Universidad de las Ciencias Informáticas

Facultad 2

Mercado de Datos para el primer año de la carrera Ingeniería en Ciencias Informáticas.

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

Autores:

Yenisey Silverio Jover

Yenlis González Galá

Tutores:

MSc. Yusnier Reyes Dixson

Ing. Reydel Capote Coipel

Co-tutor:

Ing. Daynier Ramiro García Prats

La Habana, 2017

"Año 59 de la Revolución"

...si supiese qué es lo que estoy haciendo, no lo llamaría INVESTIGACIÓN...

Albert Einstein

Declaración de autoría

Declaramos que somos los únicos autores del trabajo titulado: Mercado de Datos educativo para el primer año de la carrera Ingeniería en Ciencias Informáticas y autorizamos a la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmamos la presente a los _____días del mes de _____del año_____.

_____ _____

Yenlis González Galá **Yenisey Silverio Jover**

_____ _____

MSc. Yusnier Reyes Dixson **Ing. Reydel Capote Coipel**

Agradecimientos

Quiero agradecer a mis padres por haber confiado en mí desde el inicio e impulsarme a continuar con mis estudios.

A mi hermano Yeisel por estar a mi lado en cada momento y darme su apoyo.

A mis padrinos, por ser los mejores del mundo y preocuparse por mi educación y bienestar.

A mi primo Ramón (papo), que por circunstancias de la vida no está presente físicamente, pero que estará siempre en mi corazón.

A Lizsandra por ser mi compañera, mi amiga, mi hermana y estar presente en cada momento importante.

A mi abuela querida, a mis primos Dianelys, Noemi, Yesi, Yurien, Yuliet y Osmay, a mis tíos José, Luisa, Ramón, Teresa, Margui, Sonia, y todos los demás miembros de la familia por estar a mi lado.

A todos los amigos que he conocido en esta maravillosa universidad, en especial a mis compañeras de apartamento: Yuliet, Yanira, Daynela, Sandra y mi compañera de tesis, Yenlis.

A mis compañeros de aula: Yeniel, Luis Angel, Yosiel, Julién, Bárbaro y a todos los demás que no he mencionado.

A todos los profesores que de una forma u otra contribuyeron a mi formación como profesional.

A mis tutores por ser compañeros y guías en este bello trabajo que hemos realizado juntos.

Por último y no menos importante un agradecimiento para mí, por el esfuerzo y dedicación para cumplir este sueño.

Yenisey

Agradecimientos

Quiero agradecer ante todo a mis padres por su educación, sacrificio y apoyo en todos los momentos de mi vida, porque siempre han estado a mi lado, por ayudarme a cumplir mis metas y a confiar en mí.

Por haberme prestado su total dedicación y su constante preocupación lo que me ha motivado a lograr lo que he logrado y a ser la persona que soy hoy en día.

Les agradezco a mi hermano y a mi sobrinito que adoro, por ser especial y marco en mi vida esa primera etapa de mi universidad.

A toda mi familia.

A mis amigos, que saben quiénes son, porque me han apoyado en alguna etapa de mi vida y me han enseñado el significado de la amistad.

A todos los amigos y compañeros con los que he compartido durante estos cinco años, a los que hoy se encuentran cerca de mí y también a los que no, con todos he vivido momentos inolvidables y todos me han ayudado de una forma u otra, siempre los recordaré.

A todas las nuevas amistades que me ha regalado esta Universidad.

A todos mis profesores, por contribuir con mi formación personal y profesional.

A mis tutores,

A mi compañera de tesis, tenía que mencionarte.

En especial, a alguien muy, pero muy importante para mí, alguien que en muy poco tiempo se ha colado en lo más profundo de mi mente y de mi cuerpo, alguien que me ha hecho llorar, reír, pensar,

meditar, aprender y disfrutar momentos y experiencias que nunca pensé. Le agradezco a mi novio, que ha estado a mi lado en esta etapa final dándome todo su apoyo, preocupándose por mí. Te quiero infinitamente.

A todos los que ayudaron a que esta tesis se cumpliera.

A los que se tomaron la molestia de venir y a los que no….

Yenlis

Dedicatoria

De Yenisey:

A mis padres: Magaly y Alejandro por su amor, apoyo y sacrificio.

A mi hermano Yeisel y mi abuela Ernestina por la confianza.

A mi primo Ramón, Dianelys y mi tío Luis Miguel.

De Yenlis:

A mis padres, mi hermano, mi sobrino, mi novio.

A toda mi familia.

Resumen

Durante el proceso de enseñanza aprendizaje, la tecnología juega un papel preponderante, destacándose en la Universidad de las Ciencias Informáticas (UCI) el uso de Akademos[1] y el Centro de Calidad e Innovación de la Educación (CICE). Estos sistemas les permiten a los profesores y directivos retroalimentarse del comportamiento de sus estudiantes y de esta forma implementar sus propias estrategias educativas. Actualmente la información se encuentra dispersa en varias fuentes de datos, lo que imposibilita el análisis integral de los estudiantes. Por esta razón el propósito de esta investigación es desarrollar un Mercado de Datos que permita integrar toda la información de los estudiantes en una única base de datos para facilitar la toma de decisiones de directivos y profesores. Para la construcción de la solución se seleccionaron herramientas y tecnologías libres fundamentadas en la migración de la sociedad cubana a este tipo de ambiente de desarrollo. Se seleccionó la metodología Hefesto para guiar el proceso de desarrollo del mercado. Igualmente se definieron los procesos de extracción, transformación y carga (proceso ETL) de los datos correspondientes al modelo propuesto y se realizaron pruebas para validar la calidad de la solución.

Palabras claves: mercado de datos, proceso ETL, toma de decisión

[1] Sistema de gestión universitaria de la Universidad de las Ciencias Informáticas (UCI)

Índice de contenido

Índice de Figuras

Índice de Tablas

Introducción

Las universidades son organizaciones con una importante responsabilidad social, ya que en ellas se genera y transmite gran parte del conocimiento que apoya el desarrollo económico de cualquier sociedad. La trascendencia del encargo social de las universidades y el alto costo de la enseñanza en dichas instituciones, sobre todo las relacionadas con las ramas tecnológicas, demandan eficiencia, eficacia y calidad en los procesos que en éstas se desarrollan. Para conseguir este propósito, la gestión de estos debe ser efectiva, basada en el uso de las tecnologías y con métodos sujetos a constante perfeccionamiento. Con el objetivo de apoyar y mejorar la gestión, algunos investigadores proponen que las instituciones de la educación superior utilicen de forma organizada la información y el conocimiento. Es con este fin que la gestión de la información y el conocimiento deviene como herramienta importante en la dirección de las universidades (Luan 2002; Heredia 2011).

La formación es un proceso principal en la universidad que a su vez es complejo, debido a la gran variabilidad en las características de los estudiantes y a la variedad de condiciones que confluyen en el mismo (J. H. Rico and Hernández 2012). Los principales trabajadores que contribuyen a la formación (profesores y directivos) son conocidos como trabajadores del conocimiento, entendidos como aquellos que usan la información como su principal entrada, transformándola a través de su conocimiento para tomar decisiones y desarrollar acciones (Cuesta 2013).

Para que los profesores y directivos sean más productivos, deben poseer un amplio conocimiento, no solo de las materias que enseñan, sino también de las condiciones bajo las cuales se desarrolla el proceso docente, para poder elaborar decisiones acertadas acerca de los métodos y procedimientos a utilizar para alcanzar los diferentes resultados (Hernández 2013). El dominio cognitivo que se desea que estos trabajadores tengan sobre la formación docente requiere una correcta gestión de la información y el conocimiento, para facilitar la identificación, captación, procesamiento y diseminación de datos adecuados para la obtención de un modelo que facilite la toma de decisiones así como la concepción de estrategias orientadoras (J. Heredia Rico and Rodríguez Hernández 2012).

El aprendizaje se produce a partir de un conjunto de procesos interrelacionados los cuales deben gestionarse basándose en mediciones objetivas que reflejen hechos. Sin embargo, frente a la dinámica de desarrollo actual de la ingeniería de alto nivel, resulta insuficiente; se necesita además lograr el acercamiento progresivo de los procesos a su nivel óptimo, de manera tal que se minimice el riesgo de no alcanzar la mejor calidad posible en los resultados. Dicho fenómeno se traduce en la urgencia de

perfeccionar el proceso docente, de manera tal que se logre guiar de forma óptima el trabajo del profesor en su labor de orientación y ayuda a los estudiantes, brindándole informaciones más completas y oportunas sobre las características de estos, que les faciliten su labor y potencien la efectividad de la misma; y permitiéndole controlar todo el proceso de formación (J. H. Rico and Hernández 2012).

En estas condiciones, los directivos del proceso deben ser capaces de usar la información que aparece durante su desarrollo, integrarla, formular esquemas para la acción y ser capaces de reunir el máximo de certidumbres para confrontar la incertidumbre. Debe crearse una inteligencia organizacional que posibilite la identificación, captación y procesamiento de datos adecuados para la obtención de un modelo del proceso que facilite la toma de decisiones y la concepción de estrategias orientadoras (J. J. H. Rico, Hernández, and Alonso 2012).

Actualmente los principales objetivos de los gestores de las universidades están dirigidos a mejorar el rendimiento de la gestión interna (disminuyendo gastos y optimizando procesos) e incrementar la calidad docente e investigadora de la universidad. Los gestores universitarios también necesitan sistemas analíticos para conocer de forma fiable qué ha sucedido, está sucediendo o puede suceder en la institución. Estas preguntas pueden hacerse a distinto nivel de granularidad: a nivel global, en un departamento, en un programa de formación, en una asignatura (o conjunto de ellas) o en sus estudiantes (Hormigo and Caralt 2014).

Teniendo en cuenta los cambios que en materia tecnológica y organizacional se evidencian en la actualidad y su influencia en Cuba, urge formar un profesional cada vez mejor preparado y por consiguiente, con un mayor nivel de competitividad; así podrá afrontar con más efectividad los retos que debe asumir en su vida laboral en las circunstancias actuales y futuras. En este sentido, la Universidad de las Ciencias Informáticas tiene como misión convertirse en una: *"…Universidad innovadora de excelencia científica, académica y productiva, que forma de manera continua profesionales integrales comprometidos con la Patria, soporte de la informatización del país y la competitividad internacional de la industria cubana del software"* (UCI 2013).

La UCI cuenta con un entorno tecnológico amplio, con distintas soluciones que favorecen el desarrollo de nuevas estrategias de análisis de datos, sin embargo, no se ha aprovechado al máximo todo el caudal de información. Leyet (2012), una investigadora del Centro de Calidad e Innovación de la Educación de la UCI expresó: *"hoy existen muchos retos en este campo, hemos acumulado muchos datos sobre el estudiante, sin embargo no los utilizamos para mejorar el proceso de enseñanza y aprendizaje"*

La universidad maneja un extenso volumen de información debido a que constantemente se generan reportes de las diferentes evaluaciones y datos personales de los estudiantes. Entre los datos personales generados se encuentran: la tenencia o no de hijos, el estado de salud, la localidad en que vive, entre otros. La recogida de los datos personales es individual y se hace a inicios de la carrera, mientras el registro de evaluaciones está en constante cambio.

Dentro de sus facultades se encuentra la Facultad Introductoria de Ciencias Informáticas (FICI), donde solo se encuentran estudiantes del primer año. A estos estudiantes se les da un tratamiento diferenciado por ser este año el que determina la eficiencia vertical de la universidad y es donde más estudiantes abandonan la carrera, repiten o suspenden.

A través de entrevistas no estructuradas realizadas a directivos y profesores de la facultad, se comprobó que el proceso de toma de decisiones generalmente se realiza de forma empírica con escaso uso de la información que se genera en el proceso de formación. Esto se debe a que la información que se tiene en cuenta para este proceso no se encuentra centralizada, sino en diferentes fuentes de datos como: Akademos, informes docentes y diagnósticos iniciales. Esta situación ha provocado que dicho proceso cada día sea más complejo desde el punto de vista de la gestión del conocimiento y se haga más difícil por diversas insuficiencias detectadas a través del análisis y de las ya mencionadas entrevistas. Algunas de estas insuficiencias son:

- ✓ La diversidad de fuentes de procedencia de los datos y la información académica de entrada a los procesos, dificulta la síntesis de la misma.
- ✓ No se puede acceder de forma inmediata a datos históricos de indicadores académicos de estudiantes y asignaturas que son utilizados en los análisis de cortes evaluativos, situación docente del año, investigaciones y valoraciones, atentando contra la necesaria visión histórica del proceso y sus principales indicadores.
- ✓ Los datos que se obtienen en el proceso de matrícula y los que se captan en los diagnósticos iniciales no son utilizados sistemáticamente para la toma de decisiones que se desarrolla en el proceso de formación de los estudiantes.
- ✓ No se cuenta con mecanismos oportunos que sean capaces de responder preguntas como: ¿qué está pasando?, ¿qué ha pasado? o ¿qué puede suceder? en el proceso formativo, lo que elimina el factor proactivo en el desarrollo del proceso.

En correspondencia con lo antes expuesto se plantea como **problema de investigación** La variedad de origen de la información académica y personal de los estudiantes de la FICI provoca limitaciones en el proceso de toma de decisiones de los profesores y directivos, por lo que se define como **objeto de estudio:** Proceso de desarrollo de los mercados de datos.

Teniendo en cuenta el problema planteado se define como **objetivo general:** Desarrollar un Mercado de Datos que centralice la información académica y personal de los estudiantes de la FICI para disminuir las limitaciones en el proceso de toma de decisiones de los profesores y directivos, enmarcado en el **campo de acción:** Proceso de desarrollo de un mercado de datos para disminuir las limitaciones en el proceso de toma de decisiones de los profesores y directivos.

Para dar solución a la situación antes planteada, se definen las siguientes **tareas de investigación**:

- ✓ Estudio de sistemas similares a la propuesta de solución, tanto a nivel nacional como internacional, para conocer aspectos regulares en la elaboración del Mercado de Datos.
- ✓ Estudio y selección de las diferentes herramientas y metodologías utilizadas para el diseño y posterior desarrollo de la solución.
- ✓ Análisis de los elementos correspondientes al diseño de mercados de datos para guiar la implementación de la solución.
- ✓ Estudio de las pruebas del modelo en V del centro DATEC (Centro de Tecnologías de Gestión de Datos) para validar la propuesta de solución.

Los métodos utilizados para la presente investigación son los siguientes:

Teóricos:

- ✓ **Analítico-Sintético:** Permitió seleccionar, de las metodologías existentes para el desarrollo de almacenes de datos, la más apropiada según las características de la FICI. También se utilizó para estudiar las diferentes herramientas y tecnologías necesarias para el desarrollo del almacén de datos.
- ✓ **Modelación:** Se utilizó para el diseño de diagramas facilitando la comprensión de los procesos desarrollados para el diseño de un almacén de datos.

Empíricos:

- **Entrevista:** Se utilizó para determinar las necesidades y requerimientos de los profesores y dirigentes de la FICI para conocer para la construcción del mercado de datos.

La presente investigación cuenta con la siguiente estructura:

Capítulo 1: Fundamentación Teórica

En este capítulo se realiza una descripción de los elementos más importantes del Mercado de Datos. Se exponen conceptos, características, ventajas, desventajas, así como las herramientas, metodologías y tecnologías existentes para el desarrollo. También se hace un análisis de los principales procesos que se realizan para la construcción de un Mercado de Datos y de los sistemas existentes en Cuba.

Capítulo 2: Diseño e implementación del Mercado de Datos

Se explican las diferentes fases de la metodología de desarrollo de almacenes de datos seleccionada para guiar el proceso. Se describe el proceso de Extracción-Transformación-Carga (ETL), en el que se limpian, transforman y cargan los datos para el posterior diseño del cubo multidimensional. Se identifican las necesidades del cliente mediante el levantamiento de requisitos, se definen las reglas del negocio y el modelado de los datos a través de sus elementos principales: dimensiones, hechos y medidas.

Capítulo 3: Visualización y Validación

En este capítulo se explica cómo se desarrolló el proceso de creación de los cubos de información, cómo son publicados y como se visualizan para permitir al usuario ver reportes y vistas de análisis mediante tablas y gráficos. Además, se explican las pruebas realizadas al Mercado de Datos para verificar el cumplimiento de las necesidades del cliente.

Capítulo 1: Fundamentación Teórica de los almacenes de datos

Este capítulo está referido a los fundamentos teóricos de los almacenes de datos. Para ello se realiza un estudio del estado del arte del tema en cuestión a nivel internacional. También se hace un estudio de los Mercados de Datos exponiendo sus elementos fundamentales como su definición, características, ventajas y desventajas. Además, se describen las principales características de la metodología, tecnologías y herramientas utilizadas para el desarrollo de la aplicación que dará solución al problema planteado.

1.1. Almacén de datos o Data Warehouse (AD)

Diversos han sido los especialistas que a lo largo de los años se dedicaron al estudio de los almacenes de datos y han dado varias definiciones sobre esta potente tecnología para la generación de reportes. Aquí se muestran dos de estas definiciones.

Según W. H. Inmon un Data Warehouse es una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia". (Inmon, 2005)

Según Ralph Kimbal un Data Warehouse es: "Una copia de los datos transaccionales, específicamente estructurados para consultas y análisis". (Kimball & Margy, 2002)

Debido a que W. H. Inmon, es reconocido mundialmente como el padre de los almacenes de datos (Bernabeu, 2007), los autores de esta investigación van a tomar partido por la definición y características de esta herramienta ofrecidas por él.

Las principales características de los almacenes de datos son: (Inmon, 2005)

- ✓ **Orientada a temas**: La información se clasifica en base a los aspectos que son de interés para la empresa, siendo así, los datos tomados están en contraste con los clásicos procesos orientados a las aplicaciones.
- ✓ **Integrado**: Implica que todos los datos de diversas fuentes que son producidos por distintos departamentos, secciones y aplicaciones, tanto internas como externas, deben ser consolidados en una instancia antes de ser agregados al AD.

- ✓ **Variante en el tiempo**: Los datos son relativos a un período de tiempo y estos deben ser integrados periódicamente, los mismos son almacenados como fotos que se corresponden a un período de tiempo.

- ✓ **No volátil**: La información es útil para el análisis y la toma de decisiones solo cuando es estable. Los datos operacionales varían momento a momento, en cambio, los datos una vez que entran en el AD no cambian.

Entre las ventajas que proporciona el uso de esta herramienta se encuentran: (Bernabeu, 2007)

- ✓ Transforma datos orientados a las aplicaciones en información orientada a la toma de decisiones.
- ✓ Mejora la entrega de información, es decir, información completa, correcta, consistente, oportuna y accesible. Información que los usuarios necesitan, en el momento adecuado y en el formato apropiado.
- ✓ Aumento de la competitividad de los encargados de tomar decisiones.
- ✓ Permite la toma de decisiones estratégicas y tácticas.

Como desventajas se tiene: (Bernabeu, 2007)

- ✓ Incremento continuo de los requerimientos del usuario.

- ✓ Requiere una gran inversión, debido a que su correcta construcción no es tarea sencilla y consume muchos recursos, además, su misma implementación implica desde la adquisición de herramientas de consulta y análisis, hasta la capacitación de los usuarios.

1.2. Mercado de datos (MD)

El Mercado de Datos es una versión especial del AD. Es un subconjunto de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones. Los datos existentes en este contexto pueden ser agrupados, explorados y propagados de múltiples formas para que diversos grupos de usuarios realicen la explotación de los mismos de la forma más conveniente según sus necesidades. Es consultado mediante herramientas OLAP (*On line AnalyticalProcessing* - Procesamiento Analítico en Línea) que ofrecen una visión multidimensional de la información. Se puede decir que los MD son pequeños AD centrados en un tema o un área de negocio específico dentro de una organización.

Existen dos tipos de MD: los dependientes que obtienen sus datos del AD y los independientes que obtienen sus datos de fuentes externas. (Quilumba, 2013)

Dentro de las ventajas de aplicar el MD a un negocio, se han seleccionado las siguientes: (Bernabeu, 2007)

- ✓ Son simples de implementar.
- ✓ Conllevan poco tiempo de construcción y puesta en marcha.
- ✓ Permiten manejar información confidencial.
- ✓ Reflejan rápidamente sus beneficios y cualidades.
- ✓ Reducen la demanda del depósito de datos.

La información que se desea gestionar es la referente a los estudiantes del primer año de la carrera Ingeniería en Ciencias Informáticas, por lo que se hace necesario implementar un MD y no un AD, debido a que un MD cubre las necesidades de una determinada área dentro de la organización, y el costo de su uso es inferior por lo que se lleva menor tiempo para construirlo y ponerlo a funcionar; mientras que un AD cubre las necesidades de la organización en su conjunto, y el costo de su uso es mayor por lo que construirlo y ponerlo a funcionar llevaría más tiempo.

1.3. Soluciones similares

En la búsqueda de una respuesta a las necesidades de los profesores de la FICI, se estudiaron diferentes mercados de datos en el ámbito universitario. Este estudio se realizó con el objetivo de tomar experiencias en cuanto a su funcionamiento y a su vez analizar en qué medida dichos sistemas brindaban solución a la problemática planteada. A continuación, se muestran varios mercados de datos que facilitan la toma de decisiones en diversas universidades:

Diseño e implementación de un Mercado de Datos OLAP para el análisis gerencial académico, que será implementado en la unidad educativa "La Colina": Proyecto de fin de carrera que utiliza un Mercado de Datos en entorno universitario. Este sistema tiene información relacionada a los docentes y estudiantes de la institución educativa. Además, muestra el horario del profesor y la hora de las clases, las notas del grado, los profesores por cada materia, la provincia, las materias comunes que tienen los profesores, entre otros elementos. (Guisado Verdezoto, 2015)

Análisis y diseño de un Mercado de Datos para el seguimiento académico de alumnos en un entorno universitario: Proyecto de fin de carrera que utiliza un Mercado de Datos en entorno universitario. El sistema permite analizar las notas medias de los alumnos y de los diferentes cursos académicos, repasándolos por tipos

de asignaturas y agrupándolos por las diferentes titulaciones a las que pertenece el alumnado. Esta información proporciona el conocimiento indispensable para saber qué titulaciones y qué cursos de éstas, son las que peor nota media tienen y, por consiguiente, los cursos en los que el alumnado obtiene peores resultados. (Rodríguez Sanz, 2012)

Análisis, diseño e implementación de un Mercado de Datos académico usando tecnología de BI para la Facultad de Ingeniería, Ciencias Físicas y Matemática: Proyecto de fin de carrera que utiliza un Mercado de Datos en entorno universitario. Este sistema tiene información relacionada con los docentes y estudiantes de la facultad. Esta información está dividida en vistas: una llamada Carga_Horaria en la que se muestra información de los docentes, materia que imparte, la carga horaria que tienen, categoría y periodos en los que han impartido clases en la Facultad. La otra vista se llama Estudiante_Indicadores que tiene información personal de los estudiantes, de que institución vienen, fecha de inscripción, matrícula y además detalla el récord académico de los estudiantes de la facultad de todas las carreras. (Quilumba, 2013)

Valoración crítica

Los sistemas encontrados son insuficientes para suplir las exigencias del producto requerido por los profesores de la FICI. En ellos, se encuentran solamente fragmentos de los requisitos del sistema que se pretende elaborar.

Las características de estos sistemas, aunque con similitudes en algunos casos, no se corresponden con la totalidad de las deseadas para el sistema, que debe ser, una aplicación que facilite la toma de decisiones. Por tanto, sería muy costoso en tiempo y esfuerzo, modificar alguno de los sistemas existentes para adaptarlo a las necesidades de los profesores de la FICI.

Concluyendo que ninguno de los sistemas analizados presentas las características ideales para ser utilizado por los profesores de la FICI, se decide comenzar la implementación de un nuevo producto, un mercado de datos, que responda a los requisitos específicos del cliente.

1.4. Proceso de ETL

ETL - este término viene del inglés de las siglas Extract-Transform-Load que significan Extraer, Transformar y Cargar. ETL es el proceso que organiza el flujo de los datos entre diferentes sistemas en una organización y aporta los métodos y herramientas

necesarias para mover datos desde múltiples fuentes a un almacén de datos, reformatearlos, limpiarlos y cargarlos en otra base de datos, MD o bodega de datos. ETL forma parte de la inteligencia de negocios, también llamado “Gestión de los Datos” (Data Management). La idea es que una aplicación ETL lea los datos primarios de unas bases de datos de sistemas principales, realice transformación, validación, el proceso cualitativo, filtración y al final escriba datos en el almacén y en este momento los datos están disponibles para analizar por los usuarios. (Calderón Gómez, Díaz Mongui, & Ariza Nieves, 2015)

El primer paso de este proceso es la extracción que consiste en extraer los datos desde los diferentes sistemas fuentes y los deja listos para ser transformados. El segundo paso sería la transformación que convierte aquellos datos inconsistentes en un conjunto de datos compatibles y estandarizados que puedan ser cargados en el almacén de datos. Este paso tiene por objetivo llevar todos los datos a un único formato. El tercer paso es la carga de los datos para el almacén después de haber sido transformados. Los datos que se cargan definitivamente para el almacén deben tener la mejor calidad.

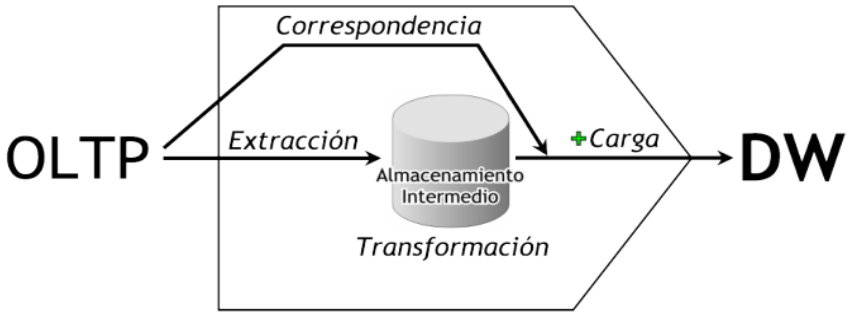


Figura 1: Proceso ETL

Este proceso es de vital importancia, pues permite a las organizaciones mover datos desde diversas fuentes, reformatearlos y cargarlos en otra base de datos que se denomina AD o MD en este caso. Mediante este proceso también se pueden analizar los datos.

1.5. Arquitectura de los Almacenes de datos

Los Almacenes de datos están compuestos por una fuente de datos y tres subsistemas propios de este tipo de sistema que son: el subsistema de integración, el subsistema de almacenamiento y el subsistema de visualización. A continuación, se explican cada una de sus componentes:

- ✓ Fuentes de datos: lugar de donde se extrae la información para poblar el AD.
- ✓ Subsistema de integración: encargado de realizar todos los procesos de ETL donde se extrae, se limpia y se integra toda la información almacenada en las fuentes de datos a través de transformaciones.
- ✓ Subsistema de almacenamiento: donde se guarda toda la información luego de haber sido transformada en el subsistema de integración.
- ✓ Subsistema de visualización: encargado de mostrar al cliente toda la información almacenada, a través de reportes y vistas de análisis permitiendo al cliente poder analizar la información procesada.

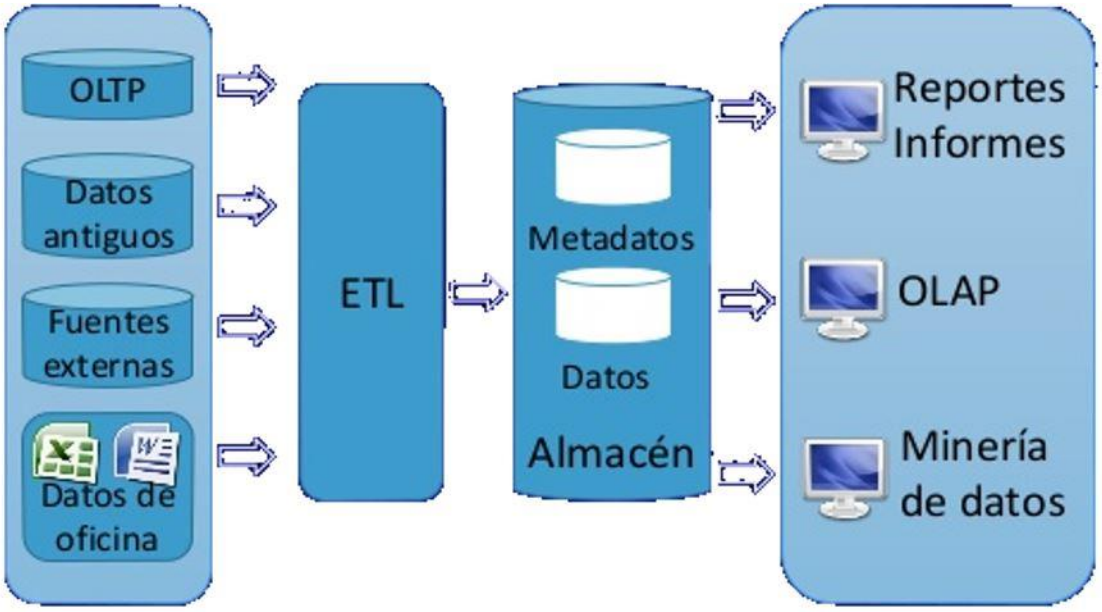


Figura 2: Arquitectura del Mercado de Datos

1.6. Modelado multidimensional

Modelo dimensional es el nombre que recibe la técnica utilizada especialmente para la construcción de MD. Esta presenta la información de una manera estándar, sencilla y sobre todo intuitiva para los usuarios, además permite el acceso a la información mucho más rápida por parte de los manejadores de bases de datos. Este modelo brinda una búsqueda rápida de los datos y la información va a almacenarse a través de tablas de dimensiones y hechos. (Sosa Bello & Salas Lóriga, 2013)

- ✓ Hechos: Es un concepto de interés primario para el proceso de toma de decisiones, corresponde a eventos que ocurren dinámicamente en el negocio

de la empresa y contiene los hechos, indicadores o medidas del negocio que se desean analizar. (Kimball & Margy, 2002)

- ✓ Dimensiones: Son objetos del negocio con los cuales se puede analizar la tendencia y el comportamiento del mismo. Las definiciones de las dimensiones se basan en políticas de la compañía, e indican la manera en que la organización interpreta o clasifica su información para segmentar el análisis facilitando la observación de los datos. (Kimball & Margy, 2002)

- ✓ Medidas: Son características cualitativas o cuantitativas de los objetos que se desean analizar en las empresas. Las medidas cuantitativas están dadas por valores o cifras porcentuales. Por ejemplo, la cantidad de estudiantes.

Las bases de datos dimensionales tienen tres variantes posibles de modelación, las cuales se mencionan a continuación:

- ✓ Esquema estrella.
- ✓ Esquema copo de nieve.
- ✓ Esquema constelación.

1.6.1. Esquema en estrella

Esquema en estrella: consta de una tabla de hechos central y de varias tablas de dimensiones relacionadas a esta, a través de sus respectivas claves. Las tablas de dimensiones solo se relacionan con la tabla de hechos y no existen relaciones entre dimensiones. Las tablas de dimensiones tendrán siempre una clave primaria simple, mientras que, en la tabla de hechos, la clave principal estará compuesta por las claves principales de las tablas dimensionales o una propia (Bernabeu, 2007). En la siguiente figura se puede apreciar un esquema en estrella estándar:

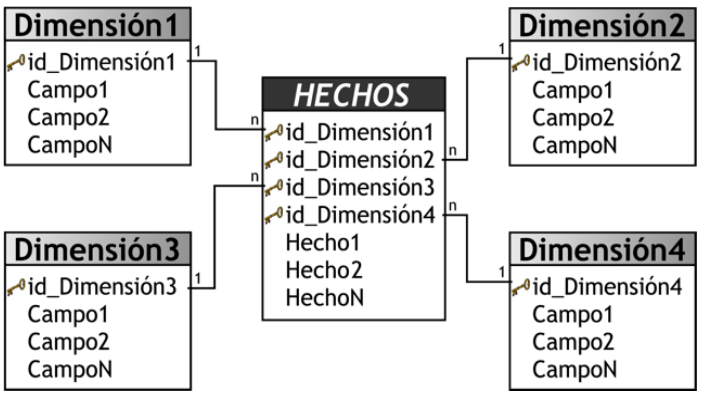


Figura 3: Esquema de Estrella

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.

1.6.2. Esquema copo de nieve

Esquema copo de nieve: este esquema representa una extensión del modelo en estrella cuando las dimensiones se organizan en jerarquías de dimensiones, es decir, 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 (Bernabeu, 2007). A continuación, se muestra un esquema copo de nieve:

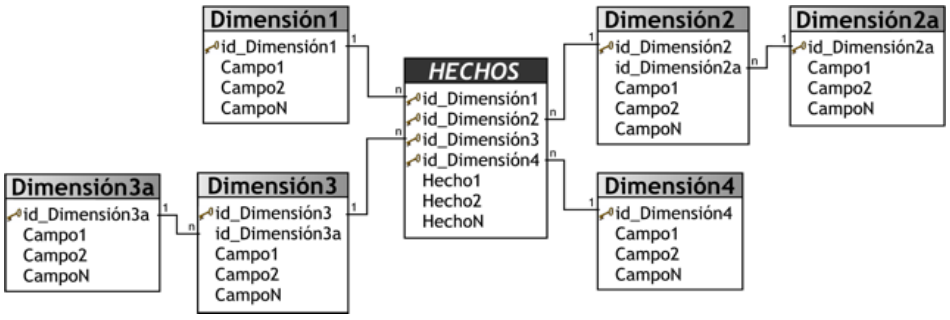


Figura 4: Esquema Copo de Nieve

1.6.3. Esquema constelación

Esquema constelación: este modelo está compuesto por una serie de esquemas en estrella, es decir, está formado por una tabla de hechos principal (“HECHOS_A”) y por una o más tablas de hechos auxiliares (“HECHOS_B”) que están relacionadas con sus respectivas tablas de dimensiones, vinculándose las tablas de hechos auxiliares con algunas dimensiones asignadas a la tabla de hecho principal y también con nuevas tablas de dimensiones. (Bernabeu, 2007). A continuación, se muestra un esquema constelación:

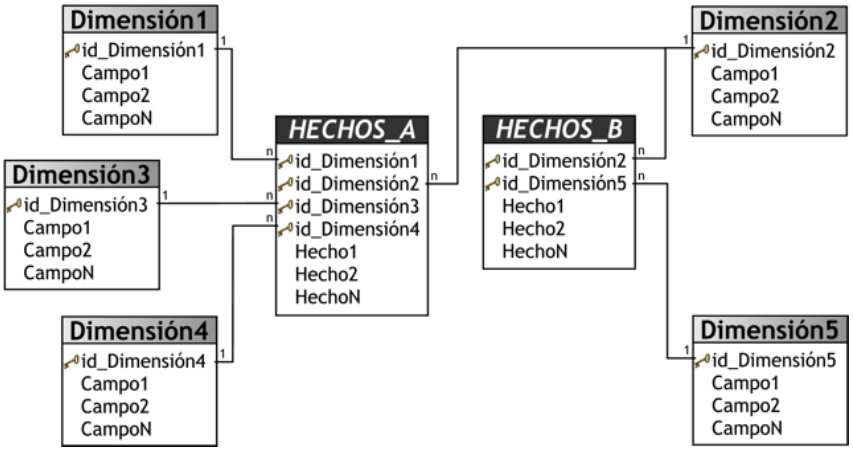


Figura 5: Esquema Constelación

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 dimensiones, ya que una misma dimensión puede utilizarse para varias tablas de hechos.
- ✓ No es soportado por todas las herramientas de consulta y análisis.

Se propone utilizar el modelado multidimensional en estrella debido a su simplicidad y velocidad en el análisis multidimensional. Es el esquema con mejor rendimiento y velocidad, que permite indexar las dimensiones de forma individualizada sin repercusión en el rendimiento de la base de datos en su conjunto. Además, este diseño permite implementar la funcionalidad de una base de datos multidimensional utilizando una base de datos relacional.

1.7. Modelo de almacenamiento OLAP

Los sistemas OLAP son bases de datos orientadas al procesamiento analítico. Este análisis suele implicar, generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil: tendencias de ventas, patrones de comportamiento de los consumidores, elaboración de informes complejos, entre otros. Este sistema es típico de un DM.

Existen varias arquitecturas para los sistemas OLAP:

- ✓ OLAP multidimensional (*MOLAP*, por sus siglas en inglés).

La arquitectura MOLAP usa unas bases de datos multidimensionales para proporcionar el análisis, su principal premisa es que el OLAP está mejor implantado almacenando los datos multidimensionalmente. Utiliza una arquitectura de dos niveles: las bases de datos multidimensionales y el motor analítico. La base de datos multidimensional es la encargada del manejo, acceso y obtención del dato.

- ✓ OLAP relacional (*ROLAP*, por sus siglas en inglés).

  La arquitectura ROLAP, accede a los datos almacenados en un AD para proporcionar los análisis OLAP. La premisa de los sistemas ROLAP es que las capacidades OLAP se soportan mejor contra las bases de datos relacionales. Utiliza una arquitectura de tres niveles. La base de datos relacional maneja los requerimientos de almacenamiento de datos, y el motor ROLAP proporciona la funcionalidad analítica. El nivel de base de datos usa bases de datos relacionales para el manejo, acceso y obtención del dato. El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios.

En la construcción del MD se utilizó el sistema ROLAP. Este ofrece ventajas como el uso total de la seguridad e integridad de los datos para grandes volúmenes de información. Además, es la arquitectura que mejor soporta el análisis OLAP contra las bases de datos relacionales.

1.8. Metodología para el desarrollo de almacenes de datos

Existen muchas metodologías de diseño y construcción para los AD. Cada fabricante de software de inteligencia de negocios busca imponer una metodología con sus productos. (Rivadera, 2010) Sin embargo, se imponen entre la mayoría tres metodologías, la de Kimball, la de Inmon y la de Hefesto.

1.8.1. Kimball

Esta metodología es muy eficaz y conduce a una solución completa en una cantidad de tiempo muy pequeña, lo que es muy efectivo en el proceso de toma de decisiones. Propone el desarrollo iterativo incremental, donde se construye una pieza a la vez.

Se basa en lo que Kimball denomina Ciclo de Vida Dimensional del Negocio. Este ciclo de vida está basado en cuatro principios básicos: (Kimball & Margy, 2002)

- ✓ Centrarse en el negocio: se centra en la identificación de los requisitos del negocio y su valor asociado, y usar los esfuerzos para desarrollar relaciones

sólidas con el negocio, agudizando el análisis del mismo y la competencia consultiva de los implementadores.

- ✓ Construir una infraestructura de información adecuada: Diseña una base de información única, integrada, fácil de usar, de alto rendimiento donde se reflejará la amplia gama de requerimientos del negocio identificado en la empresa.

- ✓ Realizar entregas en incrementos significativos: crea el almacén de datos en incrementos entregables en plazos de 6 a 12 meses. Hay que usar el valor del negocio de cada elemento identificado para determinar el orden de aplicación de los incrementos. En esto la metodología se parece a las metodologías ágiles de construcción de software.

- ✓ Ofrecer la solución completa: proporciona todos los elementos necesarios para entregar valor a los usuarios de negocios. Para comenzar, esto significa tener un almacén de datos sólido, bien diseñado, con calidad probada, y accesible. También se deberá entregar herramientas de consulta ad hoc, aplicaciones para informes y análisis avanzados, capacitación, soporte, sitio web y documentación.

La construcción de una solución de AD es sumamente compleja, y Kimball propone una metodología que ayuda a simplificar esa complejidad. Las tareas de esta metodología se muestran en la siguiente figura:

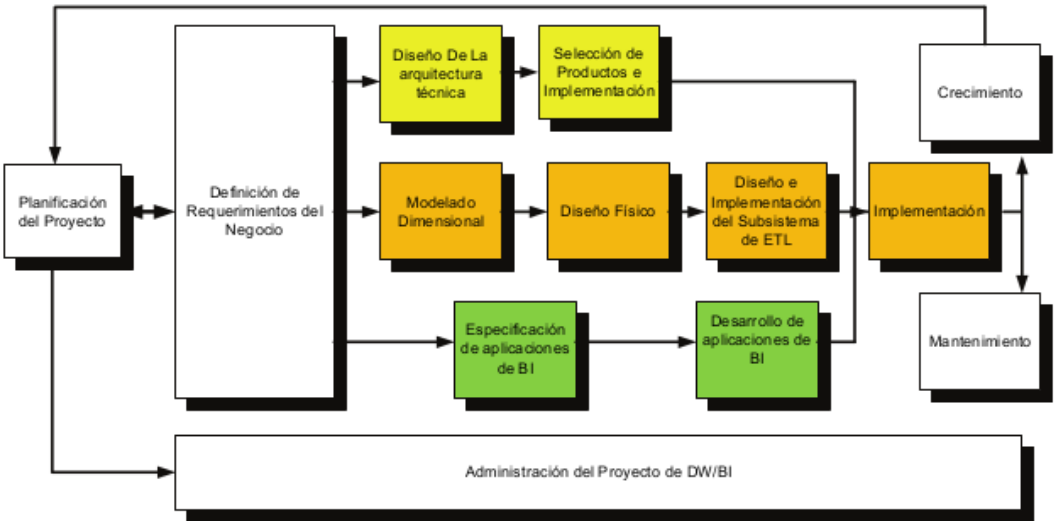


Figura 6: Ciclo de Vida de Kimball

1.8.2. Inmon

La metodología Inmon, llamada así en homenaje a su autor Bill Inmon, está basada en una arquitectura descendente (*Top-Down*, por sus siglas en inglés). Hace énfasis en los Almacenes de Datos, además está compuesta por varios niveles de áreas de interés y MD dependientes. Contiene datos del Almacén de Datos a nivel atómico, y datos del Mercado de Datos sumarizados. Esta metodología puede tener una implementación tardía y es recomendada cuando se hace demasiado difícil representar el modelo a través de dimensiones y la complejidad de la solución se hace demasiado grande. No es muy recomendable para proyectos sencillos pues va de lo general, el Almacén de Datos, a lo más específico, el MD. (Kimball I. , 2012)

1.8.3. Hefesto

Hefesto es una metodología cuya propuesta está fundamentada en una amplia investigación, comparación de metodologías existentes y experiencias propias en proceso de confección de los almacenes de datos. (Bernabeu, 2007)

La idea principal, es comprender cada paso que se realizará, para no caer en el tedio de tener que seguir un método al pie de la letra sin saber exactamente que se está haciendo, ni por qué.

Esta metodología puede resumirse a través del siguiente gráfico:

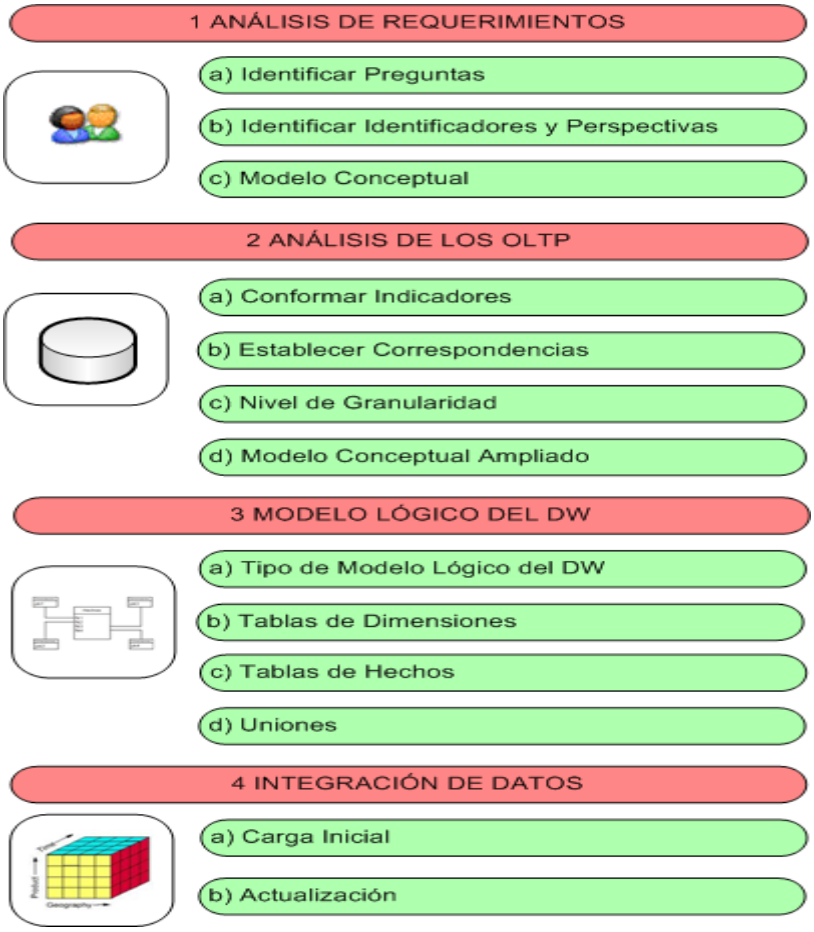


Figura 7: Fases de la Metodología de Hefesto

Como se puede apreciar, se comienza recolectando las necesidades de información de los usuarios y se obtienen las preguntas claves del negocio. Luego, se deben identificar los indicadores resultantes de las interrogantes y sus respectivas perspectivas de análisis, mediante las cuales se construirá el modelo conceptual del AD.

Después, se analizarán los OLTP (sistemas de Procesamiento Transaccional en Línea) para señalar las correspondencias con los datos fuentes y seleccionar los campos de estudio de cada perspectiva.

Una vez hecho esto, se pasará a la construcción del modelo lógico del depósito, explicitando las jerarquías que intervendrán.

Por último, se definirán los procesos de carga, transformación, extracción y limpieza de los datos fuente.

Esta metodología cuenta con las siguientes características: (Bernabeu, 2007)

- ✓ 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 AD.
- ✓ 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.
- ✓ Se aplica tanto para Almacén de Datos, como para Mercado de Datos.

Luego del análisis realizado se define Hefesto como la metodología a utilizar, esta permite construir el MD de forma sencilla, intuitiva y ordenada. Es una metodología ideal para las personas que entran por primera vez en el mundo de los almacenes de datos, posee métodos, pasos lógicos que se relacionan durante las etapas del proceso de confección. Agiliza el proceso de desarrollo del MD debido a que propone no entrar en fases extensas de reunión de análisis y requerimientos ni fases de despliegue muy largas. (Bernabeu, 2007)

1.9. Herramientas para la construcción del Mercado de datos

1.9.1. Herramientas CASE

Las herramientas CASE (en inglés: ComputerAided Software Engineering,en español: Ingeniería de Software Asistida por Computadora.) son un conjunto de métodos, utilidades y técnicas que facilitan la automatización del ciclo de vida del desarrollo de un sistema. En ellas se integran el análisis de datos y procesos integrados mediante un repositorio, generación de interfaces entre el análisis y el diseño, generación del código a partir del diseño, y control de mantenimiento.
Existen disímiles herramientas CASE entre ellas se encuentran Microsoft Project, Rational Rose, JDeveloper, MagicDraw, Microsoft Visio, BoUML. Además, existe la herramienta Visual Paradigm que se utiliza en el proyecto, la cual posee licencia en la Universidad de las Ciencias Informáticas.

Visual Paradigm 8.0 es una herramienta CASE profesional, que soporta el ciclo de vida completo de desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue. Permite dibujar todos los tipos de diagramas de

clases, código inverso, generar código desde diagramas y generar documentación, además proporciona abundantes tutoriales de UML, demostraciones interactivas de UML y proyectos UML. (Pressman, 2002) También, la herramienta es colaborativa, es decir, soporta múltiples usuarios trabajando sobre el mismo proyecto y permite control de versiones. Cabe destacar igualmente su robustez, usabilidad y portabilidad. Esta herramienta presenta numerosas características, que resultan útiles para elaborar una aplicación con gran calidad:

- ✓ Soporta aplicaciones Web.
- ✓ Varios idiomas.
- ✓ Generación de código para Java y exportación como HTML.
- ✓ Fácil de instalar y actualizar.
- ✓ Compatibilidad entre ediciones.

1.9.2. Herramientas de integración de datos

Las herramientas de integración de datos son aquellas que proporcionan de forma general una serie de funcionalidades, como, por ejemplo: (Pentaho Solution, 2015)

- ✓ Control de la extracción de los datos y su automatización, disminuyendo el tiempo empleado en el descubrimiento de procesos no documentados, minimizando el margen de error y permitiendo mayor flexibilidad.
- ✓ Acceso a diferentes tecnologías, haciendo un uso efectivo del hardware, software, datos y recursos humanos existentes.
- ✓ Proporcionar la gestión integrada del Almacén de datos y los Mercados de datos existentes, integrando la extracción, transformación y carga para la construcción del Almacén de datos corporativo y de los Mercados de datos.
- ✓ Uso de la arquitectura de metadatos, facilitando la definición de los objetos de negocio y las reglas de consolidación.
- ✓ Acceso a una gran variedad de fuentes de datos diferentes.
- ✓ Manejo de excepciones.
- ✓ Planificación, registros, interfaces a programadores de terceros, que permitirán llevar una gestión de la planificación de todos los procesos necesarios para la carga del almacén de datos.
- ✓ Interfaz independiente de hardware.
- ✓ Soporte en la explotación del almacén de datos.

Pentaho Data Integration (PDI) 6.0.0 es un herramienta de integración de datos de código abierto que se encarga de la extracción, transformación y carga (ETL) de los procesos de integración de datos (Pentaho Solution, 2015). Permite desarrollar y desplegar poderosas aplicaciones de Business Intelligence iterativas con la participación de desarrolladores y usuarios finales, combinando el desarrollo de soluciones complejas en un solo proceso, con un ahorro considerable de tiempo. PDI está conformado por varias herramientas con un propósito en específico, las cuales son:

- ✓ **Spoon:** herramienta principal de trabajo que permite el diseño de las transformaciones y los trabajos (Jobs).
- ✓ **Pan:** herramienta que permite ejecutar las transformaciones desarrolladas en el Spoon y permite ejecutar scripts desde la línea de comandos.
- ✓ **Kitchen:** permite ejecutar los Jobs diseñados en el Spoon.
- ✓ **Carte:** servidor web para ejecutar remotamente las transformaciones y los trabajos.

1.9.3. Sistema gestor de Base de datos

Sistema Gestor de Base de Datos (SGBD) son un tipo de software específico, dedicados a servir de interfaz entre la base de datos, el usuario y las aplicaciones que lo utilizan. Las principales funciones que debe cumplir un SGBD, se relacionan con la creación y mantenimiento de la base de datos, el control de accesos, la manipulación de datos de acuerdo con las necesidades del usuario, el cumplimiento de las normas de tratamiento de datos, evitar redundancias e inconsistencias y mantener la integridad.

PostgreSQL 9.2.4: Servidor de base de datos relacional, distribuido bajo licencia Distribución de Software Berkeley (BSD, por sus siglas en inglés) y con su código fuente disponible libremente. Incluye características de la orientación a objetos, como puede ser la herencia, tipos de datos, funciones, restricciones, disparadores, reglas e integridad transaccional. (Postgres SQL, 2013). Las principales características de este gestor de bases de datos son:

- ✓ Implementación del estándar (lenguaje) SQL92/SQL99.
- ✓ Soporta distintos tipos de datos: además del soporte para los tipos base, también soporta datos de tipo fecha, monetarios, elementos gráficos, datos

sobre redes (MAC, IP), cadenas de bits. También permite la creación de tipos propios.

- ✓ Incorpora una estructura de datos arreglos (array).
- ✓ Incorpora funciones de diversas índoles: manejo de fechas, geométricas, orientadas a operaciones con redes.
- ✓ Incluye herencia entre tablas (aunque no entre objetos, ya que no existen), por lo que a este gestor de bases de datos se le incluye entre los gestores objeto-relacionales.
- ✓ Permite la gestión de diferentes usuarios, como también los permisos asignados a cada uno de ellos.
- ✓ Soporta casi toda la sintaxis SQL (incluyendo sub consultas, transacciones, tipos y funciones definidas por el usuario).
- ✓ El progreso continuo del gestor de datos de código abierto PostgreSQL brinda a los consumidores la opción de instalar una base de datos no privativa.

1.9.4. Herramienta para el proceso analítico en línea

Las herramientas OLAP proporcionan a las compañías un sistema confiable para procesar datos que luego serán utilizados para llevar a cabo análisis e informes que permitan mejorar las operaciones productivas, tomar decisiones inteligentes y optimizar la competitividad en el mercado.

Mondrian es una de las aplicaciones más importantes de la plataforma Pentaho BI. Mondrian es un servidor OLAP *open source* que gestiona la comunicación entre una aplicación OLAP (escrita en Java) y la base de datos con los datos fuentes. Es decir, Mondrian actúa como “JDBC para OLAP” (Pentaho BI, 2016). Permite analizar grandes conjuntos de datos que se encuentran almacenados en el MD, pues se encarga de recibir consultas dimensionales en lenguaje de Expresiones multidimensionales (*MDX*, por sus siglas en inglés) y devolver los datos del cubo que correspondan a la consulta. El cubo se representa como un conjunto de metadatos que definen cómo se han de mapear estas consultas dimensionales a sentencias SQL, para obtener de la base de datos la información necesaria para satisfacer la consulta dimensional (Pentaho BI, 2016). Para acceder a las funcionalidades de Mondrian hay que hacer uso del cliente STPivot. Este cliente es un visor web OLAP, de código abierto, sobre la base del visor por defecto de JPivot, el cual es una librería de componentes Servidor de Páginas Java (*JSP*, por sus siglas en inglés), que se utiliza

para construir tablas OLAP generadas de forma dinámica y permite a los usuarios realizar consultas OLAP por medio del lenguaje MDX. El objetivo de STPivot es mejorar la experiencia del usuario de JPivot mediante el aprovechamiento de las bibliotecas de interfaz de usuario libre y las tecnologías jQuery y Ajax (BI 2012). La herramienta que utiliza Mondrian para crear cubos OLAP es el SchemaWorkbench; que es un entorno visual para el desarrollo y prueba de cubos OLAP Mondrian, se utiliza para la creación de los archivos XML que se usan para la construcción de los cubos. Permite la ejecución de consultas MDX contra el esquema y la base de datos.

Conclusiones del capítulo

Para apoyar el proceso de toma de decisiones de la FICI se decidió implementar un MD porque es más efectivo en áreas específicas del negocio, además de que su costo de uso es inferior al de un AD y su tiempo de construcción y puesta en marcha es menor que el de los AD. Para guiar este proceso se realizó una investigación sobre las diferentes metodologías existentes para la construcción de MD, lo que posibilitó la selección de la metodología Hefesto como apoyo para el proceso de desarrollo del MD debido a que esta permite un desarrollo rápido, sencillo y organizado. El estudio de las herramientas existentes para la construcción de MD permitió hacer una selección de las herramientas adecuadas para la realización del sistema de acuerdo a las necesidades del cliente.

Capítulo 2: Diseño e implementación del Mercado de datos.

En este capítulo se abordan una serie de elementos que posibilitan un mayor entendimiento del negocio. Se describen cada uno de los pasos a seguir para la construcción del MD haciendo uso de la metodología seleccionada, y de las técnicas de análisis de datos como el proceso ETL, utilizado para limpiar, transformar y cargar los datos; para diseñar el cubo multidimensional. Se definen las necesidades del cliente a través del levantamiento de requisitos, las reglas y modelado de los datos con sus elementos tales como dimensiones, hechos y medidas.

2.1. Propuesta del sistema

Se propone el desarrollo de un MD basado en la metodología Hefesto, que, a partir del avance de cada proceso, se obtiene una herramienta que contribuye a la toma de decisiones referente a la información de los estudiantes del primer año de la carrera de Ingeniería en Ciencias Informáticas. La información se obtiene de Akademos, sistema que contiene un módulo estudiantes donde se gestionan las notas y algunos datos personales de los estudiantes. En la siguiente figura se muestra un reporte brindado por el sistema.

FICI

FI01

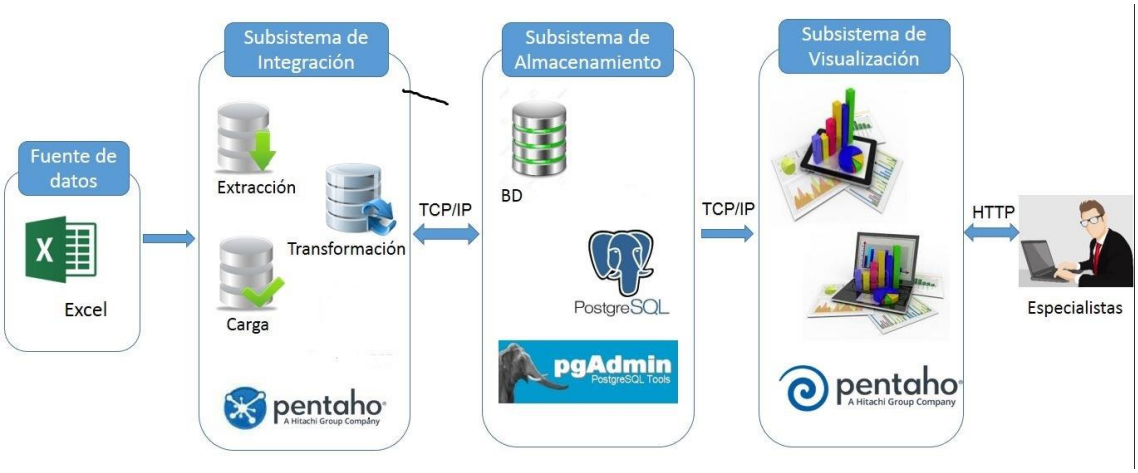
Estudiantes/Asignaturas									Algebra Lineal	Educación Física 1	Introducción a la	Introducción a las	Matemática 1	Matemática Discret	Seguridad Nacion
Nombre Completo	Estado Docente	Situacion Escolar	Sexo	Vía de Ingreso	Centro de Procedencia	Provincia	Municipio	Carnet de Identidad	Evaluación	Evaluación	Evaluación	Evaluación	Evaluación	Evaluación	Evaluación
Estudiante 1	Matriculado	Nuevo Ingreso	Masculino	Preuniversitario	IPVCE	Villa Clara	Sagua la Grande	95071037946	4	5	5	4	4	3	4
Estudiante 2	Matriculado	Nuevo Ingreso	Femenino	Preuniversitario	IPU	La Habana	Marianao	96091507538	4	4	2	4	4	4	4
Estudiante 3	Matriculado	Nuevo Ingreso	Masculino	Preuniversitario	IPR	Camagüey	Florida	95050640661	3	5	4	4	4	3	5
Estudiante 4	Matriculado	Reingreso	Masculino	Cadetes MININT	IPU	La Habana	Cerro	94031604908	3	-	3	-	3	2	-
Estudiante 5	-	-	Femenino	Preuniversitario	IPU	La Habana	Guanabacoa	96081908732	-	-	-	-	-	-	-
Estudiante 6	Matriculado	Repitente	Femenino	Preuniversitario	EMCC	La Habana	La Lisa	95051830578	-	-	-	-	3	-	-
Estudiante 7	Matriculado	Nuevo Ingreso	Femenino	Preuniversitario	IPU	Pinar del Río	Pinar del Río	96040602995	4	4	3	5	3	3	4
Estudiante 8	Matriculado	Nuevo Ingreso	Masculino	Cadetes MINFAR	EMCC	Matanzas	Jovellanos	96041811408	4	5	2	3	2	2	3
Estudiante 9	Matriculado	Nuevo Ingreso	Masculino	Preuniversitario	IPU	La Habana	Cerro	95020136200	4	5	5	5	3	2	5
Estudiante 10	Matriculado	Nuevo Ingreso	Femenino	Preuniversitario	IPU	La Habana	Boyeros	96103109478	4	4	5	4	4	2	3

Figura 8: Reporte de evaluaciones del primer semestre 1er año. Fuente: Elaboración propia

Otra fuente de donde se obtiene información es a través de los diagnósticos iniciales que se les realizan a los estudiantes al inicio de la carrera. Estos datos son almacenados en el CICE.

2.2. Arquitectura del mercado de datos

La arquitectura utilizada para el sistema propuesto en la investigación es la misma explicada con anterioridad. En este acápite se explican cada uno de sus componentes basados en la situación que presenta la investigación.



El componente Fuente de Datos: define las fuentes que se utilizarán en la obtención de los datos que utiliza el sistema. Para la versión inicial del sistema se utilizan archivos planos mayormente en formato Excel extraídos del sistema Akademos, principal fuente de información para poblar el Mercado de Datos. Otra de las fuentes utilizadas son los informes docentes que se obtienen de los profesores y los diagnósticos iniciales que se les practican a los estudiantes.

El componente Subsistema de Integración: es la sección donde se agrupan una serie de procesos que llevan a cabo tareas relacionadas con la extracción, manipulación, control, integración, limpieza de datos, carga y actualización de los datos a utilizar en el sistema, todas las tareas que se hagan desde que se toman los datos de los archivos Excel, hasta que se carguen en el sistema para su utilización en la construcción de los cubos de datos.

El componente Subsistema de Almacenamiento: en esta parte del sistema se mantienen los datos obtenidos en una base de datos temporal que se encuentra en el gestor de base de datos PostgreSQL.

El componente Subsistema de Visualización: es el área correspondiente a la interacción con el usuario, cuya funcionalidad es mostrar los datos almacenados de forma útil y transparente a través de las distintas herramientas. Este sistema se comunica directamente con el servidor de cubos a través de consultas, las cuales retornan la información requerida donde ésta es transformada y presentada para la

visualización final. Los reportes y vistas de análisis requeridos en el sistema se encuentran en esta área.

2.3. Descripción de las fases de la metodología Hefesto

La metodología Hefesto cuenta con cuatro fases, las cuales describen el proceso de desarrollo de un MD. La primera fase llamada **Análisis de requerimientos** es la encargada de recolectar las necesidades de información de los usuarios y obtener preguntas claves para el negocio, además, se identifican los indicadores y perspectivas, mediante las cuales se construirá el modelo conceptual de datos del MD. La segunda fase nombrada **Análisis de los OLTP** permite conformar los indicadores para luego establecer las correspondencias con los datos fuentes, se obtienen además los diferentes niveles de granularidad y se construye el modelo conceptual ampliado. La tercera fase es **Modelo Lógico del MD** en la que después de realizado el análisis, se construye el modelo lógico del MD, se crean las tablas de hechos y dimensiones para posteriormente realizar las uniones entre estas y especificar las jerarquías que intervienen. La última fase es la **Integración de datos** en la cual se definen los procesos de carga, transformación, extracción y limpieza de los datos fuentes, a partir de la carga inicial y de sus posteriores actualizaciones.

2.3.1. Fase 1: Análisis de los requerimientos

En esta fase se identifican los requerimientos de usuarios a través de preguntas, las cuales permiten dar cumplimiento a los objetivos trazados. Se analizan las preguntas para poder determinar cuáles serán los indicadores y perspectivas que serán utilizados en la construcción del MD y al final se confecciona el modelo conceptual donde se visualiza el resultado obtenido en este paso.

Identificar requisitos

Se realizaron entrevistas informales a diferentes profesores y dirigentes de la FICI con el objetivo de obtener las necesidades de información de los usuarios, los resultados que se esperan obtener y los reportes que consideren importantes para la FICI.

De estas entrevistas realizadas se obtuvieron las siguientes preguntas claves:

- ✓ Se desea conocer la cantidad de estudiantes, matriculados en una asignatura, por sexo y raza en un período de tiempo determinado.
- ✓ Se desea conocer la cantidad de estudiantes, matriculados en una asignatura, por estado de salud en un período de tiempo determinado.

- ✓ Se desea conocer la cantidad de estudiantes, matriculados en una asignatura, por situación social en un período de tiempo determinado.
- ✓ Se desea conocer la cantidad de estudiantes, matriculados en una asignatura, por centro de procedencia en un período de tiempo determinado.
- ✓ Se desea conocer la cantidad de estudiantes, matriculados en una asignatura, por provincia y municipio en un período de tiempo determinado.
- ✓ Se desea conocer la cantidad de estudiantes, matriculados en una asignatura, por opción en que solicito la carrera y número de escalafón en un período de tiempo determinado.
- ✓ Se desea conocer la cantidad de estudiantes, matriculados en una asignatura, por cantidad de hijos y tipo de zona en un período de tiempo determinado.
- ✓ Se desea conocer la cantidad de estudiantes, matriculados en una asignatura en un período de tiempo determinado.
- ✓ Se desea conocer cantidad de desaprobados por asignaturas, provincia, municipio y centro de procedencia en un período de tiempo determinado.
- ✓ Se desea conocer cantidad de desaprobados por asignaturas y tipo de evaluación en un período de tiempo determinado.
- ✓ Se desea conocer cantidad de desaprobados por asignaturas, opción en que solicito la carrera y número de escalafón en un período de tiempo determinado.
- ✓ Se desea conocer cantidad de desaprobados por asignaturas con examen final en un período de tiempo determinado.
- ✓ Se desea conocer cantidad de estudiantes con 3 por asignaturas, provincia, municipio y centro de procedencia en un período de tiempo determinado.
- ✓ Se desea conocer cantidad de estudiantes con 3 por asignaturas, opción en que solicito la carrera y número de escalafón en un período de tiempo determinado.
- ✓ Se desea conocer cantidad de estudiantes con 3 por asignaturas con examen final en un período de tiempo determinado.
- ✓ Se desea conocer cantidad de estudiantes con 4 por asignaturas, provincia, municipio y centro de procedencia en un período de tiempo determinado.

- ✓ Se desea conocer cantidad de estudiantes con 4 por asignaturas, opción en que solicito la carrera y número de escalafón en un período de tiempo determinado.
- ✓ Se desea conocer cantidad de estudiantes con 4 por asignaturas con examen final en un período de tiempo determinado.
- ✓ Se desea conocer cantidad de estudiantes con 5 por asignaturas, provincia, municipio y centro de procedencia en un período de tiempo determinado.
- ✓ Se desea conocer cantidad de estudiantes con 5 por asignaturas, opción en que solicito la carrera y número de escalafón en un período de tiempo determinado.
- ✓ Se desea conocer cantidad de estudiantes con 5 por asignaturas con examen final en un período de tiempo determinado.

Identificar indicadores y perspectivas

Una vez que se establecen las preguntas claves, se debe proceder a la descomposición de las mismas para la obtención de los indicadores que se utilizarán y las perspectivas de análisis que van a intervenir. Para ello hay que tener en cuenta que los indicadores son por lo general medidas numéricas que representan lo que se desea analizar concretamente, mientras que las perspectivas se refieren a los objetos mediante los cuales se quiere examinar los indicadores, con el fin de responder las preguntas planteadas (Bernabeu, 2007).

A partir de las preguntas que se obtuvieron de la entrevista realizada se pueden definir los siguientes indicadores con sus respectivas perspectivas:

Indicadores

Los indicadores son valores numéricos y representan lo que se desea analizar concretamente, por ejemplo: saldos, promedios, cantidades, sumatorias, fórmulas, entre otras. (Bernabeu, 2007)

Perspectivas

Las perspectivas son objetos mediante los cuales se quiere examinar los indicadores, con el objetivo de responder a las preguntas planteadas, por ejemplo: clientes, proveedores, sucursales, países, productos, rubros, entre otras. (Bernabeu, 2007)

A partir de los requisitos identificados en la entrevista realizada se obtuvieron los siguientes indicadores con sus correspondientes perspectivas:

Diseño e implementación del Mercado de datos

Cantidad de estudiantes matriculados en una asignatura, por sexo y raza en un tiempo determinado.

Indicador **Perspectiva**

Cantidad de estudiantes en una asignatura por estado de salud en un tiempo determinado.

Indicador **Perspectiva**

Cantidad de estudiantes en una asignatura por situación económica en un tiempo determinado.

Indicador **Perspectiva**

Cantidad de estudiantes en una asignatura por centro de procedencia en un tiempo determinado.

Indicador **Perspectiva**

Cantidad de estudiantes en una asignatura por provincia y municipio en un tiempo determinado,

Indicador **Perspectiva**

Cantidad de estudiantes en una asignatura por opción de solicitud de la carrera y número de escalafón en un tiempo determinado.

Indicador **Perspectiva**

Cantidad de estudiantes en una asignatura por cantidad de hijos y tipo de zona en un tiempo determinado.

Indicador **Perspectiva**

Cantidad de desaprobados por asignaturas, provincia, municipio y centro de procedencia en un tiempo determinado.

Indicador **Perspectiva**

Cantidad de desaprobados por asignaturas y tipo de evaluación en un tiempo.

Indicador **Perspectiva**

determinado

Modelo conceptual

A partir de los indicadores y perspectivas obtenidas en el paso anterior se construirá el modelo conceptual. Este permitirá observar con claridad cuáles son los alcances del proyecto, para luego poder trabajar sobre ellos. Además, permite ser presentado y explicado con facilidad ante los usuarios debido a que posee un alto nivel de definición de los datos (Bernabeu, 2007). A continuación se presenta el modelo conceptual donde a la izquierda están representadas las perspectivas y a la derecha los indicadores.

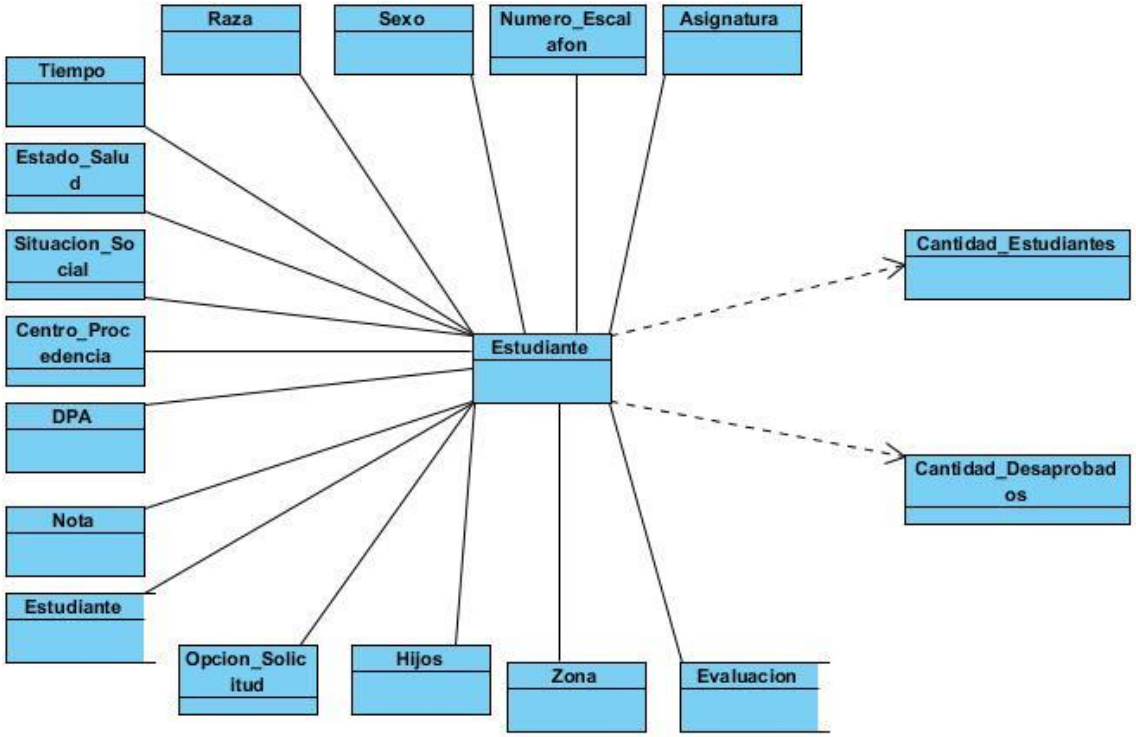


Figura 9: Modelo conceptual

2.3.2. Fase 2: Análisis de los OLTP

El objetivo de esta fase es examinar los OLTP disponibles que contengan la información requerida para poder identificar las correspondencias entre el modelo conceptual y las fuentes de datos. Además, se definen los campos que serán incluidos en cada perspectiva y se ampliará el modelo conceptual con la información obtenida en este paso.

Conformar indicadores

Este paso se realiza para obtener los hechos que componen los indicadores con sus respectivas fórmulas.

Hechos	Función de sumarización	Aclaraciones
Cantidad de estudiantes	distinct-count	Cuenta el total de estudiantes.
Cantidad de desaprobados	distinct-count	Cuenta el total de estudiantes desaprobados.

Tabla 1: Especificación de las medidas

Establecer correspondencias

El objetivo de este análisis, es el de examinar los OLTP disponibles que contengan la información requerida para poder identificar las correspondencias entre el modelo conceptual y las fuentes de datos.

Para el caso de los indicadores, deben explicarse cómo se procederá a su cálculo, y más aún si son fórmulas u operaciones complejas.

Para establecer la correspondencia entre la base de datos y el DM las relaciones identificadas son las siguientes:

- ✓ El campo “sexo_cod” de la tabla “Estudiantes” se relaciona con la perspectiva “Sexo”.
- ✓ El campo “raza_cod” de la tabla “Estudiantes” se relaciona con la perspectiva “Raza”.
- ✓ El campo “asignatura_cod” de la tabla “Asignatura” se relaciona con la perspectiva “Asignatura”.
- ✓ Los campos “año_cod”,” semestre_cod” y “mes_cod” de la tabla “Tiempo” se relacionan con la perspectiva “Tiempo”.
- ✓ El campo “estado_salud_cod” de la tabla “Estudiantes” se relaciona con la perspectiva “Estado_Salud”.
- ✓ El campo “situación_social_cod” de la tabla “Estudiantes” se relaciona con la perspectiva “Situacion_Social”.
- ✓ El campo “centro_procedencia_cod” de la tabla “Estudiantes” se relaciona con la perspectiva “Centro_Procedencia”.

- ✓ Los campos “provincia_cod” y “municipio_cod” de la tabla “Estudiantes” se relacionan con la perspectiva “Dpa”.
- ✓ El campo “opcion_solicitud_cod” de la tabla “Estudiantes” se relaciona con la perspectiva “Opcion_Solicitud”.
- ✓ El campo “escalafon_cod” de la tabla “Estudiantes” se relaciona con la perspectiva “Numero_Escalafon”.
- ✓ El campo “hijos_cod” de la tabla “Estudiantes” se relaciona con la perspectiva “Cantidad_Hijos”.
- ✓ El campo “zona_cod” de la tabla “Estudiantes” se relaciona con la perspectiva “Tipo_Zona”.

Nivel de granularidad

En este paso se examinan y seleccionan los campos que contendrá cada perspectiva, ya que a través de estos se manipularán y se filtrarán los indicadores. También se deben presentar los datos de análisis disponibles para cada perspectiva y se debe decidir cuáles son relevantes y cuáles no.

- **Perspectiva “Asignatura”:**
  - ✓ asignatura_cod: identificador de la asignatura.
  - ✓ Nombre_asignatura: referido al nombre de la asignatura
- **Perspectiva “Sexo”:**
  - ✓ sexo_cod: identificador del sexo del estudiante.
  - ✓ sexo
  - ✓ sexo_descripcion: sexo del estudiante más detallado.
- **Perspectiva “Raza”:**
  - ✓ raza_cod: identificador de la raza del estudiante.
  - ✓ raza
  - ✓ raza_descripcion: raza del estudiante detallado.
- **Perspectiva “Tiempo”:**
  - ✓ año_cod: identificador del año.
  - ✓ año_nombre: nombre del año.
  - ✓ año_numero: número del año.
  - ✓ semestre_cod: identificador del semestre.
  - ✓ semestre_nombre: nombre del semestre.
  - ✓ semestre_numero: número del semestre.
  - ✓ mes_cod: identificador del mes.
  - ✓ mes_nombre: nombre del mes.
  - ✓ mes_numero: número del mes.

- **Perspectiva "Estado salud":**
  - ✓ estado_salud_cod: identificador del estado de salud del estudiante.
  - ✓ nombre_estado_salud: referido al nombre del estado de salud (bueno, regular, malo).
  - ✓ estado_salud_descripcion: estado de salud del estudiante detallado.
- **Perspectiva "Situación social":**
  - ✓ situación_social_cod: identificador de la situación social del estudiante.
  - ✓ nombre_situacion_social: nombre de la situación social (buena, regular, mala) del estudiante.
  - ✓ situacion_social_descripcion: situación social del estudiante más detallado.
- **Perspectiva "Centro de procedencia":**
  - ✓ centro_procedencia_cod: identificador del centro de procedencia del estudiante.
  - ✓ nombre_centro_procedencia: referido al nombre del centro de procedencia (IPVC, IPUEC, IPI, Concurso) del estudiante.
  - ✓ centro_procedencia_descripcion: información sobre el centro de procedencia del estudiante más detallado.
- **Perspectiva "DPA":**
  - ✓ provincia_cod: identificador de la provincia a la que pertenece el estudiante.
  - ✓ provincia_nombre: referido al nombre de la provincia.
  - ✓ provincia_descripcion: información sobre la provincia.
  - ✓ municipio_cod: identificador del municipio del estudiante.
  - ✓ municipio_nombre: nombre del municipio.
  - ✓ municipio_descripcion: información sobre el municipio.
- **Perspectiva "Opción solicitud":**
  - ✓ opción_solicitud_cod: identificador de la opción de solicitud del estudiante.
  - ✓ numero_opcion_solicitud: referido al número en que se solicitó la Carrera de Ingeniería en Ciencias Informáticas.
- **Perspectiva "Número de escalafón":**
  - ✓ escalafon_cod: identificador del número de escalafón del estudiante.
  - ✓ numero_escalafon: referido al número de escalafón del estudiante.
- **Perspectiva "Cantidad hijos":**
  - ✓ hijos_cod: identificador de cantidad de hijos del estudiante.

- ✓ cantidad_hijos: referido a la cantidad de hijos que tiene el estudiante.

- **Perspectiva "Tipo Zona":**
  - ✓ zona_cod: identificador del tipo de zona del estudiante.
  - ✓ nombre_zona: referido al nombre de la zona (rural, urbana).
  - ✓ zona_descripcion: información más detallada sobre la zona.
- **Perspectiva "Estudiante":**
  - ✓ estudiante_cod: identificador del tipo de evaluación del estudiante.
  - ✓ estudiante_nombre: referido al nombre del estudiante.
  - ✓ estudiante_ci: carnet de identidad del estudiante.
- **Perspectiva "Tipo Evaluación":**
  - ✓ evaluación_cod: identificador del tipo de evaluación del estudiante.
  - ✓ nombre_evaluacion: referido al nombre de la evaluación (PE, PP, PF, S, T, L, NF)
  - ✓ evaluacion_descripcion: información sobre la evaluación más detallada.
- **Perspectiva "Nota":**
  - ✓ nota_cod: identificador de la nota.
  - ✓ nota_numero: valor de la nota del estudiante.
  - ✓ nota_estado: referido al estado de la nota (aprobado o desaprobado).

Modelo conceptual ampliado

Se grafican los resultados obtenidos en pasos anteriores, ampliando el modelo conceptual. Para ello se coloca debajo de cada perspectiva los campos o atributos seleccionados y en los indicadores su función de sumarización.

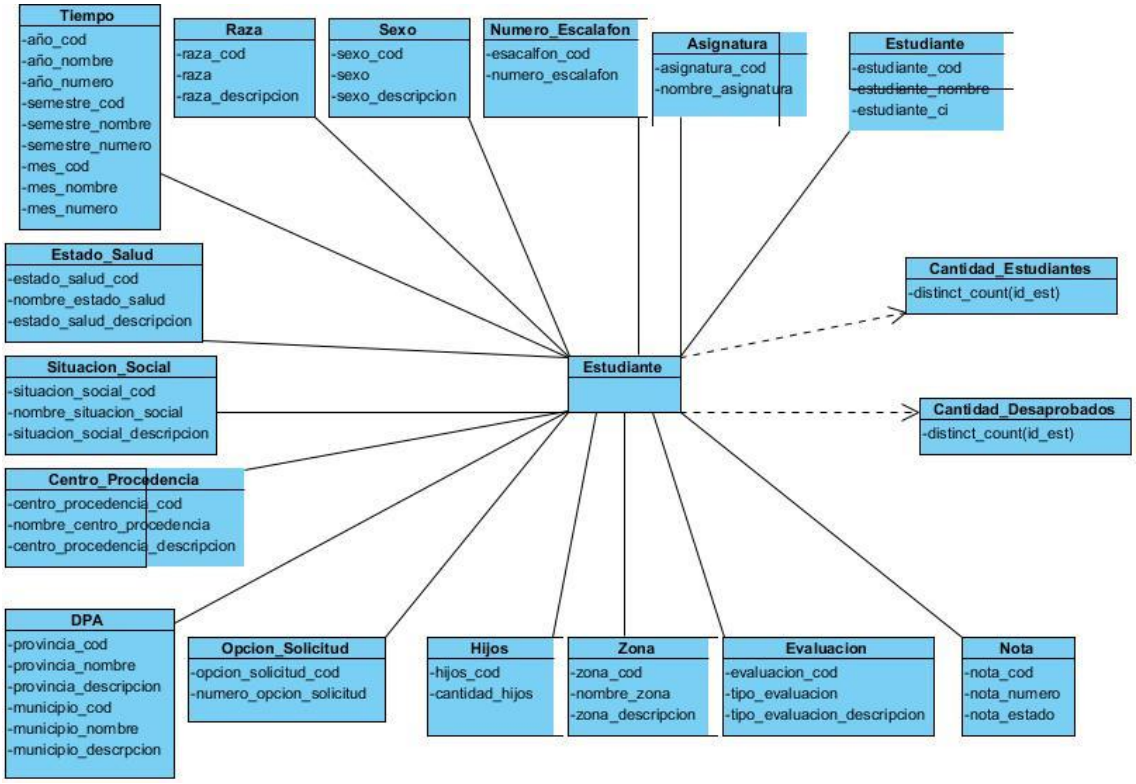


Figura 10: Modelo Conceptual Ampliado

2.3.3. Fase 3: Modelo lógico del MD

Para esta fase se debe confeccionar el modelo lógico de la estructura del MD, teniendo como base el modelo conceptual ya creado. Para ello se define el tipo de modelo que se utilizará, se diseñan las tablas de dimensiones y hechos para finalmente realizar las uniones pertinentes entre estas tablas.

Tipo de modelo lógico del MD

Para obtener la estructura del MD se selecciona el esquema en estrella, el cual se explicó en el capítulo anterior.

Tablas de dimensiones

En este paso se diseñan las tablas de dimensiones que conforman el MD. Cada perspectiva representa una tabla de dimensión. Para ello se toma cada perspectiva con sus campos relacionados y se elige el nombre que identifica a la tabla de dimensión. Se añade un campo que represente su clave principal y se definen los nombres de los campos que lo necesiten. A continuación, se muestra el ejemplo de la dimensión ¨Raza¨, para ver las demás dimensiones consultar anexos.

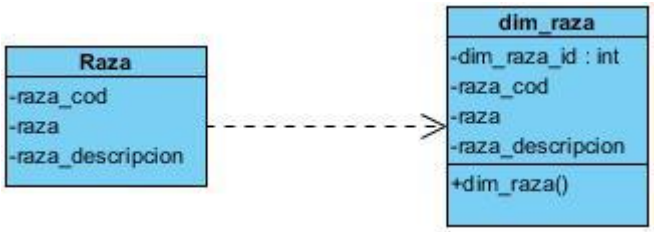


Figura 11: Dimensión Raza

Tablas de hechos

En este paso se define la tabla de hechos, que contiene los hechos a través de los cuales se construirán los indicadores de estudio. Para ello se asigna el nombre a la tabla de hechos que representa la información analizada, luego se define su clave primaria además de la combinación de claves primarias de cada tabla de dimensión relacionada y por último se crean campos de hechos a partir de los indicadores definidos en el modelo conceptual, asignándoles los mismos nombres que presentan.

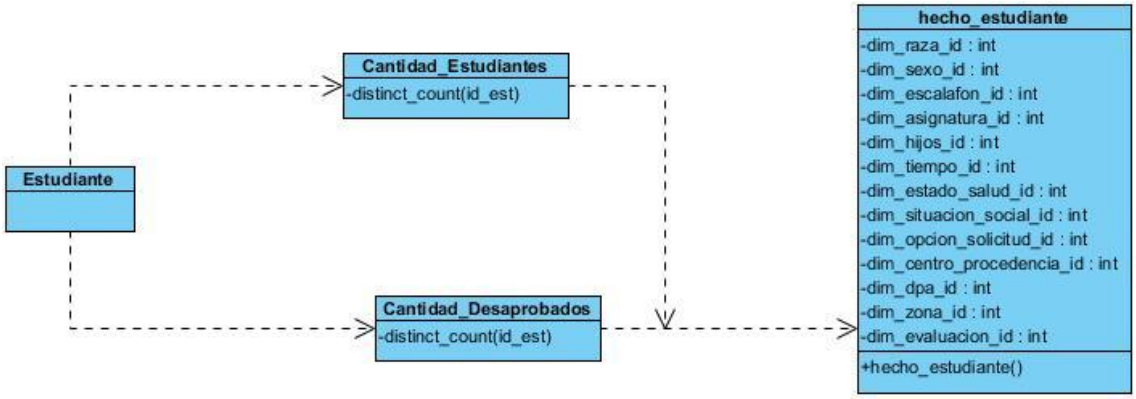


Figura 12: Diseño de la tabla de hechos estudiantes

Uniones:

Se realizan las uniones correspondientes entre las tablas de dimensiones y la tabla de hecho mostrando de forma detallada las relaciones entre ellas. Con esta unión se obtiene el modelo lógico del MD que es mostrado a continuación:

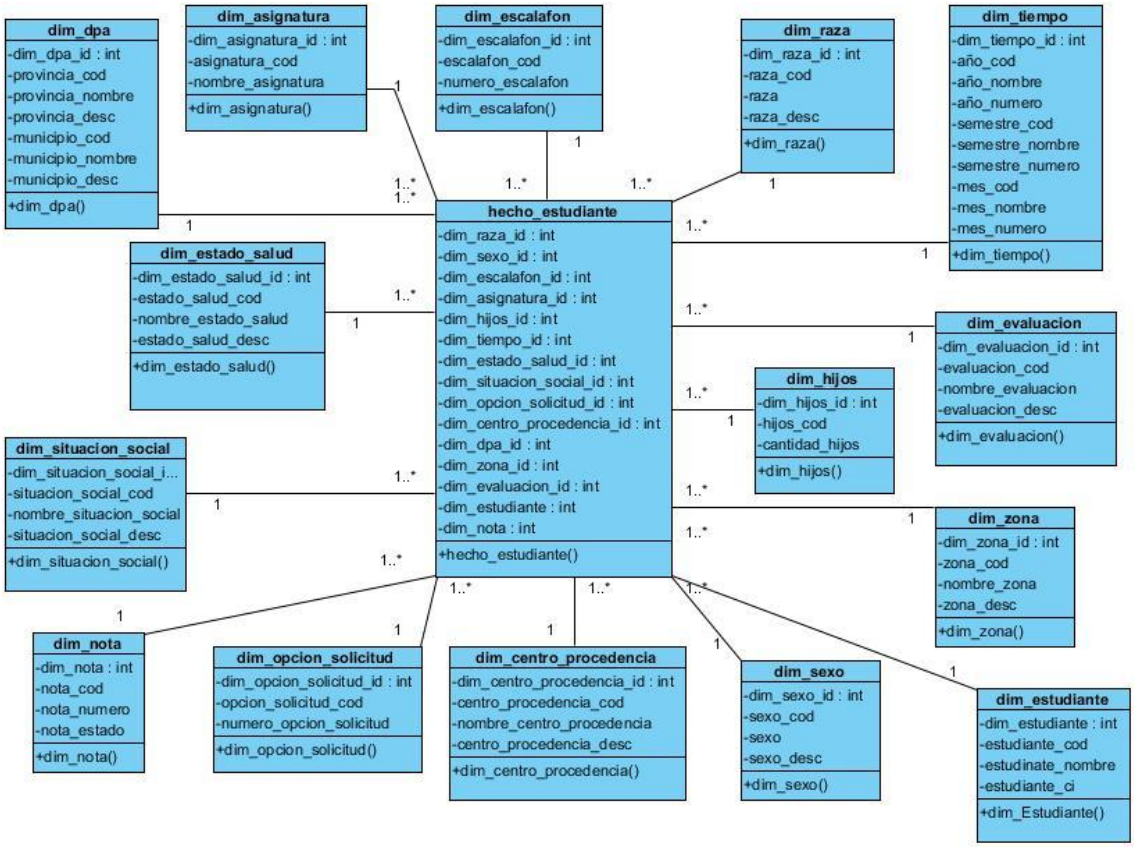


Figura 13: Uniones de las dimensiones con el hecho

Después de obtenido el modelo lógico del MD, se construye el modelo físico del mismo, en donde se especifican los tipos de datos de los diferentes campos del MD. Mediante la herramienta Visual Paradigm se puede exportar este modelo físico a una base de datos, con lo cual se obtiene la base de datos del mercado. A continuación, se muestran el modelo físico obtenido y la estructuración de la base de datos.

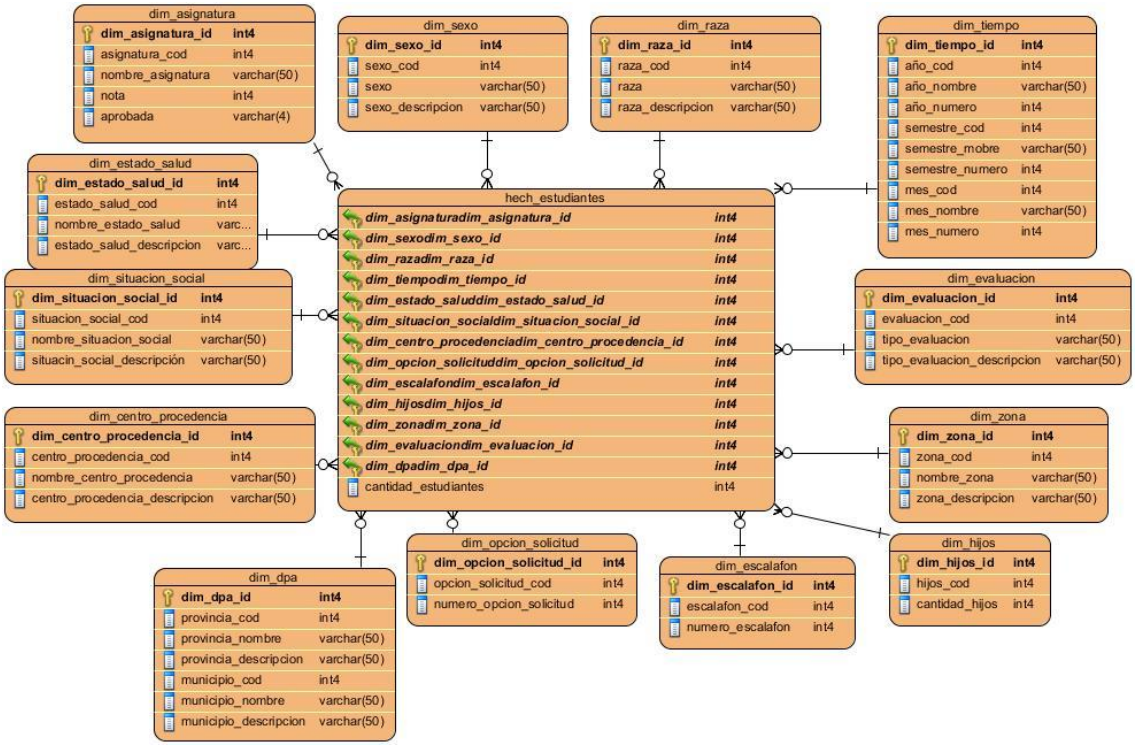


Figura 14: Modelo físico del Mercado de Datos

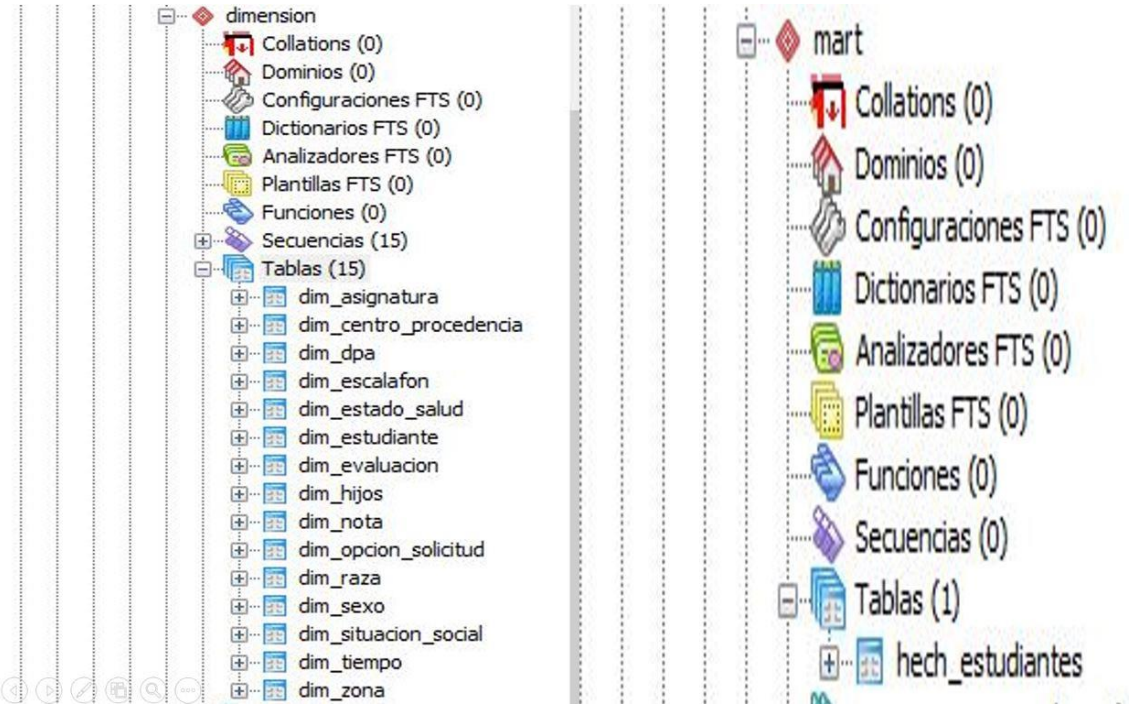


Figura 15: Estructura de la Base de Datos

La estructura que presenta la Base de Datos diseñada cuenta con 2 esquemas, uno llamado “dimensión” en el cual están almacenadas todas las dimensiones del MD y otro esquema llamado “mart” en el cual está contenido la tabla del hecho.

2.3.4. Fase 4: Integración de datos

Luego de haber construido el modelo lógico, se procede a poblarlo con datos, utilizando técnicas de limpieza, transformación y carga, procesos conocidos como ETL.

Carga inicial:

Para realizar la carga inicial se emplea la herramienta antes mencionada PDI, donde se realizan una serie de pasos para poblar el MD, iniciando el proceso a partir de la confección de las transformaciones. A continuación, se expone la transformación Centro_Procedencia (Ver Figura 16: Transformación Centro_Procedencia), para obtener las demás figuras ver Anexos: Transformaciones del MD.

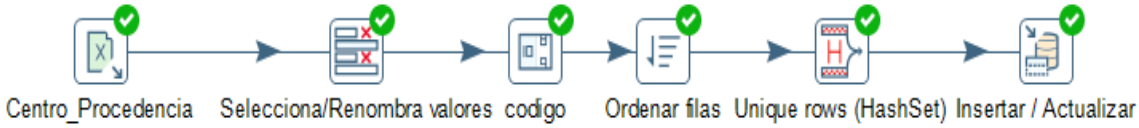


Figura 16: Transformación Centro_Procedencia

Para la carga del hecho se extrae la información de la fuente de datos (Excel) y luego se procede a una búsqueda en la Base de Datos para cargar cada una de las dimensiones. Además, se tuvo que hacer una normalización, lo que permitió convertir las filas en columnas para poder hacer las comparaciones de los datos que se encontraban en la fuente de datos y los ingresados en la Base de Datos. También fue necesaria la utilización de otros componentes como el Seleccionar/Renombrar, Partir Campo, Unión por Clave, entre otros. Esta transformación finaliza con la inserción en la Base de Datos, “ad_estudiantes”, de todos los datos. (Ver Figura 17: Ejecución de la Transformación del hecho).

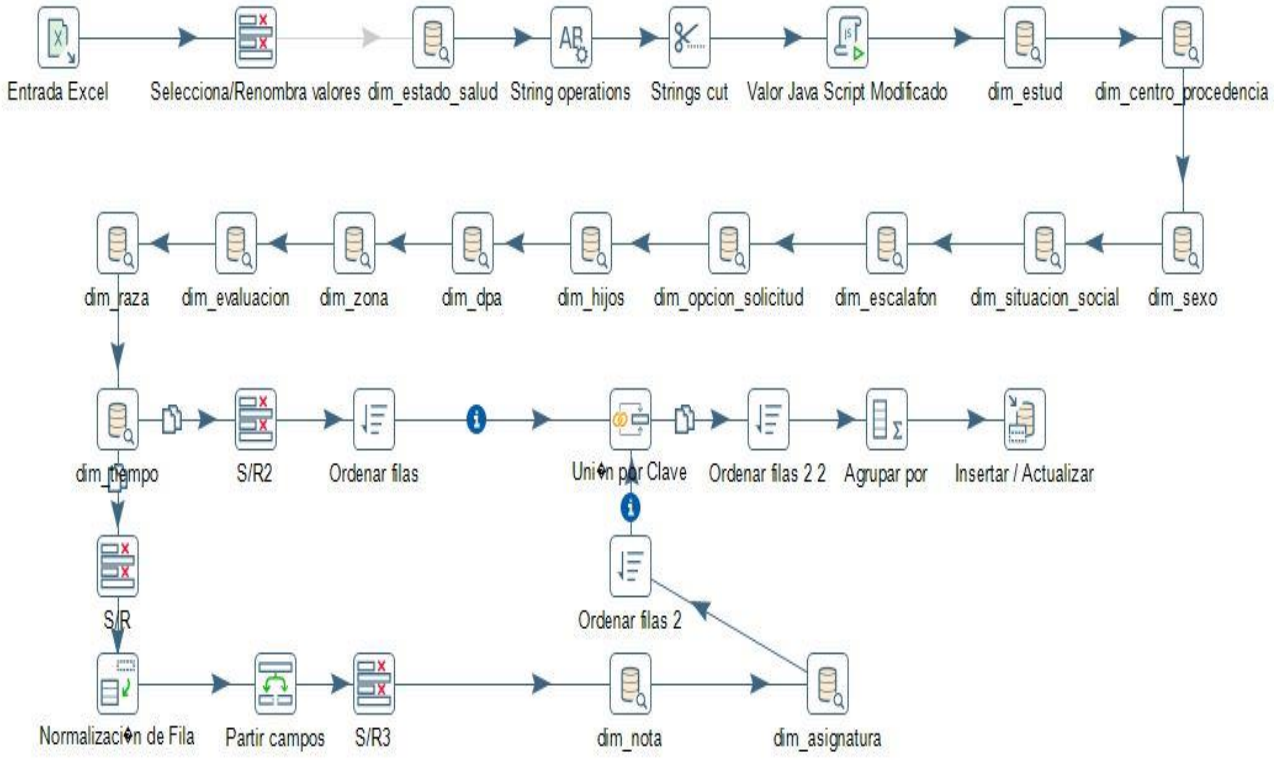


Figura 17: Ejecución de la Transformación del hecho

Conclusiones del capítulo

La selección de la metodología permitió guiar el proceso de desarrollo del mercado de datos. Se logró identificar las preguntas claves del negocio, los indicadores y perspectivas del análisis, lo que permitió elaborar el modelo conceptual de datos para observar el alcance del MD. La confección del modelo lógico del MD para realizar el proceso de extracción, transformación y carga de los datos permitió poblar el MD. El modelo físico definido facilita la comprensión del diseño estructural de la Base de Datos.

Capítulo 3: Visualización y validación

En este capítulo se describe el proceso de desarrollo del cubo OLAP, en los que se definen las dimensiones y medidas tanto físicas como calculables. Se lleva a cabo el proceso de visualización de los datos del MD mediante la herramienta que permite el análisis de los mismos mediante gráficos y tablas. Además, se le practican pruebas al MD para comprobar su rendimiento y valorar el grado de satisfacción del cliente con el resultado que se obtuvo.

3.1: Diseño de los cubos OLAP

El diseño de los cubos OLAP se realiza mediante la herramienta Pentaho Schema Workbench, la cual permite crear visualmente y probar esquemas de cubos OLAP. Además, permite generar un fichero de configuración ".XML", en el cual están definidos los cubos, medidas, tablas y dimensiones. Las dimensiones están definidas por jerarquía que presentan, la tabla para las jerarquías y sus niveles.

Para la presente investigación se modelo 1 cubo, con las características correspondientes a la tabla de hechos y tablas de dimensiones definidas. Esta estructura se puede observar en la siguiente figura:

Figura 18: Diseño del cubo OLAP en la herramienta Schema Workbench

Las medidas son utilizadas para dar respuesta a los intereses de los profesores de la FICI. Esta es agregada al cubo para obtener la cantidad de estudiantes con determinadas características, en dependencia de los que se quiera conocer.

A continuación, se muestra la medida cantidad de estudiantes, que se obtiene mediante la función distinct-count sobre la dimensión estudiantes, donde va mostrando los estudiantes que existen en la Base de Datos sin repeticiones.

Measure for 'hech_estudiantes' Cube	
Attribute	**Value**
name	cantidad_estudiantes
description	
aggregator	distinct-count
column	dim_estudiante_id
formatString	
datatype	Numeric
formatter	
caption	cantidad_estudiantes
visible	☑

Figura 19: Medida cantidad de estudiantes

3.2: Visualización

Una vez finalizado el proceso ETL y todos los datos históricos estén cargados en el mercado y disponibles para ser consultados, se procede a la implementación del subsistema de visualización. Este permite el análisis de los datos acumulados para transformarlos en información que permita generar conocimiento para agilizar el proceso de toma de decisiones. Esta visualización se puede realizar mediante vistas de análisis o reportes operacionales, los que se explicaran a continuación.

3.2.1: Vistas de análisis

Las vistas de análisis las puede crear o consultar el usuario a través de la herramienta BI server una vez que haya publicado el cubo OLAP. A continuación, se muestra la vista de análisis correspondiente a la situación social y zona en la que viven los diferentes estudiantes de la FICI.

					Medidas
dim_situacion_social	dim_zona	Fecha	Notas	dim_estudiante	cantidad_estudiantes
Situacion Social	Zonas	Fechas	All dim_notas	All dim_estudiantes	390
Bien	Zonas	Fechas	All dim_notas	All dim_estudiantes	226
	Rural	Fechas	All dim_notas	All dim_estudiantes	11
	Urbana	Fechas	All dim_notas	All dim_estudiantes	215
Mal	Zonas	Fechas	All dim_notas	All dim_estudiantes	4
	Urbana	Fechas	All dim_notas	All dim_estudiantes	4
Regular	Zonas	Fechas	All dim_notas	All dim_estudiantes	160
	Rural	Fechas	All dim_notas	All dim_estudiantes	8
	Urbana	Fechas	All dim_notas	All dim_estudiantes	152

Figura 20: Vista de análisis del hecho estudiantes

3.2.2: Reportes operacionales

Los reportes creados para analizar la información fueron realizados a través de la herramienta BI Server y el plugin Saiku. Estos brindan la oportunidad al usuario de poder filtrar el reporte según la información que desee analizar. A continuación, se muestra el reporte en forma de tabla que hace referencia a la cantidad de estudiantes matriculados en la asignatura MDII por sexo y nota.

Asignatura	Notas	Sexo	cantidad_estudiantes
MDII	2	Femenino	30
		Masculino	71
	3	Femenino	42
		Masculino	93
	4	Femenino	37
		Masculino	58
	5	Femenino	21
		Masculino	38

Figura 21: Reporte generado por el Saiku

Otra manera de visualizar la información es a través de gráficos. La herramienta Saiku también ofrece esta opción, la cual facilita la comprensión de los datos a analizar por

el usuario. En la figura se observa el reporte presentado anteriormente, pero esta vez en forma gráfica.

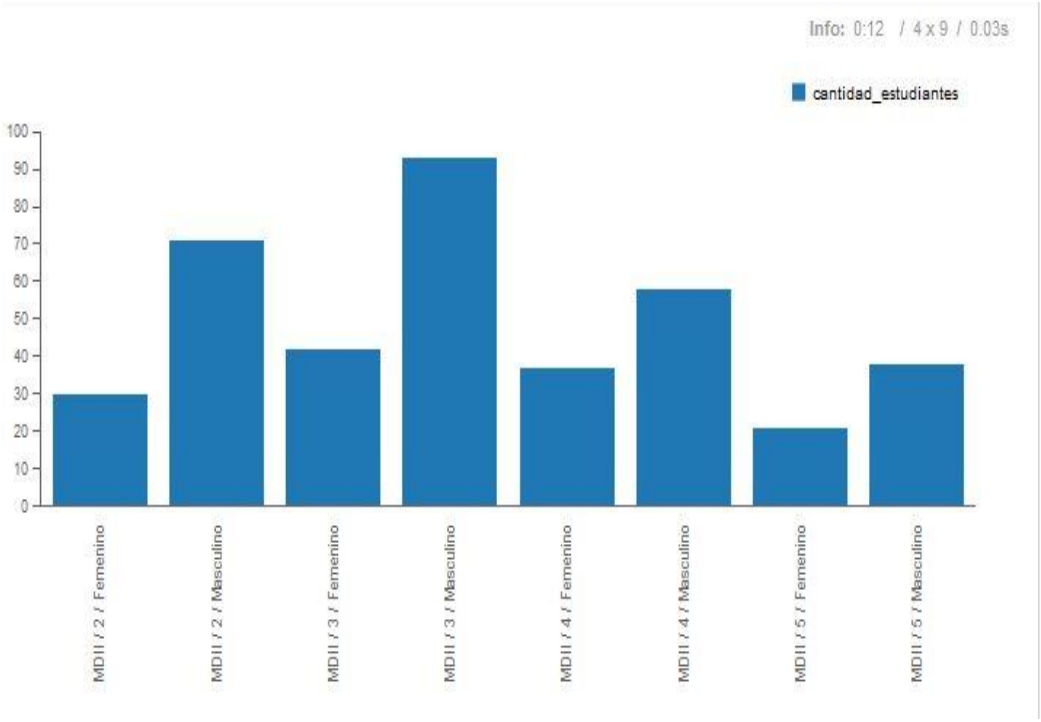


Figura 22: Reporte generado por el Saiku en forma gráfica de barra

Otro grafico generado por la herramienta Saiku referida al mismo reporte es el siguiente:

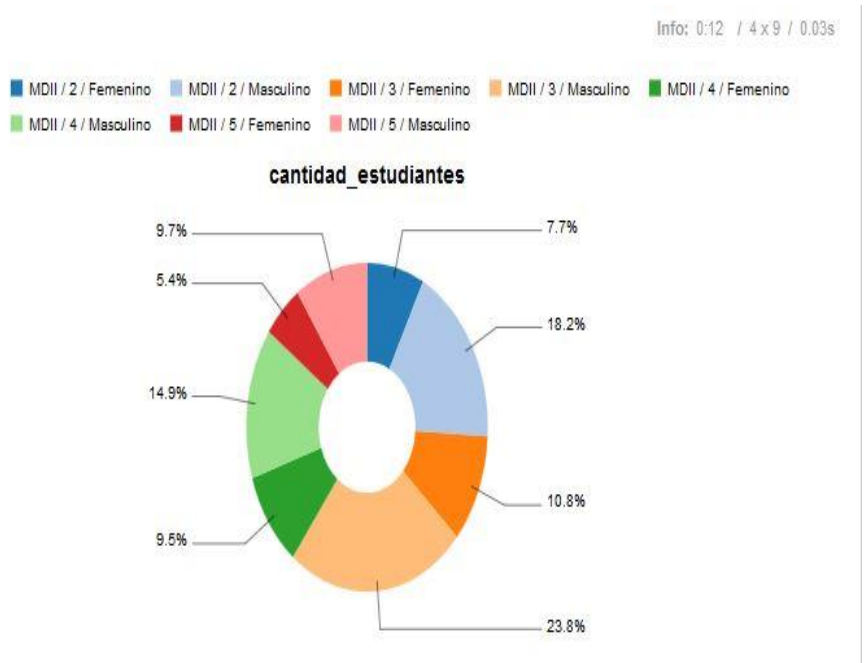


Figura 23: Reporte generado por el Saiku en forma gráfica de pastel

3.3: Pruebas

Todo proceso de creación de software está sujeto a fallos, es por esto que las pruebas de software constituyen una fase importante en el desarrollo de cualquier producto, ya que permiten comprobar que no existan fallos en la implementación del mismo, proporcionándole calidad al software. Para realizar las pruebas necesarias en el desarrollo de la solución se decidió utilizar el modelo V el cual es utilizado por DATEC, para garantizar la calidad del producto. A continuación, se muestra una representación del ciclo de vida en el modelo V. A la izquierda del mismo se puede detallar las etapas de desarrollo del software y a la derecha de este, las pruebas correspondientes a cada etapa. Las pruebas seleccionadas para validar el sistema son:

Figura 24: Modelo en V

Pruebas a utilizar en el proceso de calidad del software, las cuales surgen a partir del modelo en V:

- ✓ Pruebas unitarias: esta prueba centra el proceso de verificación en la menor unidad del diseño del software, o sea, en algún componente del software o módulo.

- ✓ Pruebas de integración: esta prueba construye el sistema a partir de distintos componentes y lo prueba una vez estén todos estos componentes integrados, debe realizarse progresivamente.
- ✓ Pruebas de aceptación: estas pruebas se realizan para comprobar que el sistema cumple con las necesidades del cliente.
- ✓ Pruebas de regresión: estas pruebas consisten en volver a ejecutar un conjunto de pruebas ya ejecutadas anteriormente, de este modo se asegura que los cambios realizados no conducen a errores adicionales.

Herramientas de pruebas

Mediante los casos de prueba el probador podrá determinar si el requisito de una aplicación es parcial o completamente satisfactorio. En el mercado de datos se diseñó un caso de prueba correspondiente a la cantidad de estudiantes con el objetivo de verificar los requisitos de información y de este modo validar el mercado de datos. A continuación, se muestra el caso de prueba:

Escenario	Descripción	Variables de entrada	Variables de salida	Respuesta del sistema	Flujo central
EC1.1 Cantidad de estudiantes matriculados en una asignatura por sexo y raza en un periodo de tiempo determinado.	Muestra la cantidad de estudiantes matriculados en una asignatura	Asignatura Sexo Raza Tiempo	Cantidad de estudiantes	Se muestra la información correspondiente al escenario	1-Se abre la aplicación. 2-Se autentica. 3-Se entra al sistema. 4-Se selecciona el reporte deseado. 5-En el área de trabajo se visualiza el reporte.

Tabla 2: Caso de prueba

Se aplicaron pruebas de aceptación donde se detectaron 6 no conformidades en la primera iteración de las que fueron resueltas 4. En la segunda iteración se detectaron las 2 no conformidades pendientes de la iteración anterior y fueron resueltas de manera satisfactoria. Las no conformidades detectadas fueron que las descripciones de las variables en los casos de prueba no se corresponden con la definición que tienen en la aplicación. Dichas no conformidades fueron resueltas satisfactoriamente.

Finalmente, el sistema es aceptado por el cliente cumpliendo con los requisitos previamente identificados.

Conclusiones del capítulo

La creación de los cubos de información posibilitó la visualización del mercado de datos creado por el equipo de desarrollo. Para mostrar los datos se utilizan vistas de análisis y reportes en forma de tabla y grafico lo que facilita la comprensión del usuario que accede a los mismos. La validación del mercado de datos permitió medir la calidad del sistema y la satisfacción del cliente.

Conclusiones

Después de finalizar la presente investigación, se obtiene el Mercado de Datos para la toma de decisiones a partir de la información de los estudiantes de la FICI cumpliendo con el objetivo trazado, se concluye que:

- ✓ El estudio de los diferentes almacenes de datos y mercado de Datos existentes para el apoyo a la toma de decisiones comprobó que las tecnologías analizadas no cumplen con los requisitos solicitados.
- ✓ El análisis de las metodologías y tecnologías para el desarrollo del Mercado de Datos permitió identificar a Hefesto como una metodología más ágil y fácil de usar para las personas con poca experiencia en el tema.
- ✓ El análisis de los requerimientos posibilitó la obtención del modelo conceptual de Mercado de Datos a partir de la definición de los indicadores y sus perspectivas.
- ✓ El estudio de las pruebas permitió identificar las más adecuadas para comprobar el buen funcionamiento del sistema.

Recomendaciones

Después de haber apreciado los resultados obtenidos y basándose en la experiencia adquirida durante la realización de la investigación y con el propósito de mejorar la propuesta plasmada en este trabajo se recomienda:

- ✓ Actualizar con frecuencia las fuentes de información para una mayor explotación del Mercado de Datos.
- ✓ Disponer de un servidor en la FICI que permita la publicación del Mercado de Datos para que los profesores y directivos tengan la posibilidad de obtener los reportes.
- ✓ Ampliar la creación del Mercado de Datos para los restantes años con el objetivo de tomar decisiones que beneficien el proceso educativo de los estudiantes.

Referencias Bibliográficas

Bernabeu, R. D. (2007). *Data Warehousing: Investigación y Sistematización de conceptos-Hefesto: Metodología propia para la construcción de un Data Warehouse.* Argentina.

Calderón Gómez, H., Díaz Mongui, M. R., & Ariza Nieves, N. J. (2015). *Diseño de herramienta de Inteligencia de Negocios para apoyar la toma de decisiones del área de ventas de un restaurante móvisushi¨SushiTruck¨.* Bogotá, Bogotá, Colombia.

Cuesta, A. (2013). *La Productividad del trabajo del trabajador del conocimiento.*

Gil Soto, E. (2012). *Data Warehouse. Antecedentes, situación actual y tendencias.* Santa Fe de Tenerife.

Guisado Verdezoto, R. M. (2015). *Diseño e Implementación de un Data Mart OLAP para el análisis gerencial académico que será implementado en la unidad educativa «La Colina».* Quito.

Hernández, Y. G. (2013). *Metodología de desarrollo para proyectos de almacenes de datos.* La Habana.

Hormigo, I. G. (2014). *Uso de analítica para dar soporte a la toma de decisiones docentes. In Aplicación De Las TIC Al Proceso De Enseñanza-aprendizaje.* España.

Inmon, W. (2005). *Building the Data Warehouse. Fouth Edition.* New York.

Kimball, I. (2012). Enfoques de desarrollo.

Kimball, R., & Margy, R. (2002). *The Data Warehouse Toolkit. Second Edition.* New York.

Leyet, O. L. (2012). *Analítica de Aprendizaje: Definiciones, Procesos y Potencia. In Seminario de Analítica de Aprendizaje.* La Habana.

Luan, J. (2002). *Data Mining and Knowledge. Management in Higher Education. Potencial Applications.* Toronto.

Matamoros Zapata, R. (2012). *Implantación en una empresa de un sistema Business Intelligence SaaS / On Demand a través de la plataforma LITEBI.* Valencia.

Pentaho BI. (2016).

Pentaho Solution. (2015).

Postgres SQL. (2013).

Pressman, R. (2002). Ingeniería de Software. Un enfoque práctico.

Quilumba, D. E. (2013). *Análisis, Diseño e Implementación de un Data Mart Académico usando tecnología de BI para la Facultad de Ingeniería, Ciencias Físicas y Matemática.* Quito, Quito, Ecuador.

Rico, J. H. (2012). *Rediseño de procesos de gestión de la enseñanza basado en el análisis de datos. Investigación de Operaciones.*

Rico, J. J. (2012). El análisis de datos en apoyo a la gestión de la enseñanza en la carrera Ingeniería Industrial. *Revista Ingeniería Industrial*.

Rico, J. J. (2012). *Modelo Basado En El Análisis De Datos Como Apoyo a La Gestión De La Enseñanza.* La Habana.

Rivadera, G. (2010). *La metodología de Kimball para el diseño de almacenes de datos.*

Rodríguez Sanz, M. (2012). *Análisis y Diseño de un Data Mart para el seguimiento académico de alumnos en un entorno universitario.* Madrid.

Sosa Bello, A., & Salas Lóriga, D. (2013). Data MArt para la toma de decisiones referentes a las Reacciones Adversas a Medicamentos, en el Ministerio de Salud Pública, desde el producto Synta. La Habana.

UCI. (2013). *Modelo Del Profesional y Objetivos Generales.*

Uvidia Fassler, M. I. (2012). *Análisis de Técnicas para Tuning de un Data Warehouse en un sistema de toma de decisiones utilizando Microsoft SQL Server.* Riobamba.

Bibliografía

Aguilar, S. M., & Lemus, J. L. (2014). *Pentaho –BI.*

Bouman, R., & Van Dongen, J. (2012). *Business Intelligence and Data Warehousing with Pentaho and MySQL.*

Brizuela, L., Ismael, E., & Castro Blanco, Y. (2015). Metodologías para desarrollar Almacén de Datos. *Revista De Arquitectura e Ingeniería*.

Casters, M., Bouman, R., & Van Donged, J. (2010). *Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration.*

Dapena Bosquet, I., Muñoz San roque, A., & Sánchez Miralles, Á. (2012). *Sistemas de Información Orientados a la toma de Decisiones: el enfoque multidimensional.* Madrid.

Eckerson, W., & White, C. (2013). *Evaluating ETL and Data Integration Platforms.*

ETL Tools Info - Data warehousing and Business Intelligence. (2017).

Expresiones MDX en Analysis Services. (2017).

Hernández, R. A., & Coello González, S. (2001). *El proceso de investigación científica.*

Hernando Velazco, R. (2016). *Almacenes de Datos (Data Warehouse).*

Imhoff, C., Galemmo, N., & Geiger, J. (2013). *Mastering Data Warehouse Desing.* Canadá.

Importancia de la utilización de un Data Warehouse. (2015).

Kimball, I. (2012). *Enfoques de desarrollo.*

Kimball, R. (2010). *El Juego de Herramientas del Almacén de Datos.*

Kimball, R., & Margy, R. (2013). *The Data Warehouse Toolkit. Third Edition.* New York.

Luis Cano, J. (2014). *Business Intelligence: Competir Con Información.*

Luján Mora, S. (2012). *Diseño de almacenes de datos con UML.*

Mantilla Hernández, J. H. (2012). *Metodología de diseño de cubos olap para la inteligencia de negocios usando Mondrian y Jpivot a partir de una base de datos transaccional.* Bucaramanga.

Manual Admin y utilización de información para decisiones empresariales. (2016).

Méndez, A., & Martire, A. (2012). *Fundamentos de Data Warehouse.* Buenos Aires.

Ponniah, P. (2001). *Data Warehousing Fundamentals.* New York.

Pulvirenti, A. S., & Roldán, M. C. (2012). *Pentaho Data Integration 4 Cookbook.*

Reyes Dixson, Y., & Nuñez Maturel, L. (2015). *La inteligencia de negocio como apoyo a la toma de decisiones en el ámbito académico (Business Intelligence as decision support system in academic environment).* La Habana.

Romina Mateo, L., & Bossero, J. C. (2012). *Utilización de técnicas de Data Warehouse para la toma de decisiones en el Área Académica.* Buenos Aires.

Thomsen, E. (2012). *OLAP Solutions. Second Edition.* New York.

Uvidia Fassler, M. I. (2012). *Análisis de Técnicas para Tuning de un Data Warehouse en un sistema de toma de decisiones utilizando Microsoft SQL Server.* Riobamba.

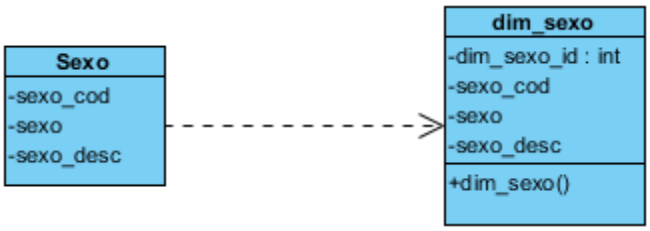
Vitier Urquizu, R. (2013). *Almacén de datos operacional para contribuir a la toma de decisiones.* La Habana.

Anexos

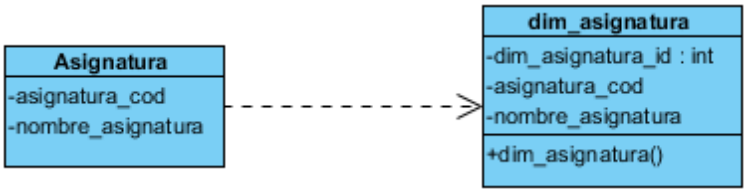
1. Tabla de dimensiones

1.1 Dimensión raza

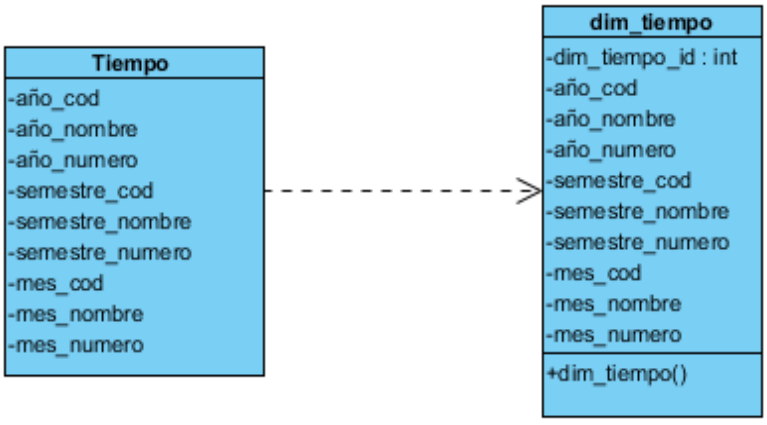
1.2 Dimensión sexo



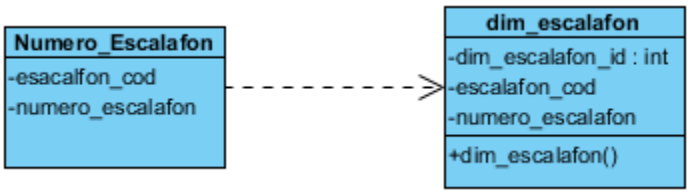
1.3 Dimensión asignatura



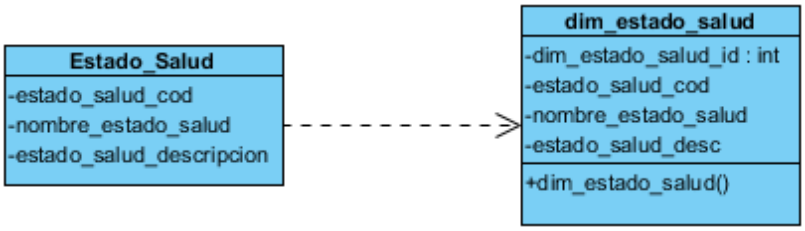
1.4 Dimensión tiempo



1.5 Dimensión Número Escalafón

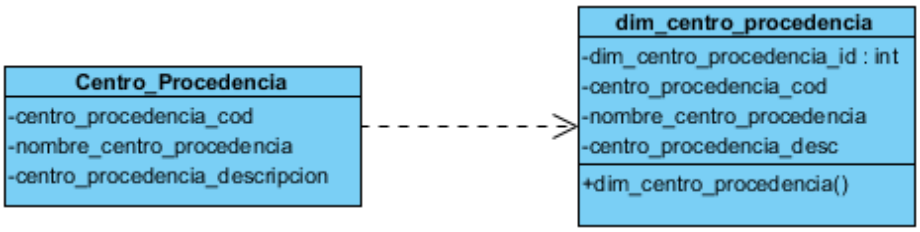


1.6 Dimensión estado de salud

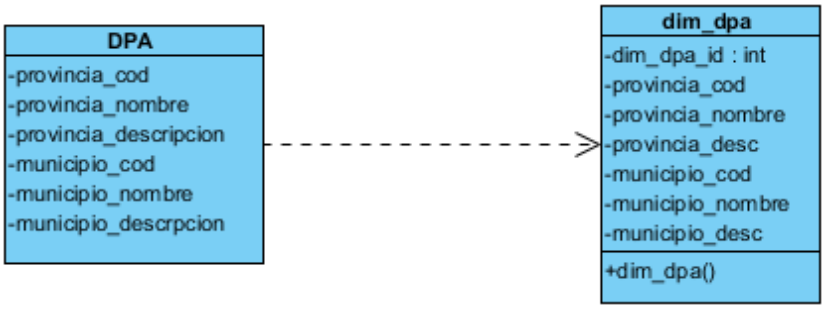


1.7 Dimensión situación social

1.8 Dimensión centro de procedencia

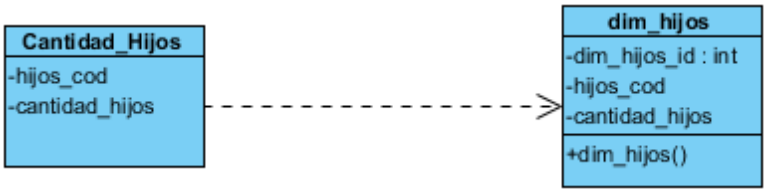


1.9 Dimensión DPA

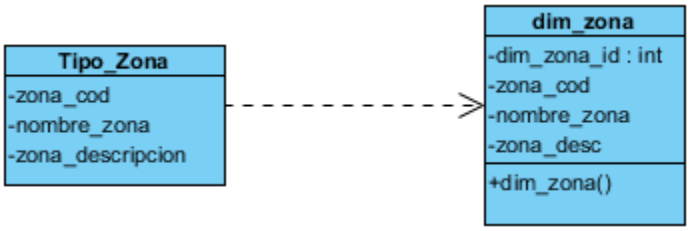


1.10 Dimensión opción solicitud

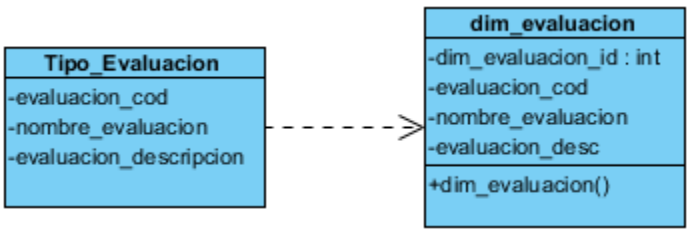
1.11 Dimensión cantidad de hijos



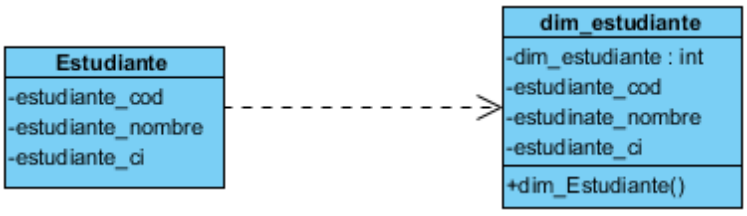
1.12 Dimensión tipo de zona



1.13 Dimensión Evaluación

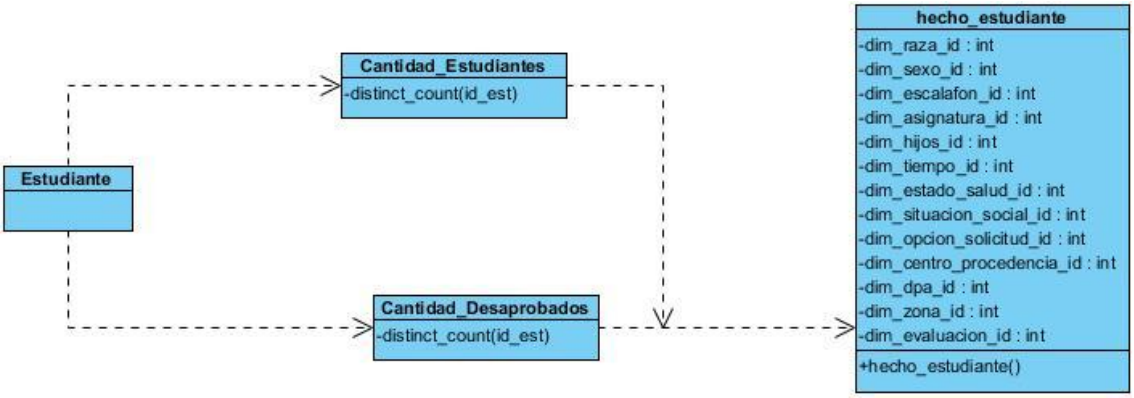


1.14 Dimensión Estudiante



1.15 Dimensión Nota

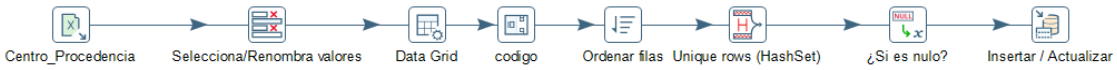
2. Tabla del hecho Estudiantes



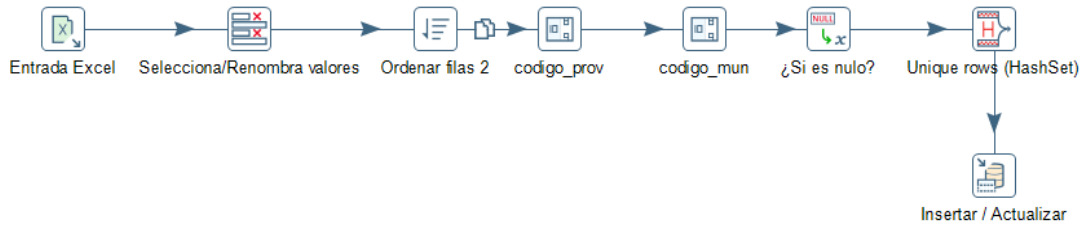
3. Transformaciones del Mercado de datos

3.1 Transformación asignatura

3.2 Transformación centro de procedencia



3.3 Transformación DPA



3.4 Transformación número de escalafón

3.5 Transformación estado de salud

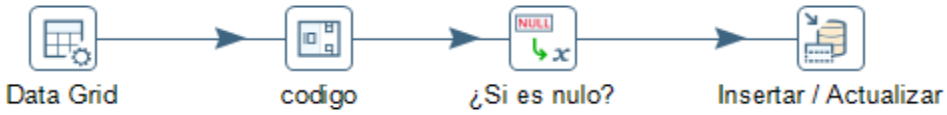
3.6 Transformación cantidad de hijos

3.7 Transformación opción de solicitud

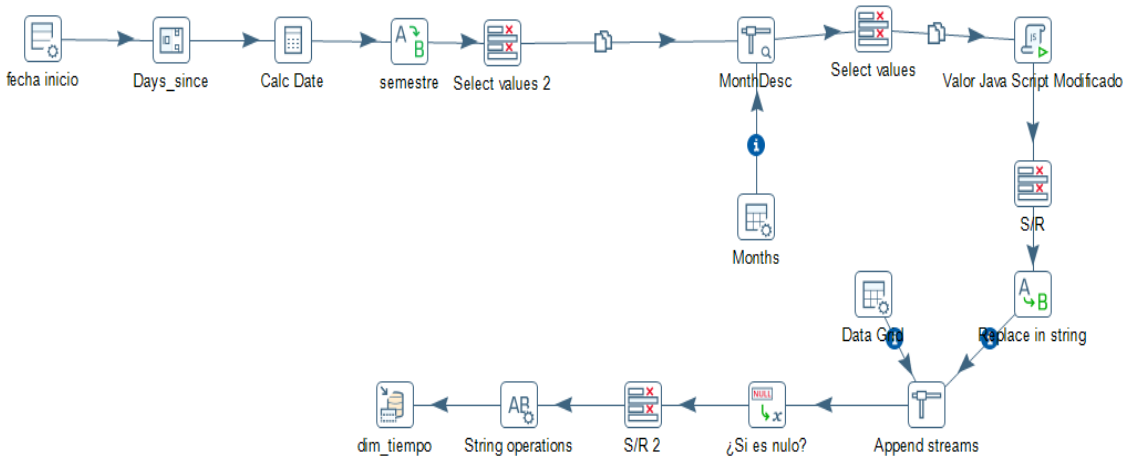
3.8 Transformación raza

3.9 Transformación sexo

3.10 Transformación situación social

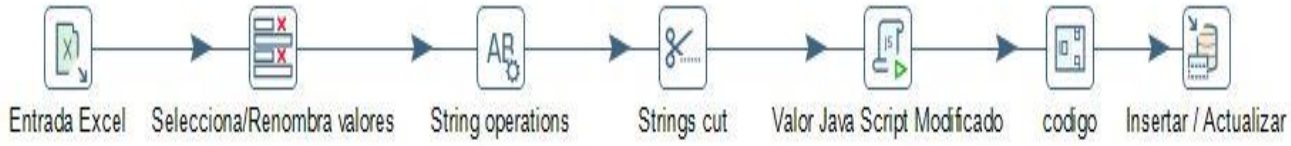


3.11 Transformación tiempo



3.12 Transformación tipo de zona

3.13 Transformación estudiantes



3.14 Transformación notas

4. Carta de aceptación

UCI Universidad de las Ciencias Informáticas

Acta de aceptación

ACTA DE ACEPTACIÓN

Producto: Mercado de Datos para el primer año de la carrera Ingeniería en Ciencias Informáticas.

Involucrados en el proceso:

- **Estudiantes:** Yenisey Silverio Jover.
  Yenlis González Galá.
- **Tutores:** MSc. Yusnier Reyes Dixson
  Ing. Reydel Capote Coipel

Observaciones del proceso:

Las no conformidades detectadas en el proceso de revisión fueron resueltas. Se comprobó la correcta implementación del Mercado de datos para el primer año de la carrera Ingeniería en Ciencias Informáticas. Por tanto, se acepta con fecha de 22 de mayo de 2017 la solución propuesta.

Lista de productos que son aceptados y que deben ser entregados:

- Diseño del Mercado de datos.
- Proceso de extracción, transformación y carga de los datos.
- Implementación de las vistas de análisis OLAP.

Entrega	**Recibe**
Nombre y Apellidos: Yenisey Silverio Jover, Yenlis González Galá	**Nombre y Apellidos:** MSc. Yusnier Reyes Dixson
Cargo: Estudiante	**Cargo:** Profesor
Firma:	**Firma:**

Comentarios: Los productos aceptados deben ser entregados al cliente previo a la defensa.