varios MD (4).

Nuestro país no se encuentra ajeno a dicho progreso, en la Universidad de las Ciencias Informáticas (UCI), como parte del control de las áreas estratégicas establecidas por el Ministerio de Educación Superior (MES), se encuentran establecidas las estrategias de seguimiento a la capacitación de los profesionales.

Estas estrategias de seguimiento a la superación de los RRHH va desde tener una notable actividad científica expresada en los resultados alcanzados en su desempeño profesional, la participación en cursos de postgrados, la obtención de resultados científicos y que los mismos hayan sido publicados en revistas referenciadas u otros medios, hasta la planificación por semestre de cada una de estas actividades científicas-investigativas a realizar en el año (5).

La facultad tres tiene establecido entre sus objetivos: lograr un incremento en la actividad científica-investigativa, en función de repercutir en la elevación de la calidad de los procesos sustantivos de la misma y en particular en la obtención de nuevos productos con alto valor agregado como parte de la estrategia de superación de los RRHH (6).

Todo este proceso de superación se refleja en cada uno de los departamentos del centro de producción suscritos a la facultad, dándole cumplimiento a través de eventos y publicaciones realizadas por cada uno de los profesionales del centro. Actualmente este proceso de control de los parámetros de Ciencia, Tecnología e Innovación (CTI), del Centro para la Informatización de la Gestión de Entidades (CEIGE), es realizado por el Jefe de Investigación y Postgrado de la facultad. Este proceso se torna engorroso, debido a que para generar la información que debe archivar y entregar dicho directivo, actualmente, se requiere más tiempo del disponible y además, el proceso se dificulta debido a que se presentan los siguientes problemas:

- El proceso de control CTI es realizado de forma manual, por lo que durante el mismo se cometen disímiles errores y, en ocasiones, la información se entrega de forma tardía.
- La información referente a estos eventos de CTI llevados a cabo por cada uno de los profesionales, se encuentra dispersa en el centro, por lo que es difícil localizar la información necesaria para elaborar los informes pertinentes.

- El empleo de herramientas ofimáticas para la elaboración de informes, filtrado y posterior análisis de la información, no permite llevar un histórico de los datos generados en los diferentes cursos, además las posibilidades de análisis se ven limitadas.
- La carencia de una herramienta capaz de resumir y estandarizar los datos referentes a CTI del centro, dificulta el trabajo de los directivos.
- El constante crecimiento de la información referente a la superación de los profesionales en CEIGE, trae consigo que el manejo de dicha información sea cada vez más complejo.

Debido a todos los problemas planteados anteriormente se dificulta la correcta revisión de los objetivos CTI trazados en el centro.

Por lo antes expuesto, se plantea el siguiente **problema a resolver**: ¿Cómo agilizar la obtención de los indicadores de CTI del CEIGE?

Definiéndose como **objeto de estudio** el análisis y almacenamiento de la información de indicadores de CTI y como **campo de acción** los mercados de datos para el análisis y almacenamiento de la información de indicadores CTI.

Para dar solución al problema se plantea entonces como **objetivo general** de esta investigación: Agilizar el proceso de obtención de indicadores CTI mediante el desarrollo de un MD, permitiendo comprobar el progreso de estas actividades realizadas por el personal del CEIGE.

Para darle cumplimiento al objetivo general se plantearon los siguientes **objetivos específicos**:

1. Fundamentar la investigación mediante la elaboración del Marco Teórico para sustentar los conceptos y la propuesta de desarrollo del MD.
2. Realizar el análisis, diseño e implementación del mercado de datos para mejorar el análisis de la información.
3. Realizar pruebas de funcionamiento y rendimiento para validar el MD.

Finalmente se define como **idea a defender:** El desarrollo de un mercado de datos para el análisis de los indicadores de CTI de CEIGE posibilitará agilizar la obtención de dichos indicadores.

Entre los **métodos científicos** utilizados destacan:

Como **métodos teóricos**: el **Analítico-Sintético** que se utilizó en la revisión de documentos, libros, artículos e informes para la extracción de elementos importantes que están relacionados con la investigación, con el objetivo de lograr una mayor visión sobre el tema, permitiendo tener un enfoque global del mismo. Se aplicó además el método **Inductivo–Deductivo** que se utilizó en la profundización de los diferentes aspectos relacionados con el tema de investigación; el método **Histórico-Lógico** que permitió conocer cómo se realiza actualmente el proceso CTI en el CEIGE, con el fin de capturar los requisitos de dicho proceso a tener en cuenta en la solución; y el método de **Modelación**, para realizar el diseño del modelo conceptual, modelo lógico y modelo de datos del MD.

Se emplearon **métodos empíricos** como: **Entrevistas estructuradas** y **Observaciones** para la recopilación de información sobre las particularidades de los diferentes procesos del negocio a tener en cuenta en la solución.

El presente trabajo consta de 3 capítulos estructurados de la siguiente manera:
Capítulo 1: Fundamentos teóricos sobre el desarrollo de un mercado de datos. Se abordan los principales conceptos relacionados con los MD, así como una descripción de todas las herramientas, tecnologías y la metodología a utilizar para dar solución al problema planteado.

Capítulo 2: Análisis y diseño del Mercado de Datos. Se describen los pasos de la metodología utilizada con el objetivo de facilitar la construcción del MD. En principio se identifican las necesidades de información del cliente para el diseño del modelo conceptual del MD y se concluye con la obtención del modelo de datos del MD.

Capítulo 3: Implementación y validación del Mercado de Datos. Se describe el proceso de implementación del MD, abordando los procesos de Extracción, Transformación y Carga (ETL) mediante la utilización de las herramientas descritas en el capítulo primero. También se realiza la validación del MD, a partir de la construcción de cubos de Procesamiento Analítico en Línea, que muestran el correcto funcionamiento de la solución y se realizan pruebas de rendimiento.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA SOBRE MERCADO DE DATOS

1.1 Introducción del Capítulo

Durante este capítulo, se realiza un estudio del estado del arte y la fundamentación teórica de los MD relacionados fundamentalmente con CTI. Además se refleja la importancia de la superación de los RRHH como elemento clave dentro de la organización. Las diferentes metodologías y herramientas empleadas en el desarrollo del MD, así como las formas de almacenarla y procesarla.

1.2 Superación de los RRHH

En la IN hay tres aspectos de fundamental importancia para el desarrollo y éxito de cualquier organización.

- La gestión de la información, la cual incluye el proceso de planificar, organizar, controlar y dirigir todos los medios informativos dentro de la organización. Mediante el mismo se garantiza un constante flujo de la información y que la misma llegue a cada uno de los miembros de la entidad.
- La gestión del conocimiento, la cual tributa al desarrollo del conocimiento del personal y un mejor empleo del mismo.
- La toma de decisiones, que permite encaminar a la empresa por el mejor camino posible y lograr alcanzar los objetivos trazados.

Actualmente muchas entidades se han percatado que la aplicación de la IN influye en el éxito competitivo entre organizaciones, pero existe otro factor fundamental y que juega un rol principal dentro de la empresa: los RRHH. Donde, cada día son más los recursos y el tiempo empleado por la entidad para incrementar la calidad, el conocimiento y la eficiencia laboral de los RRHH como parte de su superación profesional.

Por tanto, el propósito de la organización debe estar enfocado a combinar o crear una red de relaciones entre sus RRHH y que los mismos realicen actividades correlacionadas con la entidad.

Por lo antes expuesto, se concluye que mediante una correcta preparación, superación y organización del personal se puede alcanzar el éxito dentro una organización.

1.3 Almacenes de Datos

1.3.1 Definición

Los AD son sistemas con el fin de recoger datos de los distintos entornos transaccionales de la compañía, filtrarlos y procesarlos para almacenarlos, proporcionando de esta forma una plataforma sólida de datos consolidados e históricos para su posterior análisis. Además son herramientas que soportan el proceso de toma de decisiones (7).

Definición de Bill Inmon:

Un AD se define como "una colección de datos orientada a temas, integrada, variante en el tiempo y no volátil, usada principalmente para la toma de decisiones" (8).

Definición de Ralph Kimball:

Un AD es "una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis". También Kimball definió que un AD no es más que: "la unión de todos los MD de una entidad" (4).

1.3.2 Características del AD

Los AD presentan características que los identifican tales como:

- **Integrados**: Se construye mediante la integración de fuentes de datos múltiples, y heterogéneas. Se aplican técnicas de limpieza e integración.
- **Orientados a temas**: Está organizado entorno a las materias de negocio: cliente, producto, ventas. El modelado y el análisis de los datos se enfocan de cara a la toma de decisiones, no para las operaciones del día a día o en el procesamiento de transacciones. Proporciona una vista simple y concisa de las materias del negocio excluyendo aquellos datos que no se necesitan para el proceso de toma de decisiones.
- **Variables en el tiempo**: Cada clave del almacén contiene una referencia a la fecha explícita o implícitamente y el horizonte de tiempo para el almacén es significativamente más largo que el de los sistemas operacionales.
- **No volátil**: Es un almacén de datos físicamente separado del entorno operacional. Las actualizaciones de la BD operacional no ocurren en el entorno del almacén. No se requieren mecanismos de control de la concurrencia y recuperación. Se requieren dos operaciones nada más: carga de los datos y acceso a datos (8).

Se puede concluir que un AD es un sistema capaz de almacenar información el cual está

compuesto por uno o varios MD.

1.4 Mercados de Datos

Un mercado de datos es un subconjunto de datos de un almacén, relativo a los requisitos de un departamento o área de negocio concreto. Este subconjunto de datos puede funcionar de forma autónoma, o bien enlazado al Almacén de Datos (9).

La idea de lo que es un AD o un MD difiere mayormente en la dimensión de la información que contengan. Ambos pueden compartir el mismo proceso de construcción, la arquitectura, las funciones que realizan. Por esta razón, a partir de este punto del documento, se hablará indistintamente de AD o MD (10).

Enfoques para el desarrollo de un AD

Kimball propone para la creación de los AD una arquitectura ascendente (bottom-up[1]), o sea, el plantea que se debe crear un MD para cada departamento o área de negocio, y el AD sería la unión de los MD resultantes (4).

Mientras que Inmon propone un enfoque descendente (top-down[2]), en el cuál plantea que primero se construye el AD y a partir del mismo se extraen los MD (8).

Los MD presentan las siguientes **características**:

- El diseño del MD se realiza siguiendo una estructura consistente según las necesidades de los usuarios.
- Contiene solo el mínimo de información histórica.
- Tiene el grado de granularidad[3] según la profundidad del análisis necesario.
- Supone costes adicionales en hardware, software y accesos de red.
- Se centran en los requisitos de los usuarios asociados a un departamento o área de negocio concretos.
- Son sencillos a la hora de utilizarlos y comprender sus datos (4).

[1] **Buttom-up:** Estrategia de procesamiento de información donde las partes individuales se diseñan con detalle y luego se enlazan para formar componentes más grandes (10).

[2] **Top-down:** Estrategia de procesamiento de información donde se formula un resumen del sistema, sin especificar detalles. Luego cada parte del sistema se refina diseñando con mayor detalle (10).

[3] **Granularidad:** Representa el nivel de detalle al que se desea almacenar la información sobre el negocio que se esté analizando. Mientras mayor sea el nivel de detalle de los datos, se tendrán mayores posibilidades de análisis (10).

En resumen, un MD es creado con el objetivo de brindar información relevante a la entidad y dar soporte al proceso de toma de decisiones, y de esta forma satisfacer las necesidades del cliente. Mientras que un AD sería la unión de varios MD.

1.4.1 Ventajas de los MD

Los MD aportan facilidades e inmediatez en el manejo de la información. Las ventajas más significativas son:

- Menor cantidad de datos, lo cual implica que procesa más rápido la carga de datos así como las consultas.
- La aplicación cliente que pide la consulta es independiente del servidor que la procesa y del servidor de bases de datos que almacenan la información.
- Los costos que implica la construcción de un MD son mucho menores a los de la implementación de un AD.
- El enorme retorno de la inversión es evidencia de la gran ventaja competitiva que acompaña a esta tecnología. La ventaja competitiva se consigue por permitir a los responsables de la toma de decisiones tácticas y estratégicas el acceso a datos que pueden mostrar información que antes era no disponible, desconocida o sin explotar, en clientes, tendencias y demandas.
- Incremento de la productividad de los responsables de la toma decisiones.
- Permite analizar los datos del negocio desde la perspectiva de su evolución en el tiempo, predecir tendencias de evolución del negocio, identificar nuevas oportunidades de negocio y tomar decisiones estratégicas, además de reducir los costos materiales y humanos en la toma de decisiones (11).

1.4.2 Desventajas de los MD

Las desventajas más sobresalientes asociadas a los MD son las siguientes:

- No permite el manejo de grandes volúmenes de información por lo que muchas veces se debe recurrir a un conjunto de MD para cubrir todas las necesidades de información de la empresa.
- Baja estimación de los recursos necesarios para la carga de datos.
- Problemas ocultos en los sistemas fuente.
- Alto mantenimiento.
- Proyectos de larga duración.

- Complejidad de integración.

1.5 Formas de modelar un MD

1.5.1 Modelo Multidimensional

El objetivo del análisis multidimensional es ganar comprensión en el conocimiento contenido en las bases de datos. Su principal ventaja es que facilitan los análisis complejos y la visualización de los datos en el MD para procesos de toma de decisiones, reduciendo la confusión y disminuyendo las interpretaciones erróneas (12).

Además, ya que los datos están almacenados físicamente en una estructura multidimensional o base de datos n-dimensional, la velocidad de estas operaciones es varias veces superior y más consistente de lo que es posible en otras estructuras de bases de datos. La combinación de simplicidad y velocidad es uno de los principales beneficios del análisis multidimensional (12).

El modelo está basado en la noción de dimensión que permite especificar diferentes maneras de estudiar la información, de acuerdo con las perspectivas del negocio bajo las cuales el análisis puede ser realizado. Cada dimensión se organiza en una jerarquía de niveles, correspondiendo a dominios de datos en diferentes niveles o granularidades. Un esquema multidimensional consiste en un conjunto de tablas de hechos que se definen respecto a combinaciones particulares de niveles (12).

Una instancia multidimensional asocia medidas, que corresponden a los datos a ser estudiados, con coordenadas simbólicas a las tablas de hechos. Finalmente, en una dimensión, los valores con un gran nivel de detalle pueden hacer roll-up (agruparse) a valores más generales (12).

1.6 Tipos de Esquemas para el modelamiento de un MD

1.6.1 Esquema en estrella

El esquema en estrella, consta de una tabla de hechos[4] central y de varias tablas de dimensiones[5] relacionadas con ésta, a través de sus respectivas claves. En la siguiente figura se puede apreciar un esquema en estrella estándar:

[4] Las tablas de hechos es donde las mediciones numéricas del negocio son almacenadas.
[5] Las tablas dimensionales son aquellas donde las descripciones textuales de las dimensiones del negocio son almacenadas.

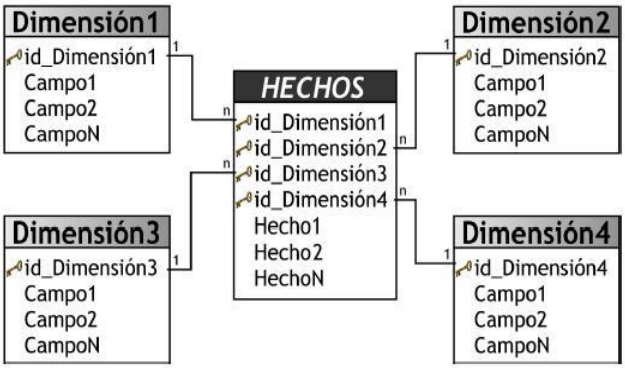


Figura 1: Esquema en estrella.

El esquema en estrella es el más simple de interpretar y optimiza los tiempos de respuesta ante las consultas de los usuarios. Este modelo es soportado por casi todas las herramientas de consulta y análisis, y los metadatos son fáciles de documentar y mantener, sin embargo es el menos robusto para la carga y es el más lento de construir (15).

A continuación se destacarán algunas características de este modelo, que ayudarán a comprender mejor el porqué de sus ventajas:

- Posee los mejores tiempos de respuesta.
- Su diseño es fácilmente modificable.
- Existe paralelismo entre su diseño y la forma en que los usuarios visualizan y manipulan los datos.
- Simplifica el análisis.
- Facilita la interacción con herramientas de consulta y análisis.

1.6.2 Esquema Copo de Nieve

Este esquema representa una extensión del modelo en estrella cuando las tablas de dimensiones se organizan en jerarquías de dimensiones el cual se puede apreciar en la siguiente figura:

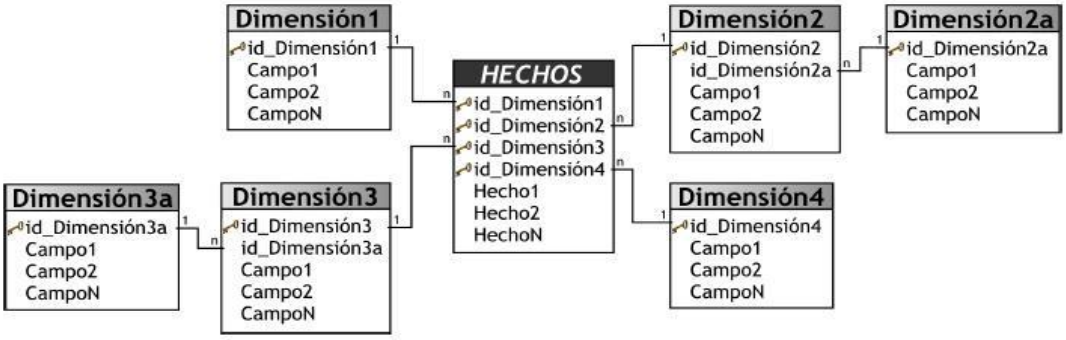


Figura 2: Esquema Copo de Nieve.

Como se puede apreciar en la figura 2, existe una tabla de hechos central que está relacionada con una o más tablas de dimensiones, quienes a su vez pueden estar relacionadas o no, con una o más tablas de dimensiones (15).

Este modelo es más cercano a un modelo de entidad relación, que al modelo en estrella, debido a que sus tablas de dimensiones están normalizadas. Una de los motivos principales de utilizar este tipo de modelo, es la posibilidad de segregar los datos de las tablas de dimensiones y proveer un esquema que sustente los requerimientos de diseño. Otra razón es que es muy flexible y puede implementarse después de que se haya desarrollado un esquema en estrella.

Se pueden definir las siguientes características de este tipo de modelo:

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

A pesar de todas las características y ventajas que trae aparejada la implementación del esquema copo de nieve, existen dos grandes inconvenientes de ello:

- Si se poseen múltiples tablas de dimensiones, cada una de ellas con varias jerarquías, se creará un número de tablas bastante considerable, que pueden llegar al punto de ser inmanejables.

- Al existir muchas uniones y relaciones entre tablas, el desempeño puede verse reducido (15).

1.6.3 Esquema Constelación

Este modelo está compuesto por una serie de esquemas en estrella, y tal como se puede apreciar en la siguiente figura, está formado por una tabla de hechos principal (“HECHOS_A”) y por una o más tablas de hechos auxiliares (“HECHOS_B”), las cuales pueden ser sumariadas de la principal. Dichas tablas yacen en el centro del modelo y están relacionadas con sus respectivas tablas de dimensiones.

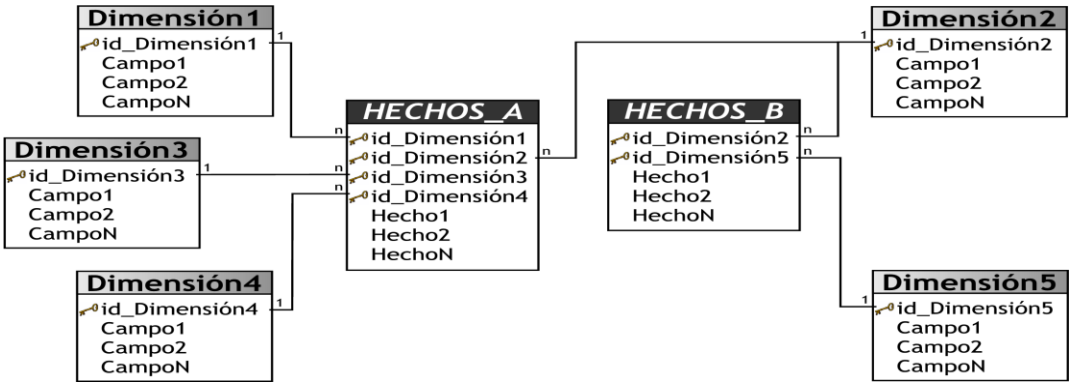


Figura 3: Esquema Constelación.

No es necesario que las diferentes tablas de hechos compartan las mismas tablas de dimensiones, ya que, las tablas de hechos auxiliares pueden vincularse con solo algunas de las tablas de dimensiones asignadas a la tabla de hechos principal, y también pueden hacerlo con nuevas tablas de dimensiones (15).

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

- Permite tener más de una tabla de hechos, por lo cual se podrán analizar más aspectos claves del negocio con un mínimo esfuerzo adicional de diseño.
- Contribuye a la reutilización de las tablas de dimensiones, ya que una misma tabla de dimensión puede utilizarse para varias tablas de hechos.
- No es soportado por todas las herramientas de consulta y análisis.

Para la realización del trabajo se empleará el tipo de esquema constelación de hechos,

teniendo en cuenta el análisis del negocio y las necesidades primarias expuestas por el cliente en los primeros encuentros.

Otro aspecto de gran importancia durante el desarrollo de un MD, es la forma en que se va a realizar el procesamiento de la información. El Procesamiento Analítico en Línea (OLAP[6]) es una de estas variantes, la cual se profundizará a continuación.

1.7 Procesamiento Analítico en Línea

La tecnología OLAP facilita el análisis de datos en línea en un MD, proporcionando respuestas rápidas a consultas analíticas complejas. OLAP es utilizado generalmente para ayuda en la toma de decisiones y presenta los datos a los usuarios a través de un modelo de datos intuitivo y natural. Con este estilo de presentación los usuarios finales pueden ver y entender con mayor facilidad la información de sus BD, lo que permite a las organizaciones reconocer el valor de sus datos. Generalmente los esquemas de las BD tienen cierta complejidad para el usuario final, debido a ello la concepción de las consultas puede ser una tarea ardua (16).

Principales **características** de OLAP:

- Rápido: proporciona información al usuario a una velocidad constante (principalmente en cinco segundos o menos).
- Análisis: realiza análisis estadísticos y numéricos básicos de los datos.
- Compartida: permite compartir los datos potencialmente confidenciales a través de una gran cantidad de usuarios, implementando, para esto, los requerimientos de seguridad necesarios.
- Multidimensional: permite ver la información en determinadas vistas o dimensiones.
- Información: accede a todos los datos necesarios, donde quiera que éstos residan, y mientras no esté limitada por el volumen (17).

Cubo OLAP

Un cubo OLAP se utiliza para el procesamiento analítico en línea, y es una estructura de datos que proporciona un análisis rápido de la información. Se puede considerar como una ayuda para manipular y analizar datos desde varias perspectivas. La estructura de datos del cubo puede ayudar a superar algunas limitaciones de las bases de datos relacionales (18).

[6] **OLAP (Procesamiento Analítico en Línea):** Es una solución utilizada en el campo de la IN, la cual consiste en consultas a estructuras multidimensionales que contienen datos resumidos de grandes Bases de Datos o Sistemas Transaccionales.

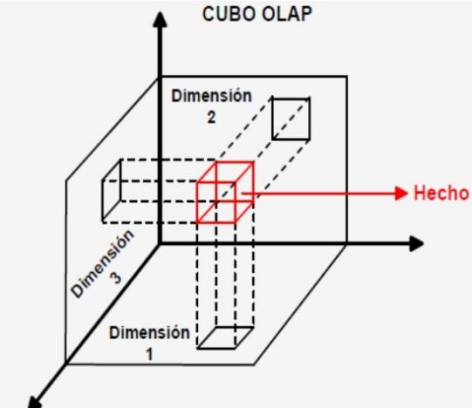


Figura 4: Estructura de un Cubo OLAP.

Dentro de este tipo de procesamiento existen tres modos diferentes de almacenar los datos y a partir de los cuales se consulta la información:

- Procesamiento Analítico en Línea Relacional (ROLAP).
- Procesamiento Analítico en Línea Multidimensional (MOLAP).
- Procesamiento Analítico en Línea Híbrido (HOLAP).

1.7.1 Modos de Almacenamiento

Los diferentes tipos de implementación o formas de almacenamiento de los datos a tener en cuenta para la construcción de un MD son los siguientes:

1.7.1.1 ROLAP

En ROLAP se utiliza una arquitectura de tres niveles. La BD relacional maneja el almacenamiento de datos, el motor OLAP proporciona la funcionalidad analítica, y alguna herramienta especializada es empleada para el nivel de presentación. El nivel de aplicación es el motor OLAP, que ejecuta las consultas de los usuarios. El motor OLAP se integra con el nivel de presentación a través del cual los usuarios realizan los análisis OLAP. Después de que el modelo de datos para el MD se ha definido, los datos se cargan desde los sistemas transaccionales.

Figura 5: Tipo de procesamiento ROLAP.

Los usuarios finales ejecutan sus análisis multidimensionales, a través del motor OLAP, el cual transforma sus datos a consultas en SQL ejecutadas en las BD relacionales y sus resultados son devueltos a los usuarios.

La arquitectura ROLAP es capaz de usar datos precalculados (si éstos están disponibles), o de generar dinámicamente los resultados desde la información elemental (menos resumida). Esta arquitectura accede directamente a los datos del MD y soporta técnicas de optimización para acelerar las consultas como tablas particionadas, soporte a la desnormalización, soporte de múltiples reuniones, pre calculado de datos e índices (16).

1.7.1.2 MOLAP

Un sistema MOLAP usa una BD multidimensional (BDMD), en la que la información se almacena multidimensionalmente. Utiliza una arquitectura de dos niveles: la BDMD y el motor analítico. La BDMD es la encargada del manejo, acceso y obtención de los datos. El nivel de aplicación es el responsable de la ejecución de las consultas OLAP. El nivel de presentación se integra con el de aplicación y proporciona una interfaz a través de la cual los usuarios finales visualizan los análisis OLAP.

Figura 6: Tipo de procesamiento MOLAP.

La información procedente de los sistemas transaccionales se carga en el sistema MOLAP. Una vez cargados los datos en la BDMD, se realiza una serie de cálculos para obtener datos agregados a través de las dimensiones del negocio, poblando la estructura de la BDMD. Luego de llenar esta estructura, se generan índices y se emplean algoritmos de tablas resumen[7] para mejorar los tiempos de accesos de las consultas. Una vez que el proceso de poblado ha finalizado, la BDMD está lista para su uso. Los usuarios solicitan informes a través de la interfaz y la lógica de aplicación de la BDMD obtiene los datos (16).

1.7.1.3 HOLAP

Se han desarrollado soluciones de OLAP híbridas que combinan el uso de las arquitecturas ROLAP y MOLAP. En una solución con HOLAP, los registros detallados (los volúmenes más grandes) se mantienen en la BD relacional, mientras que los agregados lo hacen en un almacén MOLAP independiente (19).

Figura 7: Tipo de procesamiento HOLAP.

1.7.2 Comparación entre sistemas OLAP

Cada sistema OLAP tiene ciertos beneficios, algunas implementaciones MOLAP son propensas a la ruptura de la base de datos; este fenómeno provoca la necesidad de grandes cantidades

[7] Tabla resumen es una estructura de datos que asocia llaves o claves con valores.

de espacio de almacenamiento para el uso de una base de datos MOLAP cuando se dan ciertas condiciones: elevado número de dimensiones, resultados pre calculados y escasos datos multidimensionales.

Por lo general MOLAP ofrece mejor rendimiento debido a la especializada indexación y a las optimizaciones de almacenamiento. MOLAP también necesita menos espacio de almacenamiento en comparación con los especializados ROLAP porque su almacenamiento especializado normalmente incluye técnicas de compresión.

ROLAP es generalmente más escalable, sin embargo, el gran volumen de pre procesamiento es difícil de implementar eficientemente, por lo que con frecuencia se omite; el rendimiento de una consulta ROLAP puede afectar el tiempo de respuesta del MD. Desde la aparición de ROLAP van apareciendo nuevas versiones de bases de datos preparadas para realizar cálculos, las funciones especializadas que se pueden utilizar tienen más limitaciones.

HOLAP engloba un conjunto de técnicas que tratan de combinar MOLAP y ROLAP de la mejor forma posible. Generalmente puede pre-procesar rápidamente, escala bien, y proporciona una buena función de apoyo.

Definición del modo de almacenamiento a utilizar

Teniendo en cuenta la necesidad del cliente de visualizar los datos almacenados en el sistema, se utilizará tanto para el modo de almacenamiento como para el procesamiento de la información ROLAP. Permitiendo acelerar las consultas realizadas al MD y devolver los resultados en el menor tiempo posible.

1.8 Sistemas para el Almacenamiento y Análisis de Información

1.8.1 Sistemas Internacionales

1.8.1.1 Sistema Servidor de Análisis SyBase IQ

Sybase IQ es un servidor analítico privativo diseñado específicamente para análisis avanzado, almacenamiento de datos y entornos de inteligencia comercial. Al tener la capacidad para trabajar con volúmenes masivos de datos estructurados y no estructurados, es perfecto para grandes volúmenes de datos.

Se distingue de las bases de datos convencionales por su arquitectura orientada en columnas y basada en cuadrículas, su compresión de datos patentada y su optimizador avanzado de

consultas, y se puede implementar en una amplia variedad de parámetros para proporcionar un rendimiento, una flexibilidad y una economía incomparables aún en los entornos de informes y análisis más desafiantes (20).

Ventajas del Sistema:

- Brinda resultados más rápidos para aceleración de informes, obtención de datos y análisis predictivo.
- Ofrece la mejor relación de rendimiento/precio de la industria.
- Brinda una arquitectura basada en columnas abierta y flexible.

Diseñado desde el comienzo para el análisis, Sybase IQ ofrece un método único en comparación con un sistema OLTP. Es posible lograr mejoras en el rendimiento de tipo analítico de los sistemas transaccionales mediante técnicas de ajuste y optimización, actualizaciones de tecnología o hardware adicional. Sin embargo, estos métodos son costosos y presentan limitaciones. Las consultas pueden demorar horas o días para encontrar información de utilidad.

El servidor analítico de Sybase IQ proporciona una disminución de requisitos de disco y CPU en comparación con los sistemas tradicionales de administración de bases de datos basados en filas, los cuales deben ser adaptados para admitir el análisis y almacenamiento de grandes volúmenes de datos. Constituye una alternativa sorprendentemente sencilla y económica que garantiza que las empresas saquen el mayor provecho de sus datos a diario (20).

1.8.1.2 Sistema Integrado de Información sobre Investigación Científica y Tecnológica de México (SIICYT)

Surge como un instrumento de planeación en materia de Ciencia y Tecnología, para impulsar de una manera eficiente las actividades científicas y tecnológicas del país. En este programa se definen tres objetivos estratégicos:

- Contar con una política de Estado en Ciencia y Tecnología.
- Incrementar la capacidad científica y tecnológica del país.
- Elevar la competitividad y la innovación de las empresas.

El SIICYT es una herramienta que refuerza la integración y solidez del Sistema Nacional de Ciencia y Tecnología. Integrará los esfuerzos de diferentes instituciones educativas, centros de investigación, organismos públicos, empresas y personas físicas y morales del sector público y

privado. A fin de promover el desarrollo y la vinculación de la ciencia básica y la innovación tecnológica, así como convertir a la ciencia y la tecnología en un elemento fundamental de la cultura general de la sociedad.

Desde su implantación se ha notado gran incremento en cuanto a respuesta con respectos al resto de los sistemas empleados en el país, haciendo del mismo un sistema insigne para el país y su economía (21).

1.8.1.3 System Center 2012 - Service Manager

El almacenamiento de datos en System Center 2012 – Service Manager proporciona tres funciones principales:

- Descarga los datos desde la base de datos principal de Service Manager para mejorar el rendimiento de la base de datos de Service Manager.
- Almacena datos a largo plazo.
- Proporciona datos para informes.

El almacenamiento de datos que se incluye con Service Manager es en realidad su propio grupo de administración. Estos componentes están integrados en una plataforma común, que a su vez consta de lo siguiente:

- Una base de datos basada en modelos, para almacenar la información de configuración relacionada con el almacenamiento de datos y para almacenar provisionalmente los datos una vez extraídos de la base de datos de Service Manager.
- El servidor de administración, que consta de lo siguiente:
  - Servicio de acceso a datos.
  - Servicio de administración.
  - Servicio de configuración de administración.

El almacenamiento de datos se diseñó para:

- Ser totalmente extensible mediante módulos de administración.
- Utilizar los procedimientos recomendados de almacenamiento de datos, tales como el modelado dimensional con hechos[8] y dimensiones[9].
- Operar a gran escala.

[8] Representa un evento o actividad específica del negocio. Es aquel dato que se obtiene de la intersección de las dimensiones y siempre debe estar relacionado con al menos una dimensión.
[9] Son las características de un hecho que permite su análisis posterior, en el proceso de toma de decisiones.

El almacenamiento de datos se ha diseñado y generado como componente de una plataforma que permite a los usuarios colocar datos de todos los productos con el fin de obtener una visión global de sus inversiones en tecnologías de la información (TI) (18).

1.8.1.4 Sistema Nacional de Ciencia, Tecnología e Innovación de Colombia (SNCTI)

Colciencias es el Departamento Administrativo de Ciencia, Tecnología e Innovación. El cual promueve las políticas públicas para fomentar la CT+I en Colombia. Las actividades alrededor del cumplimiento de su misión implican concertar políticas de fomento a la producción de conocimientos, construir capacidades para CT+I, y propiciar la circulación y usos de los mismos para el desarrollo integral del país y el bienestar de los colombianos.

El SNCTI es un sistema en el cual forman parte las políticas, estrategias, programas, metodologías y mecanismos para la gestión, promoción, financiación, protección y divulgación de la investigación científica y la innovación tecnológica, así como las organizaciones públicas, privadas o mixtas que realicen o promuevan el desarrollo de actividades científicas tecnológicas y de innovación (22).

Estrategias de la Política Nacional de SNCTI:

- Crear condiciones donde el conocimiento sea el instrumento del desarrollo.
- Consolidación institucional SNCTI.
- Fomento a la apropiación social CTI.
- Dimensiones regional e internacional.
- Apoyo a la formación avanzada de investigadores.
- Consolidación de capacidades para CTI.
- Acelerar el crecimiento económico (23).

1.8.2 Sistemas Nacionales

1.8.2.1 Sistema de Información Docente del Ministerio de Educación Superior (SIGENU)

El proyecto SIGENU surge en Junio de 2004 a solicitud de la dirección del Ministerio de Educación Superior de Cuba (MES), como requisito a las necesidades de automatización de los procesos fundamentales de la gestión académica de una Institución de Educación Superior (IES) en todas sus modalidades de estudio. El sistema está compuesto por varios módulos que gestionan la información de un estudiante desde que se matrícula hasta que se gradúa o causa baja definitiva.

Características Técnicas

Consta de dos tipos de aplicaciones, un sistema transaccional y un sistema para la toma de decisiones. Entre las funcionalidades principales del sistema de gestión se encuentra la inscripción de un estudiante, registro de asignaturas a cursar, registro de evaluaciones, control de bajas y emisión de reportes oficiales. Para la toma de decisiones tiene como centro de la arquitectura la tecnología AD y es explotado utilizando la tecnología OLAP. Éste soporta el almacenamiento de información agregada y nominalizada de cualquier estudiante que pertenezca a una Institución de Educación Superior de Cuba, aunque no utilice como sistema de gestión académica el proyecto SIGENU.

Entorno de Desarrollo

El proyecto se desarrolla con herramientas de software libre con el paradigma de modelado MDA[10] sobre la plataforma J2EE[11]. El producto se encuentra en explotación sobre el Gestor de Base de Datos PostgreSQL, aunque su diseño e implementación es independiente del Gestor de bases de Datos y del Sistema Operativo.

Generalización de su aplicación

Comenzó su implantación hace 5 años en tres centros, desde hace 3 años se encuentran en explotación en las 17 Instituciones de Educación Superior que pertenecen al MES y en sus 169 sedes universitarias municipales en las cuales se ha notado mejoría en la realización de los procesos docentes académicos que aborda el sistema. En el curso 2008-2009 fue implantando en el Instituto Superior de Relaciones Internacionales (ISRI) y el Instituto Superior de Diseño Industrial (ISDI). El volumen de datos que gestiona el sistema en estos momentos es aproximadamente 400 000 estudiantes (24).

1.8.2.2 Sistema de Gestión de la Información de CTI

Es un sistema para la gestión de la información desarrollado para la UCI por el departamento DESPROD, que permite evaluar la producción científica de los profesores y especialistas de la UCI y en específico de CEIGE, potencia los resultados científicos y de innovación, premia el trabajo en equipo y se adapta a las características del centro. Para la realización del sistema se utilizó el Modelo de Desarrollo de Software de CEIGE, así como tecnologías Web basadas en el marco de trabajo Sauxe.

[10] Arquitectura Dirigida por Modelos (Model-Driven Architecture)
[11] Plataforma Java, Edición Empresa (Java Platform, Enterprise Edition)

Para la realización del mismo se empleó la librería JavaScript ExtJS, para la programación del lado del cliente. El marco de trabajo Zend, empleado en el desarrollo de múltiples soluciones y servicios webs con PHP[12]. Doctrine, otra de las herramientas empleadas, es una librería para PHP que permite trabajar con un esquema de base de datos como si fuese un conjunto de objetos, y no de tablas y registros (25).

1.8.3 Análisis de los sistemas de gestión descritos

Luego de un análisis de los sistemas informáticos identificados con el propósito de gestionar grandes volúmenes de datos en una organización, no ha sido posible adoptar ninguno de éstos como solución al problema planteado.

Los sistemas internacionales abordados en la investigación presentan un inconveniente común, son privativos lo cual descarta cualquier alternativa de adquirirlos para su implantación ya que el cliente no cuenta con los recursos necesarios para su pago. Por otra parte, en el caso específico del Sistema Servidor de Análisis Sybase IQ, de tipo OLTP, lo cual queda definido en la investigación realizada, no es adecuado para el negocio ya que la idea es realizar análisis de información y para esto lo más propicio sería un sistema modelado multidimensionalmente.

En cuanto a los sistemas nacionales el principal inconveniente que presentan es que ambos poseen aristas diferentes al negocio, o sea, presenta dimensiones que no pueden ser empleadas para calcular los indicadores que necesita el usuario. Llevar a cabo una modificación de los mismos para lograr adaptarlos al negocio, sería un proceso engorroso y extenso donde se tendría que volver a realizar un análisis del sistema, redefinir los requisitos e implementar todo prácticamente desde cero. Por tanto, lo factible es desarrollar uno nuevo enfocado en los requisitos del usuario y que dé cumplimiento a los mismos.

1.9 Metodologías para el desarrollo de un MD

1.9.1 Ralph Kimball

La propuesta de Kimball se basa en dividir el mundo de IN entre los Hechos y las Dimensiones, esta metodología es eficaz y conduce a una solución completa en un corto período de tiempo. Además, tiene abundante documentación y se puede encontrar una respuesta a casi todas las preguntas que se puedan tener (26).

[12] Herramientas de Página de Inicio Personal (Personal Home Page Tools)

Entre sus características principales, Kimball plantea un enfoque ascendente (bottom-up), donde se debe crear por cada departamento un MD, y "El Almacén de Datos es la unión de todos los Mercados de Datos de una entidad" (26).

1.9.2 Hefesto

Entre sus principales directrices plantea que la construcción e implementación de un Almacén de Datos puede adaptarse muy bien a cualquier ciclo de vida de desarrollo de software, con la salvedad de que para algunas fases en particular, las acciones que se han de realizar serán muy diferentes. Lo que se debe tener en cuenta, es no entrar en la utilización de metodologías que requieran fases extensas de reunión de requerimientos y análisis, fases de desarrollo monolítico que conlleve demasiado tiempo y fases de despliegue muy largas. Lo que se busca, es entregar una primera implementación que satisfaga una parte de las necesidades, para demostrar las ventajas del MD y motivar a los usuarios (27).

En la figura 8 se muestran cada una de las fases de la metodología de Hefesto.

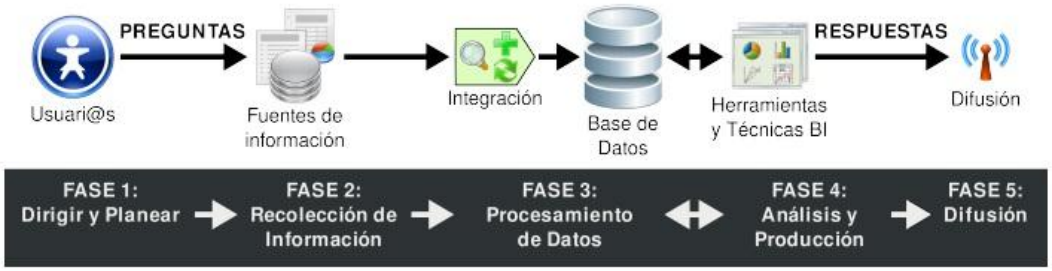


Figura 8: Fases de la Metodología Hefesto.

1.9.3 Empresa de Tecnologías de la Información para la Defensa (XETID)

La metodología XETID es un híbrido entre la metodología de Hefestos y la de Kimball adoptada por el centro de datos de la UCI, adaptándola a las necesidades propias de la entidad. Se decidió realizar un híbrido entre estas dos metodologías y no adoptar una de las dos, porque se quiere aprovechar la facilidad que brinda Hefestos de ser una metodología sencilla y entendible con la de Kimball de generar artefactos para la documentación a la hora de realizar el MD (15).

Es importante destacar que algunos de los artefactos que adoptados por el centro de la XETID fueron modificados pues se determinó que no hacía falta documentar todo lo que éste requería. Además la metodología rectora fue la de Hefestos y no Kimball. Teniendo en cuenta esto se puede decir que las **características** de este híbrido son:

- Los objetivos y resultados esperados en cada fase se distinguen fácilmente y son sencillos de comprender.
- Se basa en los requerimientos del usuario, por lo cual su estructura es capaz de adaptarse con facilidad y rapidez ante los cambios en el negocio.
- Reduce la resistencia al cambio, ya que involucra al usuario final en cada etapa para que tome decisiones respecto al comportamiento y funciones del MD.
- Utiliza modelos conceptuales y lógicos, los cuales son sencillos de interpretar y analizar.
- Es independiente del tipo de ciclo de vida que se emplee para contener la metodología.
- Es independiente de las herramientas que se utilicen para su implementación.
- Es independiente de las estructuras físicas que contengan el MD y de su respectiva distribución.
- Cuando se culmina con una fase, los resultados obtenidos se convierten en el punto de partida para llevar a cabo el paso siguiente (15).

La metodología de XETID establece cuatro fases las cuales se muestran en la figura 9.

Figura 9: Fases de la Metodología de XETID.

1.9.4 Selección de la metodología adecuada

Por todo lo antes planteado, se decide emplear la metodología de XETID, ya que la misma genera la documentación necesaria para el desarrollo del MD. Es una metodología bien estructurada, la cual guía al desarrollador por cada una de las fases que se encuentran bien definidas y logran darle cumplimiento al trabajo propuesto. Utiliza modelos sencillos de analizar e interpretar, los cuales le permiten al cliente verificar el estado en que se encuentra el desarrollo del sistema.

Luego de la selección de la metodología para el desarrollo del MD es necesario realizar el proceso de Extracción-Transformación-Carga de los datos. El cual permitirá a grandes rasgos obtener la información de la fuente de datos, se realiza la limpieza de los mismos para evitar inconsistencia y finalmente se cargan en el MD.

1.10 Proceso ETL

El éxito de un MD se debe sobre todo al buen estudio del negocio a partir de los conocimientos extraídos de las fuentes externas de datos. La integración de los datos es el proceso donde se extraen datos desde múltiples fuentes. Éstos, son limpiados, asegurando la calidad y consistencia de los mismos, se homogenizan los datos de sistemas divergentes para que puedan ser utilizados de una forma conjunta y finalmente se generan los datos en el formato apropiado para ser utilizados posteriormente; luego son cargados en el sistema de almacenamiento para ser analizados.

Este proceso tiene gran importancia, ya que de no realizarse se podrían obtener datos incorrectos que afectarían la toma de decisiones de la organización. Las principales técnicas de integración de datos utilizadas en los sistemas MD son ETL (Extract-Transform-Load) y ELT (Extract-Load-Transform). Para el MD asociado a este trabajo se seleccionó la técnica ETL ya que esta consiste en la extracción, transformación y posteriormente la carga de los datos no siendo así con la técnica ELT donde ocurre la extracción y carga de los datos antes de la transformación de los mismos, por lo que la técnica escogida es la que se adecúa a las metodologías de desarrollo que se utilizarán en el proceso de construcción del MD (27).

La técnica de integración de datos ETL consta de 3 procesos fundamentales.

1.10.1 Extracción

Es aquí, donde, basándose en las necesidades y requisitos del usuario, se exploran las diversas fuentes que se tengan a disposición, y se extrae la información que se considere relevante para el negocio.

Si los datos operacionales residen en un SGBD Relacional, el proceso de extracción se puede reducir a consultas en SQL o rutinas programadas. En cambio, si se encuentran en un sistema no convencional o fuentes externas, ya sean textuales o hojas de cálculos como es el caso del trabajo llevado a cabo, la obtención de los mismos puede ser un tanto dificultoso, debido a que se tendrán que realizar cambios de formato y volcado de información a partir de alguna herramienta específica (27).

Una vez que los datos son seleccionados y extraídos, se guardan en un almacenamiento intermedio, lo cual permite, entre otras ventajas:

- Manipular los datos sin interrumpir ni paralizar los OLAP, ni tampoco el MD.

- No depender de la disponibilidad de los OLAP.
- Almacenar y gestionar los metadatos que se generarán en los procesos ETL.
- Facilitar la integración de las diversas fuentes, internas y externas.

El almacenamiento intermedio constituye en la mayoría de los casos una base de datos en donde la información puede ser almacenada por ejemplo en tablas auxiliares o tablas temporales. Los datos de estas tablas serán los que finalmente (luego de su correspondiente transformación) poblarán el MD (27).

1.10.2 Transformación

Esta función es la encargada de convertir aquellos datos inconsistentes en un conjunto de datos compatibles y congruentes, para que puedan ser cargados en el DW. Estas acciones se llevan a cabo, debido a que pueden existir diferentes fuentes de información, y es vital conciliar un formato y forma única, definiendo estándares, para que todos los datos que ingresarán al MD estén integrados.

Los casos más comunes en los que se deberá realizar integración, son los siguientes:

- Codificación.
- Medida de atributos.
- Convenciones de nombramiento.
- Fuentes múltiples.

Además de lo antes mencionado, esta función se encarga de realizar los procesos de Limpieza y comprobar la Calidad de Datos (27).

1.10.3 Carga

Esta función se encarga, por un lado de realizar las tareas relacionadas con:

- Carga Inicial (Initial Load).
- Actualización o mantenimiento periódico (siempre teniendo en cuenta un intervalo de tiempo predefinido para tal operación).

La carga inicial, se refiere precisamente a la primera carga de datos que se le realizará al MD. Por lo general, esta tarea consume un tiempo bastante considerable, ya que se deben insertar registros que han sido generados aproximadamente, y en casos ideales, durante más de un año. Los mantenimientos periódicos mueven pequeños volúmenes de datos, y su frecuencia

está dada en función del gránulo del MD y los requerimientos del usuario. El objetivo de esta tarea es añadir al depósito aquellos datos nuevos que se fueron generando desde la última actualización (27).

Antes de realizar una nueva actualización, es necesario identificar si se han producido cambios en las fuentes originales de los datos recogidos, desde la fecha del último mantenimiento, a fin de no atentar contra la consistencia del MD. Para efectuar esta operación, se pueden realizar las siguientes acciones:

- Cotejar las instancias de los OLAP involucrados.
- Utilizar disparadores en los OLAP.
- Recurrir a Marcas de Tiempo (Time Stamp), en los registros de los OLAP.
- Comparar los datos existentes en los dos ambientes (OLAP y MD) (27).

A continuación se visualiza, en la figura 10, cómo se realiza el proceso de ETL.

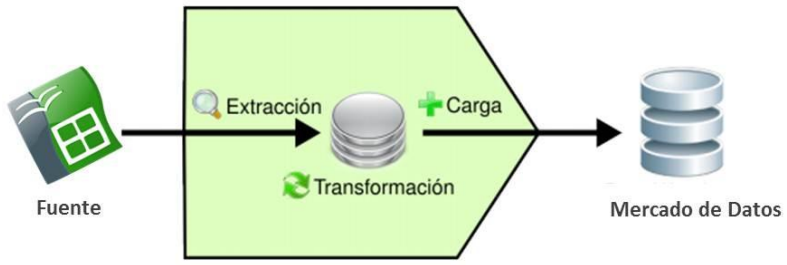


Figura 10: Proceso ETL.

1.11 Herramientas de desarrollo

1.11.1 Herramienta Case Visual Paradigm

Visual Paradigm para UML (VP-UML) es una herramienta de diseño de software diseñada para proyectos de software ágiles y que soporta el ciclo de vida completo de desarrollo del software. Soporta los estándares de modelado como UML, SysML, ERD, DFD, BPMN y ArchiMate. Permite elaborar diagramas de clases, código inverso, generando código a través de estos, así como documentación. Facilita la creación de software y sistemas que se destacan en la experiencia del usuario mediante el apoyo eficaz a la identificación de casos de usos, recopilación de requisitos, el flujo de los acontecimientos y la generación de especificación de requisitos (28).

Para el modelado de este sistema se utilizará la versión 5.0 para UML 8.0, por las grandes

posibilidades que brinda en la Universidad de las Ciencias Informáticas (UCI), lugar donde se desarrolla la investigación, donde existe una licencia para el manejo del mismo lo que posibilita su utilización.

Luego de realizarse el análisis y diseño del MD se necesita emplear una herramienta que sea capaz de crear cada una de las tablas de hechos y dimensiones que se definan, y permita posteriormente su manejo. Para esto entonces se hace la selección de un Sistema Gestor de Base de Datos (SGBD).

1.11.2 Sistema Gestor de Base de Datos PostgreSQL

Creado por el proyecto POSTGRES de la universidad de Berkeley, PostgreSQL es un Sistema Objeto-Relacional, ya que incluye características de la orientación a objetos, como puede ser la herencia, tipos de datos, funciones, restricciones, disparadores, reglas e integridad transaccional (29).

Ventajas de PostgreSQL:

- DBMS (Sistemas Manejadores de Bases de Datos) Objeto-Relacional: Aproxima los datos a un modelo Objeto-Relacional, y es capaz de manejar complejas rutinas y reglas.
- Cliente/Servidor: Usa una arquitectura proceso por usuario cliente/servidor. Hay un proceso maestro que se ramifica para proporcionar conexiones adicionales para cada cliente que intente conectarse a PostgreSQL.
- Altamente Extensible: Soporta los tipos de datos base, así como: tipo, fecha, monetarios, elementos gráficos, datos sobre redes (MAC, IP…) y cadenas de bits.
- Soporte SQL Compresivo: Soporta la especificación SQL99 e incluye características avanzadas tales como las uniones (joins) SQL92.
- Integridad Referencial: Es utilizada para garantizar la validez de los datos de la base de datos.
- Lenguajes Procedurales: Tiene soporte para lenguajes procedurales internos, incluyendo un lenguaje nativo denominado PL/pgSQL. Este lenguaje es comparable al lenguaje procedural de Oracle, PL/SQL.
- MVCC (Multi-Version Concurrency Control) Control de Concurrencia Multi-Versión: Es la tecnología que PostgreSQL usa para evitar bloqueos innecesarios, es decir, permite la lectura sin que sea bloqueada por los que escriben que están actualizando registros.
- Write Ahead Loging (WAL): Esta característica incrementa la dependencia de la base de

datos al registro de cambios antes de que éstos sean escritos en ella. Esto garantiza que en caso de que la base de datos falle o presente un problema, existirá un registro de las transacciones a partir del cual se podrá restaurar la base de datos desde el punto en que se quedó.

- Es un gestor bajo licencia Berkeley Software Distribution (BSD), que posee una gran escalabilidad, haciéndolo idóneo para su uso en sitios web. Además, por su arquitectura de diseño, escala muy bien al aumentar el número de CPUs y la cantidad de RAM.
- Sus tablas pueden llegar a 32 TB, sus tuplas 1.6 TB y los campos a 1GB de tamaño respectivamente.
- El tamaño de la base de datos es ilimitada (29).

Desventajas de PostgreSQL:

- Consume recursos y carga con facilidad el sistema.
- Velocidad de respuesta un poco deficiente al gestionar bases de datos relativamente pequeñas, aunque esta misma velocidad la mantiene al gestionar bases de datos realmente grandes (29).

Una vez creadas cada una de las tablas de hechos y dimensiones del sistema para almacenar la información necesaria para el negocio, se necesita de una herramienta que sea capaz de extraer los datos de la fuente, transformarlos y posteriormente cargarlos hacia el MD. Para realizar este proceso se selecciona la suite libre de Pentaho que brinda un conjunto de herramientas para la realización de cada uno de estos procesos.

1.11.3 Suite Pentaho

La Suite Pentaho para IN es un conjunto de programas libres para dar soporte a la inteligencia empresarial (IE). Incluye herramientas integradas para generar informes, realizar minería de datos y llevar a cabo del proceso de ETL.

Pentaho se define como una plataforma de IN orientada a la solución y centrada en procesos, que incluye todos los principales componentes requeridos para implementar soluciones basadas en procesos. Las soluciones que Pentaho pretende ofrecer se componen fundamentalmente de una infraestructura de herramientas de análisis e informes integrados para procesos de negocio. La plataforma será capaz de ejecutar las reglas de negocio necesarias, expresadas en forma de procesos y actividades, y de presentar y entregar la

información adecuada en el momento que se solicite (30).

Pentaho Data Integration (PDI): También conocido como Kettle, es el componente de Pentaho responsable de los procesos de ETL. El uso más frecuente de toda herramienta ETL, es la población de almacenes de datos. Todos los procesos se crean en un entorno gráfico donde uno especifica qué hacer sin necesidad de escribir código para indicar cómo hacerlo. Como herramienta ETL, es la más popular entre las de código abierto. Soporta amplia variedad de formatos de entradas y salidas, incluyendo archivos planos, planillas de cálculo, y conexión con motores de bases de datos tanto comerciales como abiertos (31).

Pentaho BI Server: Provee el soporte y la infraestructura necesarios para crear soluciones de inteligencia empresarial a problemas de negocios. El marco proporciona los servicios básicos, incluidos autenticación, registro, auditoría, servicios web y motor de reglas. La plataforma también incluye un motor de solución que integra reportes, análisis, tableros de comandos y componentes de minería de datos (30).

1.12 Conclusiones del Capítulo

Luego del estudio realizado de las tecnologías, metodologías y estándares para el desarrollo del MD, posibilitó la selección de la Metodología de desarrollo de XETID, la cual servirá para darle cumplimiento a las necesidades del cliente en tiempo y guiar correctamente el desarrollo del trabajo. Por otra parte las herramientas seleccionadas permitirán la correcta elaboración de la solución. El estudio realizado respecto a los sistemas existentes para el análisis y almacenamiento de información CTI, permitió descartar los sistemas existentes como posibles soluciones, debido a que, en su gran mayoría son privativos o no cuentan con las mismas aristas de análisis del negocio en cuestión. Se propone como posible esquema, el de constelación de hechos.

CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS

2.1 Introducción

En el presente capítulo se identifican los requisitos del usuario a través de preguntas que explicitan los objetivos del negocio, se analizan dichas preguntas a fin de identificar cuáles serán los indicadores y perspectivas que serán tomadas en cuenta para la construcción del MD.

Se definen qué campos se incluirán en cada perspectiva y los indicadores que se calculan para darle cumplimiento a los requisitos del cliente. Además se definen las reglas del negocio y el nivel de granularidad del MD. Se confecciona además, un modelo conceptual donde se podrá visualizar la relación existente entre los indicadores y perspectivas. Finalmente se identifican las dimensiones y hechos que conllevarán al diseño de las tablas y columnas físicas del MD quedando confeccionado el modelo de datos del sistema.

2.2 Análisis

El objetivo principal de este paso, es obtener e identificar las necesidades de información claves de alto nivel, que es esencial para llevar a cabo las metas y estrategias de la entidad, y que facilitará una eficaz y eficiente toma de decisiones. Se debe tener en cuenta que dicha información, es la que proveerá el soporte para desarrollar los pasos sucesivos, por lo cual, es muy importante que se preste especial atención al relevar los datos.

Análisis de los Requisitos

El análisis de los requerimientos de los diferentes usuarios, es el punto de partida del trabajo, ya que ellos son los que deben, en cierto modo, guiar la investigación hacia un desarrollo que refleje claramente lo que se espera del depósito de datos, en relación a sus funciones y cualidades.

2.2.1 Identificación de Preguntas

La idea central de este subepígrafe es que se formulen preguntas complejas sobre el negocio, que incluyan variables de análisis que se consideren relevantes, ya que son éstas las que permitirán estudiar la información desde diferentes perspectivas.

A continuación se muestra una lista con las preguntas identificadas de mayor relevancia para el cliente, las cuales fueron aprobadas por el mismo:

1. Cantidad de trabajos presentados por tipos por cada área en un tiempo determinado.

2. Cantidad de trabajos planificados por tipos por cada área en un tiempo determinado.
3. Monto total de trabajos planificados por área en un tiempo determinado.
4. Monto total de trabajos planificados en el centro en un tiempo determinado.
5. Cantidad de publicaciones planificadas por tipos por área en un tiempo determinado.
6. Cantidad de publicaciones por tipos por área en un tiempo determinado.
7. Cantidad de registros por tipos por cada área en un tiempo determinado.
8. Cantidad de resultados introducidos por tipos por cada área en un tiempo determinado.
9. Cantidad de premios por tipos por cada área en un tiempo determinado.
10. Porciento de los trabajos presentados de los planificados en un tiempo de terminado.
11. Monto total de trabajos presentados por área en un tiempo determinado.
12. Monto total de trabajos presentados en el centro en un tiempo determinado.
13. Monto total de publicaciones por área en un tiempo determinado.
14. Monto total de publicaciones en el centro en un tiempo determinado.
15. Monto total de registros por área en un tiempo determinado.
16. Monto total de registros en el centro en un tiempo determinado.
17. Monto total de resultados introducidos por área en un tiempo determinado.
18. Monto total de resultados introducidos en el centro en un tiempo determinado.
19. Monto total de premios por área en un tiempo determinado.
20. Monto total de premios en el centro en un tiempo determinado.
21. Cantidad de trabajos no certificados por tipo por área en un tiempo determinado.
22. Cantidad de publicaciones no certificadas por tipo por área en un tiempo determinado.
23. Cantidad de premios no certificados por tipo por área en un tiempo determinado.
24. Monto total de trabajos presentados no certificados por área en un tiempo determinado.
25. Monto total de publicaciones no certificadas por área en un tiempo determinado.
26. Monto total de premios no certificados por área en un tiempo determinado.
27. Monto total de trabajos no certificados en el centro en un tiempo determinado.
28. Monto total de publicaciones no certificadas en el centro en un tiempo determinado.
29. Monto total de premios no certificados en el centro en un tiempo determinado.
30. Cantidad de profesionales por categoría docente de cada una de las áreas en un tiempo determinado.

31. Cantidad de profesionales por categoría científica de cada una de las áreas en un tiempo determinado.
32. Cantidad de profesionales por preferencias políticas de cada una de las áreas en un tiempo determinado.
33. Proyección de discusión de maestrías por área en un tiempo determinado.
34. Proyección total de discusiones de maestrías del centro en un tiempo determinado.

Una vez que se establecidas las preguntas claves, se debe proceder a su descomposición para descubrir los indicadores que se utilizarán y las perspectivas de análisis que intervendrán. Quedando definidas como perspectivas:

- Trabajos.
- Publicaciones.
- Registros.
- Resultados.
- Premios.
- Categoría docente.
- Categoría científica.
- Preferencia política.
- Área.
- Tiempo.

2.2.2 Identificar Perspectivas e Indicadores

Primeramente se debe tener en cuenta que los indicadores, para que sean realmente efectivos, deben ser valores numéricos y representar lo que se desea analizar concretamente. En cambio, las perspectivas se refieren a los objetos mediante los cuales se quiere examinar los indicadores, con el fin de responder a las preguntas planteadas.

Por cada una de las preguntas se genera un indicador en el cual para calcular su valor intervienen algunas de las perspectivas identificadas previamente. La relación existente entre indicadores y perspectivas se describen a continuación en la siguiente tabla:

Pregunta	Descripción	Indicador	Perspectivas
1	Cantidad de trabajos presentados	ctp	Trabajo, Área, Tiempo
2	Cantidad de trabajos planificados	ctplan	Trabajo, Área, Tiempo

3	Total de trabajos planificados en el centro	ttplanc	Tiempo
4	Total de trabajos planificados	ttplan	Área, Tiempo
5	Cantidad de publicaciones planificadas	cpplan	Publicación, Área, Tiempo
6	Cantidad de publicaciones	cp	Publicación, Área, Tiempo
7	Cantidad de registros	creg	Registro, Área, Tiempo
8	Cantidad de resultados	cres	Resultado, Área, Tiempo
9	Cantidad de premios	cprem	Premio, Área, Tiempo
10	Por ciento de trabajos presentados de los planificados en el centro	ptpvsplanc	Tiempo
11	Total de trabajos presentados	ttp	Área, Tiempo
12	Total de trabajos presentados en el centro	ttpc	Tiempo
13	Total de publicaciones	tp	Área, Tiempo
14	Total de publicaciones en el centro	tpc	Tiempo
15	Total de registros	treg	Área, Tiempo
16	Total de registros del centro	tregc	Tiempo
17	Total de resultados	tres	Área, Tiempo
18	Total de resultados del centro	tresc	Tiempo
19	Total de premios	tprem	Área, Tiempo

20	Total de premios del centro	tpremc	Tiempo
21	Cantidad de trabajos no certificados	ctnc	Trabajo, Área, Tiempo
22	Cantidad de publicaciones no certificadas	cpnc	Publicación, Área, Tiempo
23	Cantidad de premios no certificados	cpremnc	Premio, Área, Tiempo
24	Total de trabajos no certificados	ttnc	Área, Tiempo
25	Total de publicaciones no certificadas	tpnc	Área, Tiempo
26	Total de premios no certificados	tpremnc	Área, Tiempo
27	Total de trabajos no certificados del centro	ttncc	Tiempo
28	Total de publicaciones no certificadas del centro	tpncc	Tiempo
29	Total de premios no certificados del centro	tpremncc	Tiempo
30	Cantidad de profesionales	cpcd	Categoría Docente, Área, Tiempo
31	Cantidad de profesionales	cpcc	Categoría Científica, Área, Tiempo
32	Cantidad de profesionales	cppp	Preferencia Política, Área, Tiempo
33	Proyección de discusión de maestrías	pdm	Área, Tiempo
34	Proyección total de discusiones de maestrías del centro	pdmc	Tiempo

Tabla 1: Relación entre indicadores y perspectivas.

Una vez realizado el análisis de la relación existente indicadores-perspectivas se realiza una agrupación de los indicadores por las perspectivas comunes, los cuales se definirán a continuación.

2.2.3 Agrupación de indicadores por perspectivas comunes.

La agrupación de los indicadores teniendo en cuenta cada una de las perspectivas comunes, posibilitará posteriormente diseñar el modelo conceptual, o los modelos conceptuales, que permitirán obtener una visión clara de cómo quedará organizado el MD teniendo en cuenta los hechos y dimensiones implicadas.

En la investigación quedan definidos diez grupos, teniendo en cuenta las perspectivas comunes entre cada indicador, los mismos se reflejan a continuación en la siguiente tabla:

Grupos	Indicadores	Perspectivas
1	ctp, ctplan, ctnc	Trabajo, Área, Tiempo
2	cp, cpplan, cpnc	Publicación, Área, Tiempo
3	creg	Registro, Área, Tiempo
4	cres	Resultado, Área, Tiempo
5	cprem, cpremnc	Premio, Área, Tiempo
6	cpcd	Categoría Docente, Área, Tiempo
7	cpcc	Categoría Científica, Área, Tiempo
8	cppp	Preferencia Política, Área, Tiempo

9	ttp, ttplan, tp, treg, tres, tprem, ttnc, tpnc, tpremnc, pdm	Área, Tiempo
10	ttplanc, ptpvsplanc, ttpc, tpc, tregc, tresc, tpremc, ttncc, tpncc, tpremncc, pdmc	Tiempo

Tabla 2: Definición de los grupos por indicadores según las perspectivas comunes.

2.2.4 Definición de las Tablas de Hechos y Dimensiones

Las tablas de hechos contienen, precisamente, los hechos que serán utilizados por los analistas del negocio para apoyar el proceso de toma de decisiones.

Los hechos son datos instantáneos en el tiempo, que son filtrados, agrupados y explorados a través de condiciones definidas en las tablas de dimensiones (27). En el MD en desarrollo, se definió una tabla de hechos por cada arista de análisis que se desea estudiar, y se calculan los indicadores que responden a las preguntas formuladas. A continuación se enumeran las tablas de hechos que conformarán el MD:

1. cub_hechos_trabajos.
2. cub_hechos_publicaciones.
3. cub_hechos_registros.
4. cub_hechos_resultados.
5. cub_hechos_premios.
6. cub_hechos_categorias_docentes.
7. cub_hechos_categorias_cientificas.
8. cub_hechos_preferencias_politicas.
9. cub_hechos_cti_area.
10. cub_hechos_cti_centro.

Las tablas de dimensiones definen cómo los datos están organizados lógicamente y cómo proveen el medio para analizar el contexto del negocio. Contienen datos cualitativos y representan los aspectos de interés del usuario por los que se filtra y manipula la información almacenada en las tablas de hechos (27).

Una vez identificadas las aristas de análisis del negocio en cuestión, se definieron como dimensiones del MD, las perspectivas identificadas anteriormente. A continuación se enumeran dichas dimensiones:

- dim_trabajo.
- dim_publicacion.

- dim_registro.
- dim_resultado.
- dim_premio.
- dim_categoria_docente.
- dim_categoria_cientifica.
- dim_preferencia_politica.
- dim_area.
- dim_tiempo.

Una vez definido cada una de las dimensiones y hechos es necesario definir el nivel de granularidad del MD.

Definición del Nivel de Granularidad de las Dimensiones

La granularidad representa el nivel de detalle al que se desea almacenar la información sobre el negocio que se esté analizando. Mientras mayor sea el nivel de detalle de los datos se tendrán mayores posibilidades analíticas (27).

El nivel de granularidad definido para cada una de las dimensiones está representado en la tabla 3.

Dimensión	Nivel de Granularidad
dim_trabajo	id_trabajo, tipo_trabajo
dim_publicacion	id_publicacion, tipo_publicacion
dim_registro	id_registro, tipo_registro
dim_resultado	id_resultado, tipo_resultado
dim_premio	id_premio, tipo_premio
dim_categoria_docente	id_categoria_docente, tipo_categoria_docente
dim_categoria_cientifica	id_categoria_cientifica, tipo_categoria_cientifica
dim_preferencia_politica	id_preferencia_politica, tipo_preferencia_politica
dim_area	id_area, nombre_area
dim_tiempo	id_tiempo, año, trimestre, mes, dia

Tabla 3: Definición del Nivel de Granularidad.

Con la definición de los indicadores, las perspectivas, dimensiones, hechos y el nivel de detalle para la consulta de la información se cuentan con todas las herramientas necesarias para el diseño del MD.

2.3 Diseño

En este epígrafe quedan diseñados los modelos conceptuales asociados a cada área de análisis, así como el modelo lógico del MD, el cual permite posteriormente el diseño del modelo de datos, en el que se representan físicamente las tablas de hechos y dimensiones con cada uno de sus atributos.

2.3.1 Diseño de los Modelos Conceptuales

En esta etapa, se construye el modelo conceptual a partir de los indicadores y perspectivas obtenidas epígrafe 2.2.2 y teniendo en cuenta cada una de las áreas definidas en el epígrafe 2.2.3.

A través de estos modelos, se podrá observar con claridad cuáles son los alcances del proyecto, para luego poder trabajar sobre el mismo, además al poseer un alto nivel de definición de los datos, permite su presentación y explicación ante los usuarios con facilidad.

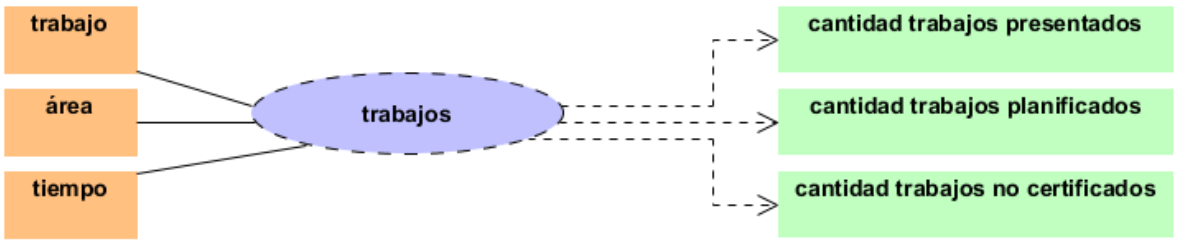


Figura 11: Modelo conceptual del área de análisis # 1 para los trabajos.

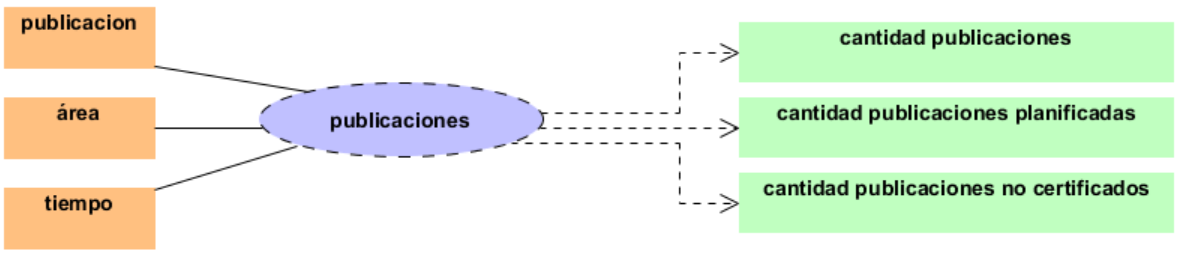


Figura 12: Modelo conceptual del área de análisis # 2 para las publicaciones.

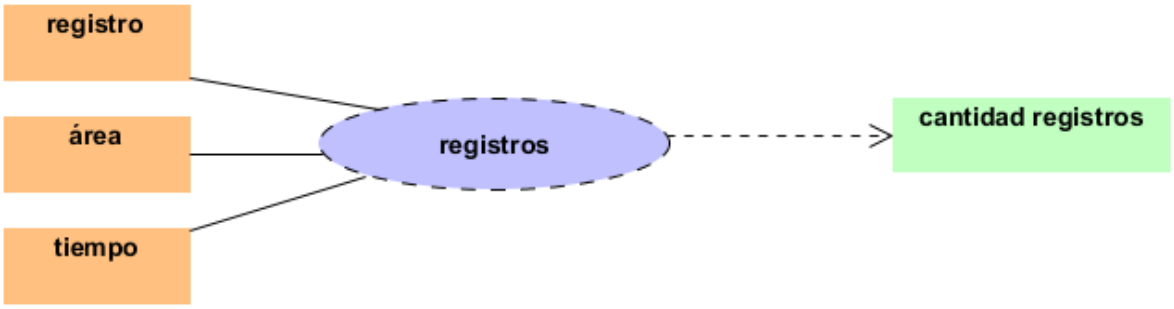


Figura 13: Modelo conceptual del área de análisis # 3 para los registros.

Figura 14: Modelo conceptual del área de análisis # 4 para los resultados.

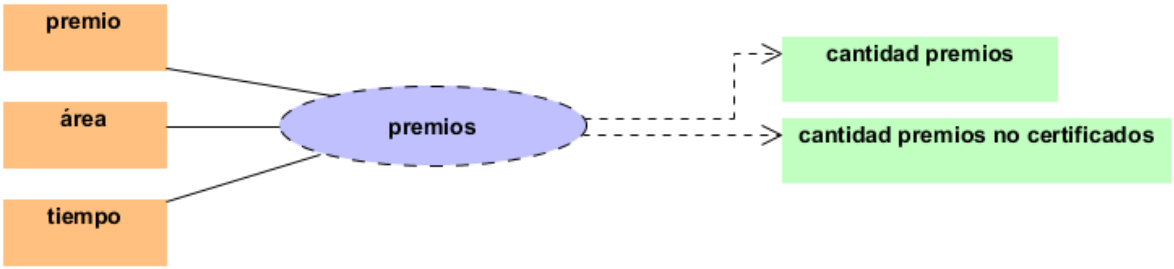


Figura 15: Modelo conceptual del área de análisis # 5 para los premios.

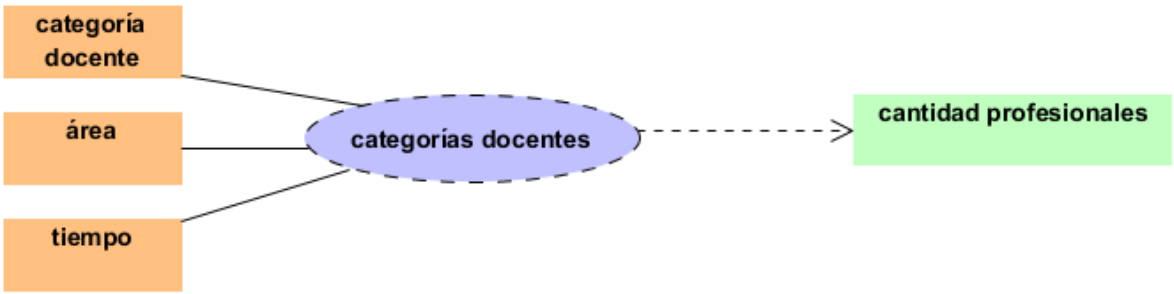


Figura 16: Modelo conceptual del área de análisis # 6 para las categorías docentes.

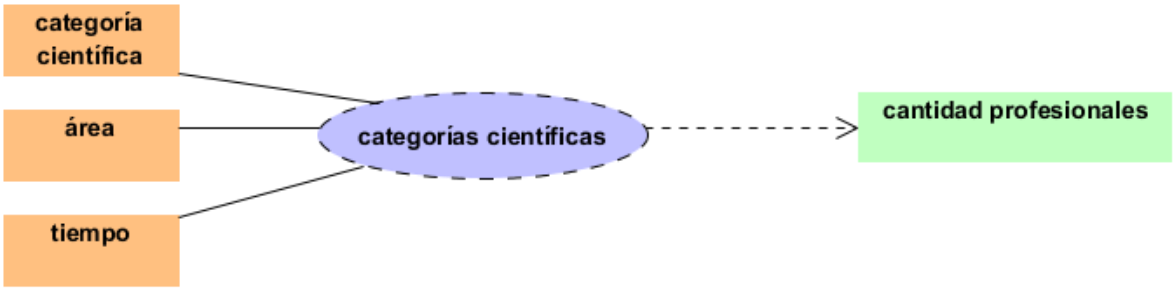


Figura 17: Modelo conceptual del área de análisis # 7 para las categorías científicas.

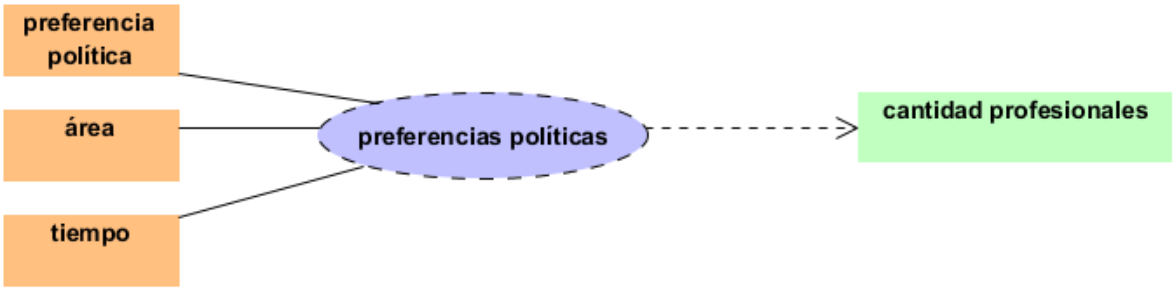


Figura 18: Modelo conceptual del área de análisis # 8 para las preferencias políticas.

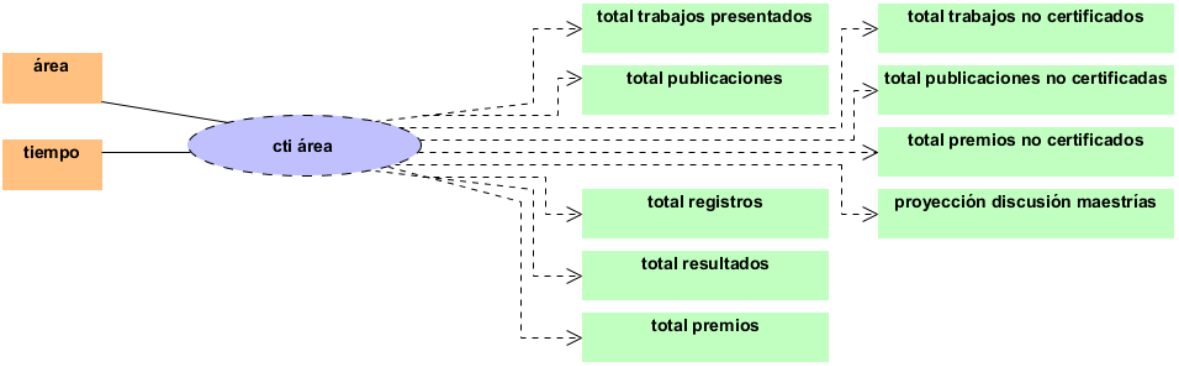


Figura 19: Modelo conceptual del área de análisis # 9 para indicadores CTI por área.

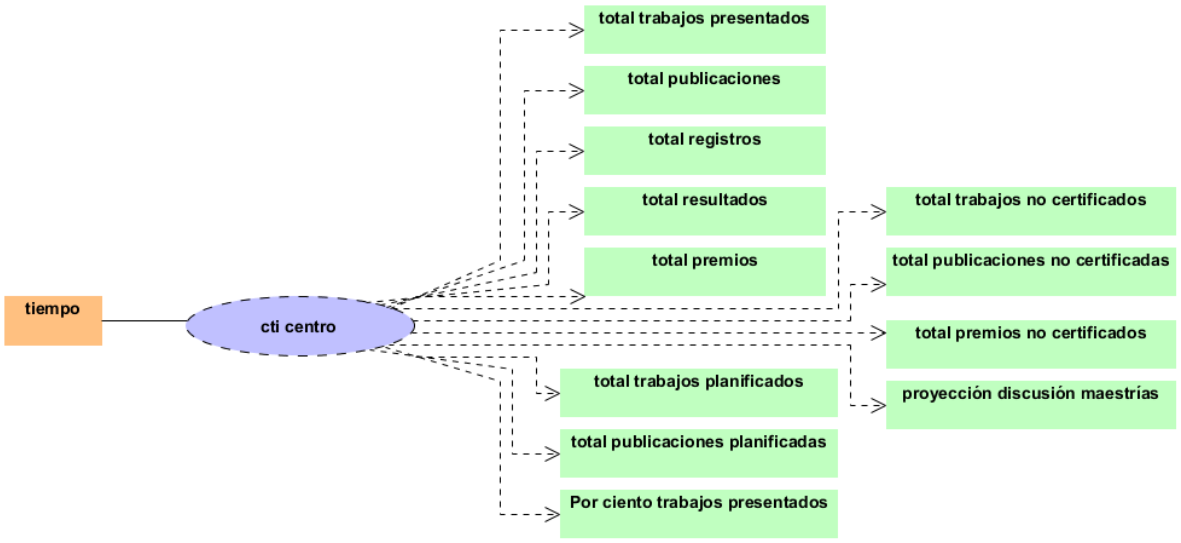


Figura 20: Modelo conceptual del área de análisis # 10 para los indicadores CTI del centro.

Luego del diseño de cada uno de los modelos conceptuales del sistema se procede a diseñar el modelo de datos del MD empleando el esquema constelación de hechos.

2.3.2 Modelo de Datos del MD

Finalmente se realizó el diseño del modelo de datos del MD. En la figura 21 se pueden observar las tablas de dimensiones y hechos incluidas y las relaciones entre ambas.

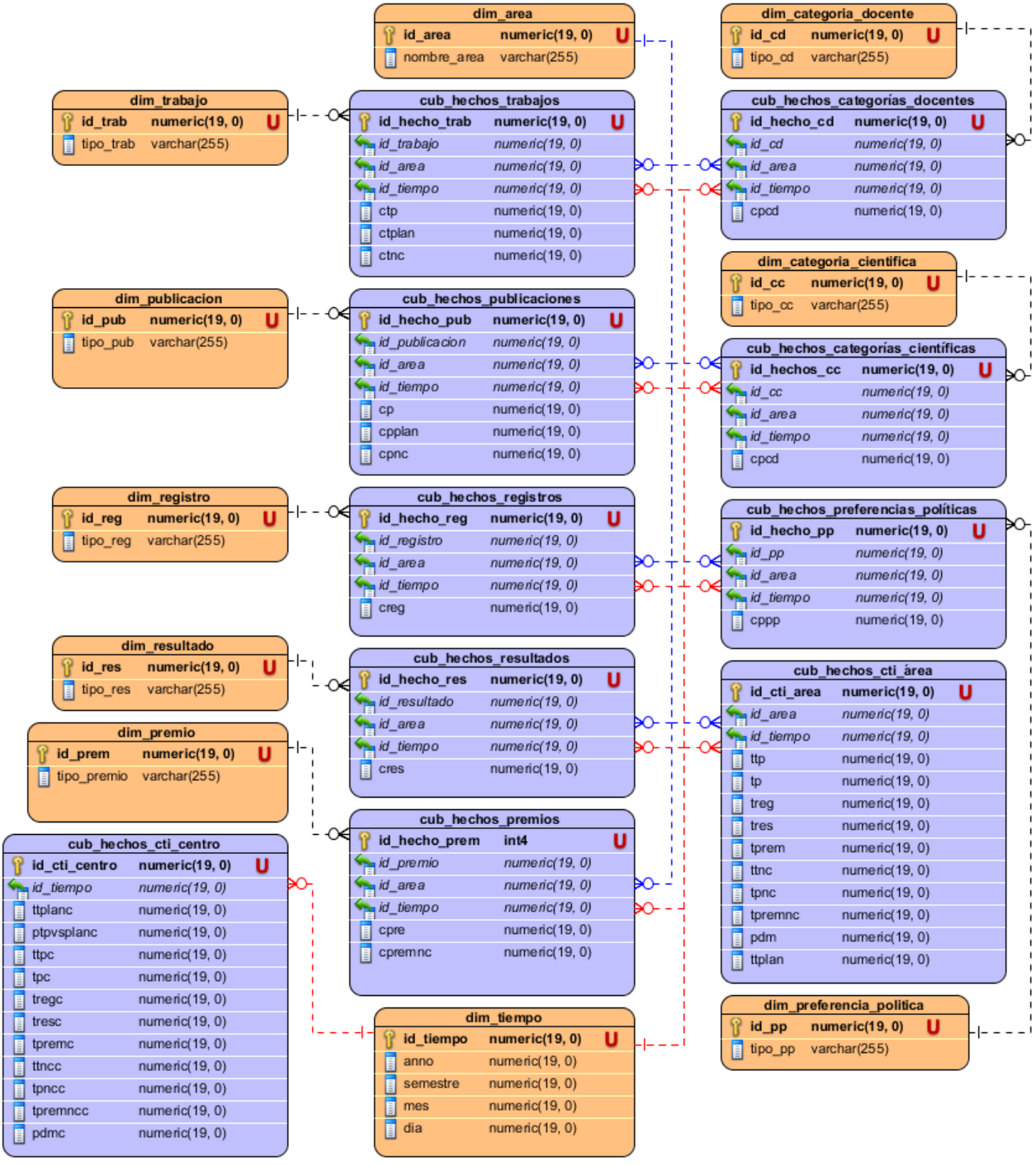


Figura 21: Modelo de Datos del MD.

2.4 Conclusiones del Capítulo

Luego de transitar por las dos primeras fases que propone la metodología de la XETID, se logró la identificación de un total de 34 preguntas, posibilitando conocer las necesidades de información del cliente, se identificaron y definieron los indicadores y perspectivas de análisis, y el nivel de granularidad de cada perspectiva, lo que permitió identificar el nivel de detalle al que se podrán realizar estudios a los datos almacenados en el sistema. La elaboración del modelo de datos del MD, posibilitó verificar que el diseño elaborado satisface las necesidades de información del cliente. Una vez generado cada uno de los artefactos pertinentes a estas fases, es posible continuar con el siguiente capítulo para llevar a cabo la implementación y validación del sistema.

CAPÍTULO 3: IMPLEMENTACIÓN Y VALIDACIÓN DEL MERCADO DE DATOS

3.1 Introducción al Capítulo

Al finalizarse el proceso de análisis y diseño del MD, resulta importante adentrarse en los temas de implementación y validación. Luego de identificadas las medidas, hechos y dimensiones correctamente, se procede a realizar el proceso ETL de los datos.

Finalizada la implementación del MD será sometido a pruebas de rendimiento con el comienzo del almacenamiento de los datos a partir del presente año y con la interacción del usuario final con el sistema. Estas pruebas permitirán la validación de la propuesta de solución planteada, comprobar la calidad de los datos contenidos en el MD y verificar el rendimiento del mismo teniendo en cuenta las exigencias del cliente.

3.2 Proceso ETL

Una vez creadas las tablas físicas del MD como resultado de las fases anteriores de la metodología aplicada, se puede proceder al llenado del MD mediante los procesos ETL. Para esto se realizará primeramente un mapeo de los datos fuentes hacia los destinos detallando claramente de dónde salen y hacia dónde se dirigen, luego se establecerán algunas restricciones y condiciones adicionales, con el fin de no obviar ningún aspecto relevante del negocio. Luego se procederá a la carga de los datos hacia las dimensiones y hechos según corresponda. Finalmente se diseñará y construirá la automatización de todo el proceso (10).

3.2.1 Mapeo de los datos fuente al destino

Se establece un mapeo de los datos desde la fuente hacia el destino de los mismos con el objetivo de evitar pérdidas de datos en el proceso ETL y establecer la relación existente entre los campos de la fuente y su correspondiente en la base de datos destino. La tabla número 4 muestra el mapeo de datos realizado.

Mapa Lógico de Datos						
Fuente				**Nombre BD Destino: CTI**		
Nombre del documento	**Nombre de la hoja**	**Nombre de la columna**	**Tipo de dato**	**Nombre de la tabla**	**Nombre del campo**	**Tipo de dato**
Balance CTI CEIGE	Eventos	Clasificación	String	dim_trabajo	tipo_trabajo	text
Balance CTI CEIGE	Publicaciones	Clasificación	String	dim_publicacion	tipo_publicacion	text
Balance CTI CEIGE	Registros	Clasificación	String	dim_registro	tipo_registro	text
Balance CTI CEIGE	Resultados introducidos	Resultado	String	dim_resultado	tipo_resultado	text
Balance CTI CEIGE	Premios	Clasificación	String	dim_premio	tipo_premio	text
Plan Perspectivo	Datos	Área	String	dim_area	nombre_area	text
Plan Perspectivo	Datos	Integración	String	dim_preferencia_politica	tipo_pp	text
Plan Perspectivo	Datos	Categoría Docente	String	dim_categoria_docente	tipo_cd	text
Plan Perspectivo	Hoja1	Grado científico	String	dim_categoria_cientifica	tipo_cc	text

Tabla 4: Mapeo de Datos.

La tabla “dim_tiempo” del MD no se incluyó en el mapeo de datos realizado ya que la misma no se construyó a partir de los datos de la fuente origen, sino que se generaron todos los campos que la integran, a partir de la definición de la fecha de inicio y la fecha de fin, a partir del 1 de enero del 2001 hasta 40000 días posteriores, o sea, hasta el día 7 de julio del 2109. Para la generación de este gran intervalo de fechas se utilizó la herramienta Pentaho Data Integration (PDI) de la suite de Pentaho.

3.2.2 Establecer condiciones adicionales y de restricciones

Para la realización del proceso ETL es necesario tener en cuenta una serie de condiciones y restricciones adicionales, mediante las que se garantizará que los datos persistentes en el MD

sean lo más confiables posible. Las restricciones establecidas son las siguientes:

1. Para el llenado de la tabla “cub_hechos_cat_cient” es necesario que los indicadores sean calculados por la categoría científica, el área y tiempo.
2. Para el llenado de la tabla “cub_hechos_cat_doc” es necesario que los indicadores sean calculados por la categoría docente, el área y tiempo.
3. Para el llenado de la tabla “cub_hechos_cti_area” es necesario que los indicadores sean calculados por el área y tiempo.
4. Para el llenado de la tabla “cub_hechos_cti_centro” es necesario que los indicadores sean calculados por el tiempo.
5. Para el llenado de la tabla “cub_hechos_prefrencias_politicas” es necesario que los indicadores sean calculados por la preferencia política, el área y tiempo.
6. Para el llenado de la tabla “cub_hechos_premios” es necesario que los indicadores sean calculados por los premios, el área y tiempo.
7. Para el llenado de la tabla “cub_hechos_publicaciones” es necesario que los indicadores sean calculados por las publicaciones, el área y tiempo.
8. Para el llenado de la tabla “cub_hechos_registros” es necesario que los indicadores sean calculados por los registros, el área y tiempo.
9. Para el llenado de la tabla “cub_hechos_resultados” es necesario que los indicadores sean calculados por los resultados, el área y tiempo.
10. Para el llenado de la tabla “cub_hechos_trabajos” es necesario que los indicadores sean calculados por los trabajos, el área y tiempo.
11. Para el poblado de las tablas de hechos será necesario realizar el poblado de las tablas dimensiones primeramente.
12. En caso de que en la fuente de datos existan campos con valores nulos, vacíos o anómalos, éstos deberán ser transformados por el nuevo valor: “desconocido”.
13. Para realizar la carga de la tabla de hechos “cub_hechos_cti_area” es necesario haber realizado la carga de las tablas de hechos “cub_hechos_trabajos”, “cub_hechos_resultados”, “cub_hechos_registros”, “cub_hechos_publicaciones” y “cub_hechos_premios”.
14. Para realizar la carga de la tabla de hechos “cub_hechos_cti_centro” es necesario haber realizado el poblado de la tabla de hechos “cub_hechos_cti_area”.

3.2.3 Cargas incrementales de datos

El proceso de carga de datos hacia el MD se llevó a cabo a partir de la extracción de los datos fuente, la transformación de los mismos para evitar valores faltantes o anómalos y finalmente la carga en las dimensiones y hechos correspondientes del MD. Para la implementación del proceso ETL se utilizó la herramienta PDI.

En la implementación del proceso ETL se crearon dos paquetes de integración, “ETL carga inicial CTI”, el cual tiene como objetivo realizar la primera carga hacia el MD y “ETL carga mensual CTI”, con el objetivo de realizar cargas incrementales a final de cada mes. Se implementan dos paquetes de integración, ya que en la carga inicial se realiza el poblado de la dimensión dim_tiempo, lo cual no es necesario ejecutar mensualmente ya que en dicha dimensión se incluye un margen de más de cien años y solo sería necesario incluir nuevas fechas una vez que se aproxime la fecha fin establecida.

En los paquetes de integración para la carga inicial y final del MD, se realizó la implementación del escenario de flujo de datos de los procesos ETL de los hechos y dimensiones. En el caso de las dimensiones se extrajeron datos de múltiples hojas de los documentos Excel fuentes, los datos fueron sometidos a transformaciones, pues se hizo necesario validar que la información que se insertara en el MD no presentara valores vacíos o anómalos. Finalmente los datos transformados fueron insertados o actualizados en el MD destino. La figura 22 muestra el proceso ETL implementado para el llenado de la dimensión “dim_trabajo”. El resto del proceso para el llenado de las dimensiones restantes se puede encontrar en los anexos desde el número 2 al 10.

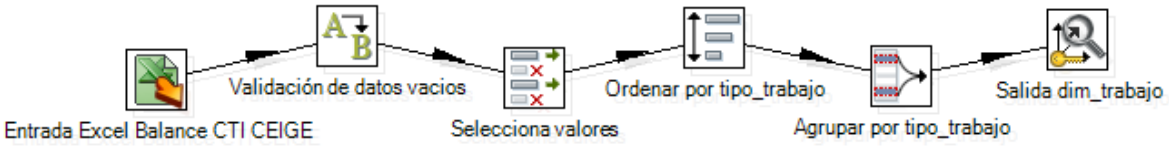


Figura 22: Proceso ETL para el poblado de la dimensión “dim_trabajo”.

Para la implementación de los procesos ETL de las tablas de hechos, se siguió básicamente el mismo proceder que en el caso de la tablas dimensiones, teniendo en cuenta los datos origen y los datos almacenados en las dimensiones, estableciéndose relación entre los mismos, con el objetivo de obtener los identificadores necesarios para el cálculo y posterior almacenamiento de los indicadores o medidas del hecho. La figura número 23 muestra el flujo del proceso ETL para

el poblado de la tabla de hechos “cub_hechos_cat_doc”, las restantes transformaciones se pueden visualizar en los anexos desde el número 11 al 21.

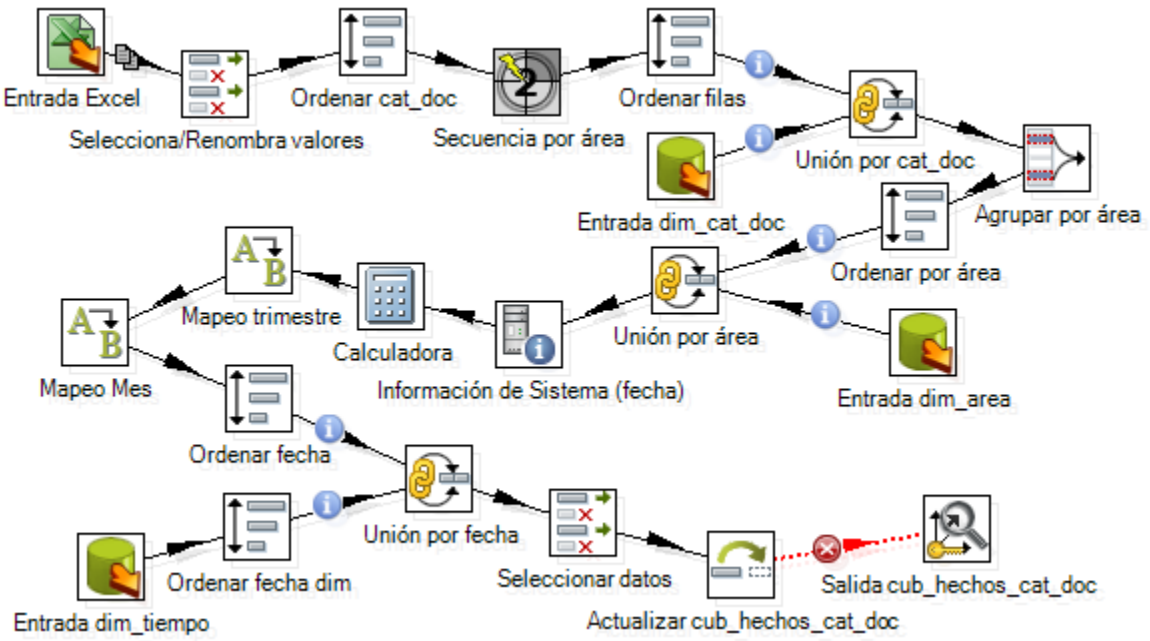


Figura 23: Proceso ETL para el poblado de la tabla de hechos “cub_hechos_cat_doc”.

3.2.4 Automatización del proceso ETL

Luego de la elaboración de cada uno de los procesos ETL para cada dimensión y hechos, se procede a la automatización de los mismos. Para llevar a cabo esto se procede a crear dos trabajos, denominados “ETL carga inicial CTI” y “ETL carga incremental CTI” mediante el uso de la herramienta PDI.

Para ambas cargas se realizan configuraciones similares, con la particularidad que en el primero se incluye la transformación para la tabla “dim_tiempo”, de allí que se denomine carga inicial. Para la segunda carga, ejecutada mensualmente los días 28 a las 22:00 horas específicamente, no se incluye el poblado de la tabla “dim_tiempo”. En la figura número 24 se muestra el escenario de control de flujo del proceso ETL para la carga inicial del MD, el anexo número 22 muestra el proceso de carga mensual.

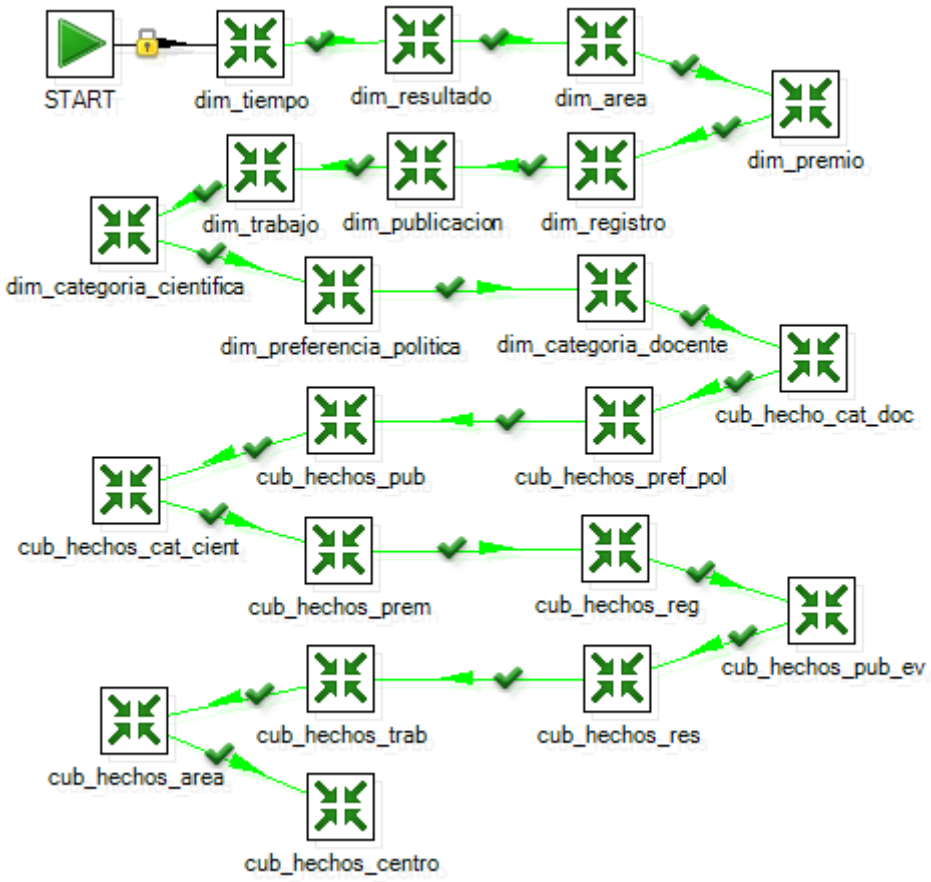


Figura 24: Proceso ETL para la carga inicial de los datos del MD.

3.3 Validación del Mercado de Datos

La validación de todo sistema informático, es el proceso de revisión que verifica que el sistema producido cumple con las especificaciones y requisitos planteados por el cliente durante el ciclo de desarrollo, y verificando además, el rendimiento del sistema en un ambiente de explotación similar al que se desea emplear.

Para cumplir con lo planteado anteriormente se realizaron dos grupos de pruebas al MD, el primer grupo se enfocó a validar el correcto funcionamiento de la solución plateada, para esto se realizaron pruebas de funcionalidad, pruebas de calidad de datos y pruebas de confiabilidad. El segundo grupo de pruebas se centró en verificar el correcto desempeño del MD para esto se realizaron pruebas de rendimiento.

3.3.1 Pruebas para validar la calidad del Mercado de Datos

3.3.1.1 Pruebas de funcionalidad

La prueba de funcionalidad se realiza con el objetivo de verificar que se pueden emplear los datos persistentes en el MD para la realización de estudios y análisis, a partir de la selección indistinta de determinadas dimensiones y medidas. Para la realización de esta prueba se construyeron un total de diez cubos OLAP, mediante la utilización de la herramienta Pentaho BI-Server.

Para la realización de la prueba de funcionalidad se diseñaron los siguientes cubos OLAP: “cub_cat_cient”, “cub_cat_doc”, “cub_cti_area”, “cub_cti_centro”, “cub_pref_pol”, “cub_premios”, “cub_publicaciones”, “cub_registros”, “cub_resultados”, “cub_trabajos”, los cuales son una representación multidimensional de las tablas de hechos del MD: “cub_hechos_cat_cient”, “cub_hechos_cat_doc”, “cub_hechos_cti_area”, “cub_hechos_cti_centro”, “cub_hechos_prefrencias_politicas”, “cub_hechos_premios”, “cub_hechos_publicaciones”, “cub_hechos_registros”, “cub_hechos_resultados” y “cub_hechos_trabajos” respectivamente. La figura 25 muestra el diseño del cubo “cub_hechos_trabajos”, con las medidas y dimensiones asociadas al mismo, el resto de las representaciones se pueden visualizar en los anexos desde el número 23 al 31.

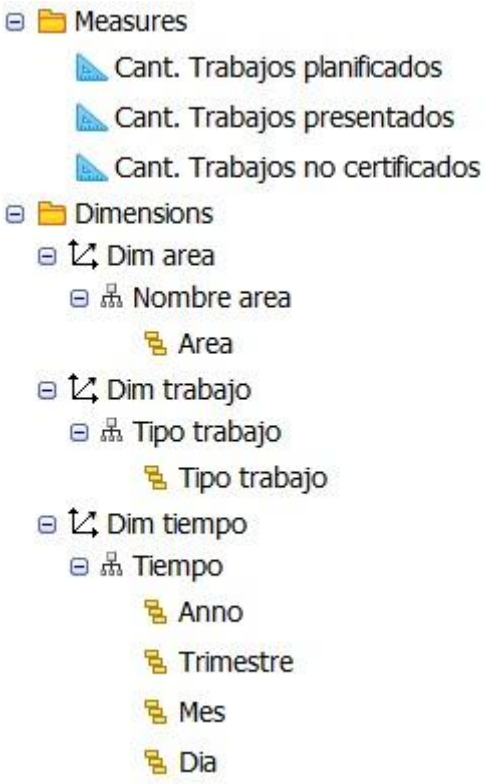


Figura 25: Representación del cubo "cub_hechos_trabajos".

Luego de construidos los cubos OLAP fue posible visualizar la información contenida en el MD desde todos los niveles de detalle que se definió en el diseño. Actualmente es posible navegar por las diferentes estructuras según los datos que se deseen estudiar, éstos pueden ser visualizados, sin necesidad de realizar manualmente consultas SQL al MD (10).

Los cubos "cub_cat_cient", "cub_cat_doc" y "cub_pref_pol" se destinaron al análisis de los profesionales respecto a la categoría científica, categoría docente y preferencias políticas respectivamente, mientras que los cubos "cub_premios", "cub_publicaciones", "cub_registros", "cub_resultados" y "cub_trabajos", se destinaron al estudio de la cantidad de premios, publicaciones, registros, resultados y trabajos respectivamente. En el caso de los cubos "cub_cti_area" y "cub_cti_centro" son destinados para el análisis total, para el análisis de las proyecciones de discusión de maestrías y porciento de trabajos presentados respectos a los planificados, con la peculiaridad de que para el análisis del cubo "cub_cti_area" se tiene en cuenta cada una de las áreas del centro, mientras que en el "cub_cti_centro" se realiza teniendo en cuenta solo el tiempo y de esta forma se representa la información a nivel de centro.

La figura 26 muestra la navegación por los datos referentes al cubo OLAP “cub_cti_centro”, mostrando la información referente al total de trabajos planificados, total de trabajos presentados, total publicaciones, total resultados, total registros, total premios, proyecciones de discusión de maestrías y el porciento de trabajos presentados respecto a los planificados según el tiempo establecido.

	Medidas							
Tiempo	Total trabajos planificados	Total trabajos presentados	Total publicaciones	Total resultados	Total registros	Total premios	Proyecciones de Maest.	Porciento trab. pres. de plan.
All Dim tiempo.Tiempos	61	52	36	11	3	14	108	85
2010							2	
2011							2	
2012							22	
2013							25	
2014	61	52	36	11	3	14	22	85
2015							19	
2016							16	

Figura 26: Navegación por los datos del cubo OLAP “cub_cti_centro”

En la figura 27 se muestra el resultado de la representación, mediante un gráfico de barras verticales, de la navegación realizada anteriormente.

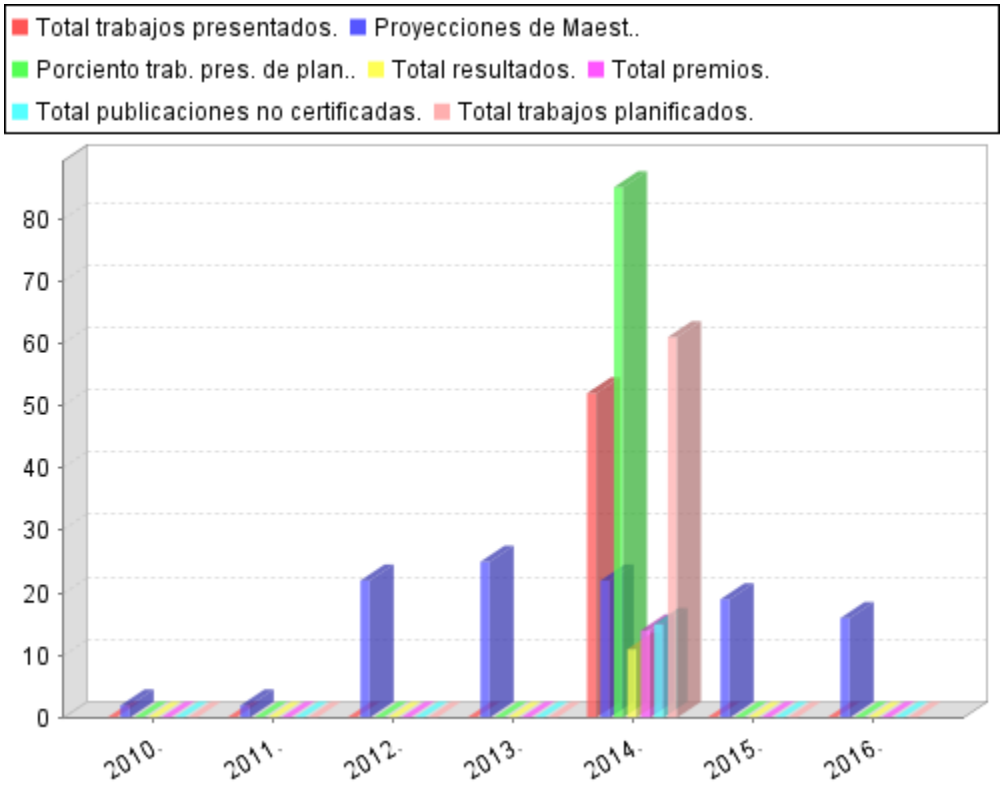


Figura 27: Gráfica con barras verticales de la navegación del cubo OLAP “cub_cti_centro”.

Lograda una correcta generación de la tabla de datos mostrada en la figura 26 y de la gráfica representada en la figura 27, se puede concluir que la prueba fue satisfactoria, ya que para la obtención de este resultado, es indispensable que el proceso ETL y la construcción de los cubos OLAP se realicen correctamente.

3.3.1.2 Pruebas de calidad de datos

Para la realización de la prueba de calidad de datos del MD, se debe tener en cuenta que en el mismo no existan valores nulos o vacíos, ya que dichos datos deben aportarle información al cliente para la realización de análisis posteriores. Con el objetivo de validar lo planteado anteriormente, a pesar de las validaciones realizadas durante el proceso de ETL, se realizaron consultas SQL a cada una de las tablas dimensiones y hechos del MD, con el fin de verificar que en las mismas no persistan campos vacíos o nulos. Dicho resultado se comparó con el índice de valores vacíos o nulos presentes en la fuente de datos, y los resultados arrojados demuestran que no existe este tipo de valores en el MD, ver figura 28.

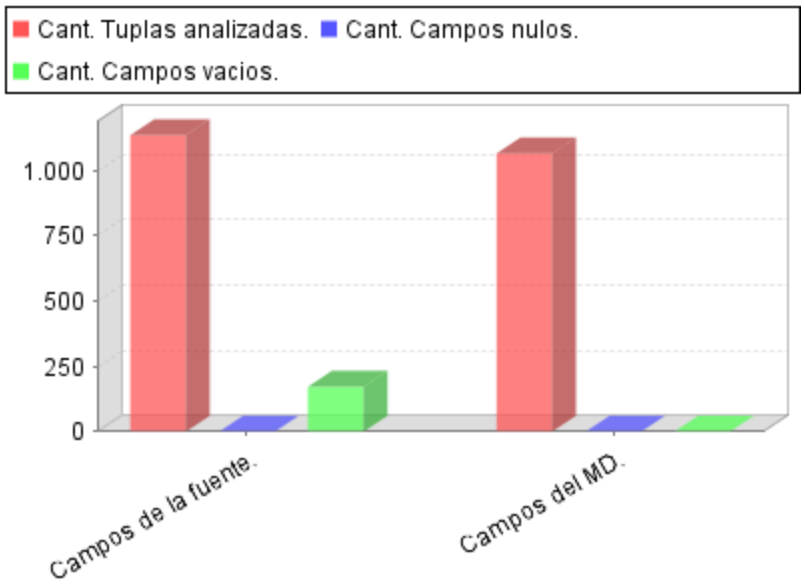


Figura 28: Comparación de valores vacíos y nulos entre la fuente y el MD.

3.3.1.3 Pruebas de confiabilidad de los datos

Concluido el proceso ETL es necesario verificar cuán confiable son los datos que persisten en el MD. Para esto se realizan pruebas mediante la confrontación entre los datos contenidos en la fuente origen y los del MD.

Con el objetivo de probar lo planteado anteriormente, se realiza una consulta a la información

almacenada en el MD y a la contenida en los documentos Excel entregados por el cliente. Esta consulta es realizada con el objetivo de conocer la cantidad de trabajos planificados en el año 2014 teniendo en cuenta cada una de las áreas que integran el centro. A continuación se muestra la información obtenida en forma textual:

En el año 2014 CEIGE planificó un total de 61 trabajos, de ellos 11, 7, 3, 1, 4, 8, 4, 16, 7 pertenecientes a las áreas ADUANA, DESPROD, DIR, FOR, INV, SOLEM, SOLFIN, SUBPROD, TEC respectivamente.

La figura 29 corresponde a la información obtenida del MD.

		Medidas
Nombre area	Tiempo	Total trabajos planificados
All Dim area.Nombre areas	2014	61
ADUANA	2014	11
DESPROD	2014	7
DIR	2014	3
FOR	2014	1
INV	2014	4
SOLEM	2014	8
SOLFIN	2014	4
SUBPROD	2014	16
TEC	2014	7

Figura 29: **Navegación por los datos del cubo OLAP "cub_cti_area".**

La tabla número 5 muestra la información correspondiente al documento Excel entregado por el cliente.

Áreas		ADUANA	DESPROD	DIR	FOR	INV	SOLEM	SOLFIN	SUBPROD	TEC
Total	61	11	7	3	1	4	8	4	16	7

Tabla 5: Reporte de los trabajos realizados en CEIGE en el año 2014.

Como se pudo observar los resultados arrojados por el análisis entre los datos orígenes, expuestos en la tabla 5, coinciden con los persistentes en el MD mostrados en la figura 29. Por

lo que se puede afirmar que el proceso ETL se realizó de forma correcta, ya que para su desarrollo y posterior análisis de los datos del cubo "cub_cti_area", es necesario haber realizado correctamente el poblado de cada una de las dimensiones y de los cubos de hechos.

3.3.2 Pruebas para validar el rendimiento del Mercado de Datos

3.3.2.1 Pruebas de rendimiento

Uno de los requisitos planteados por el cliente fue que la solución permitiera la conexión y consulta paralela a veinte usuarios y que el tiempo de respuesta no excediera de los 10 segundos. Con el objetivo de verificar el cumplimiento de este requisito se utilizó la herramienta Apache JMeter la cual permite analizar detalladamente el comportamiento del entorno montado en diferentes escenarios, y comprobar la correcta ejecución del MD según los requerimientos solicitados por el cliente.

La prueba se realizó bajo las siguientes condiciones:

- Herramienta a probar: Mercado de datos CTI para CEIGE.
- Tipo de operación a realizar: Consulta SQL.
- Hardware del Servidor de aplicaciones: 1.6 GHz de velocidad del microprocesador, 3 GB de memoria RAM y 160 GB como capacidad de disco duro.
- Sistema Operativo: Windows 7 Ultimate Service Pack 1.

La prueba se configuró para un total de 20 conexiones realizando una consulta a la semejanza de las que se efectuarán al MD en el entorno real de trabajo: select sum(ttp) as total from cub_hechos_cti_area group by id_tiempo order by id_tiempo, esta prueba arrojó los resultados que se pueden apreciar en la figura 30.

Label	# Muestras	Media	Mediana	Linea de 90%	Mín	Máx	% Error	Rendimiento
Petición JDBC	20	0	1	2	0	3	100,00%	3,4/sec
TOTAL	20	0	1	2	0	3	100,00%	3,4/sec

Figura 30: Prueba de rendimiento de una consulta al MD usando JMeter.

3.4 Conclusiones del capítulo

La herramienta seleccionada para el proceso ETL, permitió el correcto poblado de las tablas dimensiones y hechos, y la automatización de este proceso en el MD. Las pruebas de calidad de datos, posibilitó comprobar que en el MD no existen valores vacíos o nulos, comprobar el correcto funcionamiento del sistema y la confiabilidad de sus datos. La prueba de rendimiento,

permitió satisfacer los requisitos solicitados por el cliente en cuanto a tiempo de respuesta del MD.

CONCLUSIONES GENERALES

Con el fin de llevar a cabo la construcción de un MD para la obtención de los indicadores de CTI manejados por el CEIGE, facilitando el control y análisis por parte del centro, se cumplieron los objetivos trazados:

- La elaboración de la fundamentación teórica, permitió la selección de las herramientas y metodología adecuada para la construcción del MD.
- El análisis de los documentos otorgados por el cliente posibilitó la identificación de un total de treinta cuatro preguntas, facilitando la obtención de los principales requisitos de información del cliente.
- La elaboración de los modelos conceptuales y el modelo de datos del MD, permitió verificar que el diseño obtenido se ajusta a las necesidades del cliente.
- La herramienta seleccionada para la realización del proceso ETL, conllevó a la exitosa población de las tablas dimensiones y hechos del MD.
- Las pruebas realizadas al sistema, posibilitó verificar que las necesidades de información y requisitos solicitados por el cliente fueran cumplidas en su totalidad.

RECOMENDACIONES

Después de haber concluido el presente trabajo se recomienda:

- Integrar la solución al Sistema de Gestión de la Información de CTI del CEIGE.
- Incorporar la implementación de reglas de seguridad al MD, a partir de la especificación de permisos de acceso y modificación sobre la solución.

REFERENCIAS BIBLIOGRÁFICAS

1. **Carrasco, Francisco.** CIO América Latina. *CIO América Latina.* [En línea] CIO, 11 de Diciembre de 2012. [Citado el: 10 de Diciembre de 2013.] http://www.cioal.com/.

2. **Sallan, Rita, y otros, y otros.** Magic Quadrant for Business Intelligence Platforms. *Gartner.* [En línea] 23 de Enero de 2010. [Citado el: 19 de Octubre de 2013.] http://www.prostrategy-colman.ie/assets/files/downloads/Gartner_Magic_Quadrant_for_Business_Intelligence_Platforms_2010.docx.

3. *Repositorio de Integracion de Indicadores Claves de Rendimiento.* **Gil Riaño, Leandro Miguel, Triana Rodriguez, Anaelys y Ramirez Alvarez, Alexis.** 1, La Habana : Revista Estudiantil Nacional de Ingeniería y Arquitectura, 2011, Vol. 3. 2307-471X.

4. **Kimball, Ralph.** *The data warehouse toolkit: the complete guide to dimensional modeling.* 1998.

5. **Valdés, Dr. Juan Vela.** *Resolución No.128.* La Habana : s.n., 2006.

6. **Postgrado, Vicedecano Investigación.** *Estrategia de Superación.* La Habana : s.n., 2012.

7. **Ponniah, Paulraj.** *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals.* new York : Wiley-Interscience, 2001.

8. **Inmon, W. H.** *Building the data warehouse.* New York : John Wiley & Sons, 1996.

9. **López Gaona, Dr. Amparo.** Facultad Ciencias UNAM. *Facultad Ciencias UNAM.* [En línea] Febrero de 2012. [Citado el: 26 de Noviembre de 2013.] http://hp.fciencias.unam.mx/~alg/bd/introduccion.pdf.

10. **Pérez López, Yadini.** *Mercado de Datos para la gestión de la información sobre operaciones con tarjetas bancarias en el Banco Metropolitano.* Habana : s.n., 2013.

11. **Martínez Ruiz, Tomás, Navarro Quevedo, Almudena y Visuete, Michael.** *Modelos Avanzados de Bases de Datas.* Quito : Universidad de Castilla-La Mancha Escuela Superior de Informática, 2006.

12. **Alonso Llombart, Óscar.** *BI: Inteligencia aplicada al negocio.* s.l. : DAA Contenidos Digitales, S.L, 2003.

13. **Ortiz Sierra, Julio Ernesto.** *Diseño e Implementación de un Mercado de Datos para la Oficina Nacional de Estadísticas.* Habana : s.n., 2009.

14. **Hechavarría, Yunior Ricardo.** *Análisis, Diseño e Implementación del Mercado de Datos indicadores relacionados con la ciencia e innovación tecnológica para la Oficina Nacional de Estadísticas.* Habana : s.n., 2010.

15. **XETID.** *Metodología XETID para el desarrollo de un Almacen de Datos.* Habana : s.n., 2012.

16. **Tamayo, Marysol y Moreno, Francisco Javier.** Análisis del modelo de almacenamiento MOLAP frente al modelo de almacenamiento ROLAP. *Scielo.* [En línea] Septiembre de 2006. [Citado el: 25 de Noviembre de 2013.] http://www.scielo.org.co/scielo.php?pid=S0120-56092006000300016&script=sci_arttext.

17. **López Beltrán, Carlos Patricio.** *Análisis, Diseño e Implementación de un Data Mart para la Dirección Financiera y recursos Humanos de la Escuela Politécnica del Ejército para una toma de decisión efectiva.* Sangolquí : s.n., 2007.

18. **Microsoft.** Microsoft System Center. *Microsoft System Center.* [En línea] Microsoft, Diciembre de 2012. [Citado el: 09 de Diciembre de 2013.] http://technet.microsoft.com/es-es/library/hh524258.aspx.

19. **Ibarzabal, J.** *Estrategia de reporting.* 2003.

20. **Inc., SyBase.** SyBase. *SyBase.* [En línea] SyBase Inc., 2014. [Citado el: 04 de Febrero de 2014.] http://www.sybase.es/.

21. **CONACYT.** SIICYT. *Sistema Integrado de Información sobre Investigación Científica, Desarrollo Tecnológico e Innovación.* [En línea] CONACYT, 04 de Diciembre de 2003. [Citado el: 05 de Diciembre de 2013.] http://www.siicyt.gob.mx/siicyt/cms/paginas/AcercadelSIICYT.jsp?pSel=.

22. **D.C, Bogotá.** Colciencias. Departamento Administrativo de Ciencia, Tecnología e Innovación. *Colciencias. Departamento Administrativo de Ciencia, Tecnología e Innovación.* [En línea] 21 de Diciembre de 2012. [Citado el: 04 de Febrero de 2014.] http://www.colciencias.gov.co.

23. *SNCTI Relación y Perspectivas.* **Bermúdez, M.Sc. Giovanni.** 2011.

24. **SIGENU.** Sistema de Gestión de la Nueva Universidad. *Sistema de Gestión de la Nueva Universidad.* [En línea] 2013. [Citado el: 18 de Febrero de 2014.] http://sigenu.mes.edu.cu:8080/dmmes/pages/login/loginUser.faces.

25. **Roque Lavandero, Luis Ángel y Cabrales Camino, Jorge César.** *Sistema de Gestión de la información.* Habana : s.n., 2013.

26. **Curto, J.** CIF vs MD: Dos enfoques clásicos en el diseño de la arquitectura de una Data Warehouse. *CIF vs MD: Dos enfoques clásicos en el diseño de la arquitectura de una Data Warehouse.* [En línea] Lunes 12 de Enero de 2009. [Citado el: Martes 03 de Diciembre de 2013.] http://bi-businessintelligence.blogspot.com/2009/01/cif-vs-md-dos-enfoques-clsicos-en-el.html.

27. **Bernabeu, Ricardo Dario.** *Hefesto: Metodología para la Construcción de un Almacén de*

Datos. Córdoba, Argentina : s.n., 2010.

28. Visual Paradigm for UML. *Visual Paradigm for UML.* [En línea] 16 de Diciembre de 2013. [Citado el: 6 de Febrero de 2014.] http://www.visual-paradigm.com/product/vpuml/.

29. **González Ochoa, Darián.** *Diseño e Implementación de un Almacén de Datos Operacionales para la Corporación CIMEX.* Habana : s.n., 2009.

30. **Corporation, Pentaho.** *Pentaho BI Suite Enterprise Edition.* 2008.

31. **Carina Roldán, María.** *Pentaho 3.2 Data Integration Beginner's Guide.* s.l. : Packt Publishing Ltd, 2010.