Universidad de las Ciencias Informáticas

Facultad 4

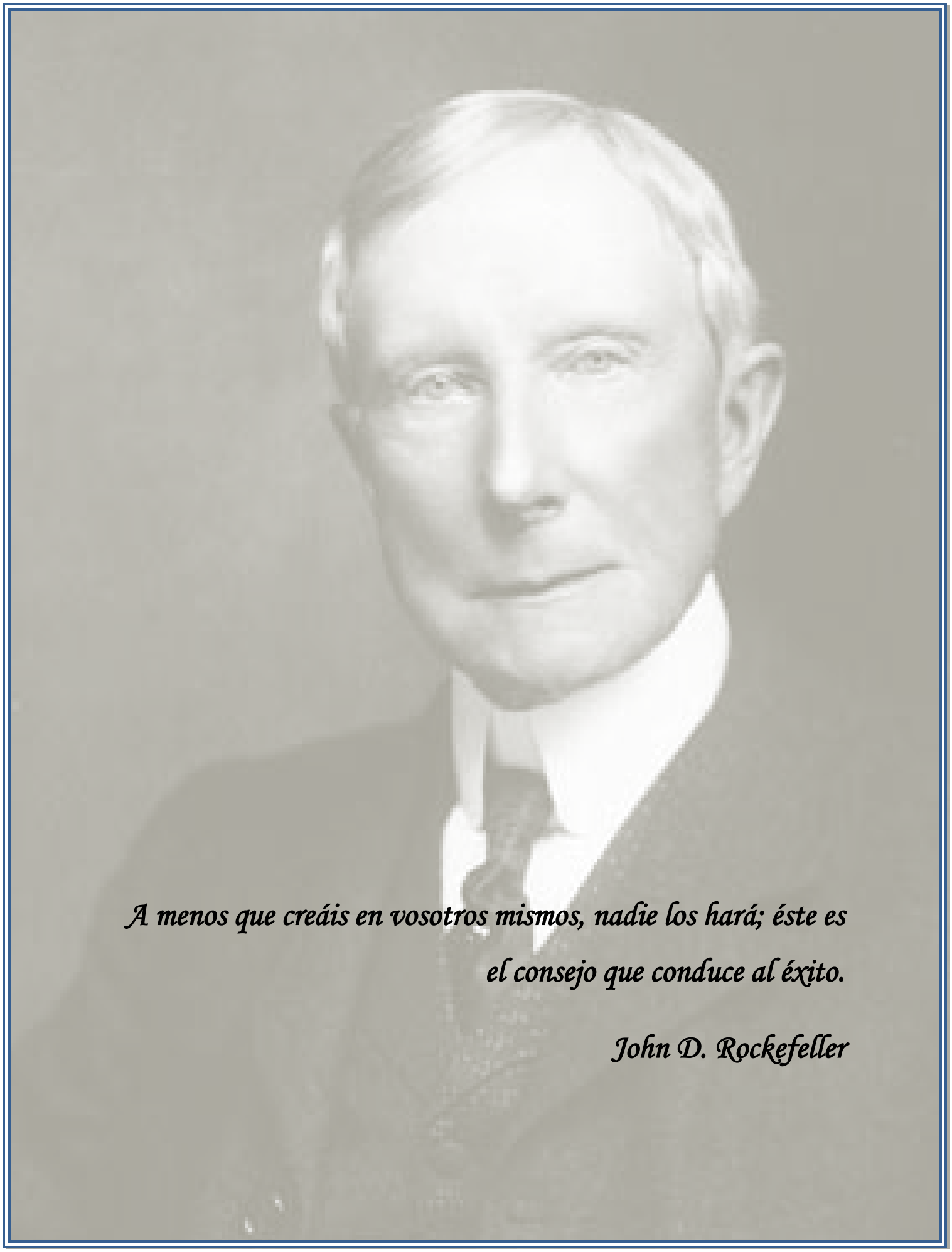
Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas

Título: Almacén de datos operacional para contribuir a la toma de decisiones basado en el uso de los recursos de la Plataforma Educativa ZERA

Autor: Ramel Vitier Urquizu

Tutores: Ing. Arcel Labrada Batista
Ing. Adrián García Sánchez

La Habana, Junio de 2013
"Año 55 de la Revolución"



A menos que creáis en vosotros mismos, nadie los hará; éste es el consejo que conduce al éxito.

John D. Rockefeller

A mi papá René, que ha exigido el máximo de mí en cada etapa de mi vida, espero te sientas muy orgulloso de tu hijo, gracias por cuidarme, educarme y guiarme hasta hacerme la persona que soy hoy. Te amo no lo dudes nunca.

A mi mamá y mi abuelo Antonio, a los que un día les prometí que me verían convertirme en ingeniero en muy poco tiempo, gracias por complacerme en todas mis malcriadeces, aconsejarme y estar ahí cada vez que lo necesité, los adoro.

A toda mi familia, mis padres, abuelos, a mi hermanita Lorena que sabe que la amo. A toda la familia del casino, mi papá Ramel y mis dos hermanos, Wicho y Flavia, los quiero con la vida, gracias por apoyarme y creer en mí.

A mi tío Ernesto por proveerme de cada tecnología que se me antojaba para apoyar mi estudio.

A Yalina, mi reina sabes que sin ti esta tesis no fuera realidad, gracias por entregarme tu cariño, algún que otro día triste, pero eso sí, los 6 meses más maravillosos que pase en esta universidad. Gracias por escogerme, nunca pierdas la Fe.

A mis tutores por tenerme toda la paciencia del mundo y haberme guiado en todo el proceso de construcción de este trabajo, no me pudieron asignar mejores tutores, muchas gracias.

A Nory por compartir conmigo 4 lindos años, en los que me ayudó al máximo, me aconsejó y me transfirió toda su experiencia en esta universidad, gracias a ti llegué hasta aquí.

A mis dos inseparables amigos de universidad, el Eduar y la Daya por acompañarme en este largo trayecto y ayudarme cada vez que me hizo falta.

A todos los amigos que hice en esta linda escuela, el Yoe, la Mary, la Kiry, el Charly, la Susy, Alien, Randy y los muchos otros que no me alcanzarían las páginas para mencionar, espero que nuestra amistad no termine aquí.

A todas las personas que de una forma u otra me ayudaron en la confección de mi tesis, muchas gracias a todos de verdad.

Declaro que soy el único autor de este trabajo y autorizo a la Facultad 4 de la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste, firma la presente declaración jurada de autoría en La Habana a los ____ días del mes _____ del año _____.

_____ _____

Ramel Vitier Urquizu
Autor

Ing. Arcel Labrada Batista
Tutor

_____

Ing. Adrián García Sánchez
Tutor

El surgimiento y desarrollo acelerado de Internet, la generación diaria de importantes datos en áreas científicas, económicas y sociales así como diversas fuentes de información, han servido para el desarrollo de la ciencia, la tecnología y otras áreas. Hoy en día una primicia de las organizaciones, instituciones y empresas es el conocimiento que tengan referente a los procesos de negocio en que están involucradas y a los recursos que ofrecen. De ahí que uno de los mayores retos actualmente sea la capacidad para gestionar e interpretar los datos que en ella se generan obteniendo una ventaja competitiva sólida. Para lograr la manipulación y análisis de este gran cúmulo de información surgen nuevas propuestas tecnológicas como los almacenes de datos, que permiten el almacenamiento en una base de datos que es diseñada para favorecer el análisis y la divulgación de la información y contribuir a la toma de decisiones de las entidades que lo utilizan.

La Plataforma Educativa ZERA desarrollada en el centro FORTES perteneciente a la Facultad 4 de la Universidad de las Ciencias Informáticas (UCI) posibilita la gestión de los procesos de enseñanza aprendizaje. Esta Plataforma Educativa no posee actualmente una herramienta que permita contribuir a la toma de decisiones en cuanto al uso que tienen los recursos publicados así como realizar y visualizar reportes. En la presente investigación se desarrolla una propuesta de almacén de datos para permitir el análisis de la información en la plataforma. Para ello se realiza un estudio de las metodologías, herramientas y procesos asociados a la construcción de estas aplicaciones, obteniéndose un almacén de datos operacional que permita contribuir a la toma de decisiones basada en el uso de los recursos.

Palabras claves: Plataforma Educativa ZERA, Recursos, Toma de decisiones, Almacén de datos.

Introducción

Actualmente con la informatización de la sociedad, ha crecido a nivel mundial la capacidad de generación y almacenamiento de la información que no puede ser analizada por los métodos tradicionales existentes. La cantidad de información que se debe gestionar diariamente es inmensa, Eric Schmidt, ex director general de Google refiriéndose a este suceso expresó: *"entre el nacimiento del mundo y el año 2003, hubo cinco exabytes de información creada. Ahora creamos cinco exabytes cada dos días"***.** (1)

Mientras mayor es la capacidad para almacenar más y más datos, mayor es la incapacidad para extraer información realmente útil de éstos. Las empresas e instituciones se han dado cuenta de la necesidad de procesar su activo más importante, la información, pues esta solo puede ser útil si los datos están ordenados, analizados y transformados en respuestas en tiempo real que puedan resolver problemas específicos. (2)

Mediante la automatización de los datos se puede descubrir información valiosa oculta, la cual se puede aprovechar y convertirse en una oportunidad de negocio, utilizando la información extraída y procesada para elaborar efectivos planes de Inteligencia del Negocio (*Business Intelligence,* ***BI*** por su siglas en inglés), logrando oportunidades de capitalizar con rapidez las ventajas competitivas de una organización y analizar de manera rápida y sencilla la información para la toma de decisiones a nivel operativo, táctico y estratégico.

Con el crecimiento explosivo de Internet ha cambiado la manera de manejar la información. Los grandes volúmenes de datos han llegado a un punto de inflexión creando un valor significativo para la economía mundial, la mejora de la productividad, la competitividad de las empresas y el sector público.

La sociedad ha llegado a una explosión de información que exige la aplicación de nuevas tecnologías para su correcto manejo. Esto conlleva a que aparezcan procesos y tecnologías nuevas que buscan suplir las necesidades de manejo de información existentes. Dentro de estas tecnologías surgen nuevos conceptos para el análisis y manejo de la información: los almacenes de datos, estos proveen un ambiente para que las organizaciones hagan un mejor uso de los datos que manejan, posibilitando un

aumento en su rendimiento, teniendo una información real y oportuna la cual ayuda a la toma de decisiones.

El análisis de los datos para la toma de decisiones no se limita solamente a los sectores económicos, han surgido durante la última década campos mucho más amplios. Por ejemplo en las escuelas, al igual que en otros ámbitos, se han ido incorporando nuevos recursos tecnológicos que ponen de manifiesto la necesidad de reconceptualizar los procesos y modelos tradicionales de enseñanza y aprendizaje.

Con este avance en el mundo de la educación se ha hecho necesario la creación y utilización de sistemas informáticos que ayuden al proceso de aprendizaje, en lo que los estudiantes cada vez más se individualicen y que conlleve de igual manera al interés hacia el estudio contribuyendo a la evolución de los procesos de enseñanza y aprendizaje, para así complementar o presentar alternativas en los procesos de la educación tradicional. (3)

Para ello se han desarrollado los sistemas de gestión de aprendizaje (*Learning Management System,* ***LMS*** por sus siglas en inglés) que comprenden cualquier actividad educativa que utilice medios electrónicos para realizar todo o parte del proceso formativo permitiendo así la capacitación de manera no presencial, eliminando las barreras de tiempo y distancia en el proceso de enseñanza aprendizaje, adecuándose a las habilidades, necesidades y disponibilidades de cada estudiante.

La introducción de los sistemas de gestión de aprendizaje ha generado grandes volúmenes de datos. El análisis de enormes cantidades de trazas dejados por los actores en el proceso de enseñanza aprendizaje puede ser usado para mejorar dinámicamente el desarrollo de analíticas del aprendizaje. La idea fundamental consiste en la interpretación del gran número de datos o señales dejados por los estudiantes, producto de su progreso en el aprendizaje, para evaluar el desempeño académico, predecir actuaciones futuras e identificar los problemas potenciales, en resumen, contribuir a la toma de decisiones de los docentes sobre el proceso de aprendizaje de sus estudiantes. (4)

La Plataforma Educativa ZERA, capaz de adaptarse a los procesos de negocio de diferentes instituciones o escuela, permite la gestión de los hiperentornos de aprendizaje, la visualización de estos, y finalmente la gestión del aprendizaje. Esta plataforma

educativa no cuenta actualmente con una herramienta que permita contribuir a la toma de decisiones. No se puede realizar por tanto un análisis del uso y aceptación de los contenidos que se brindan, por lo que se desconoce cómo enfocar los recursos y a quién dirigirlos. Se dificulta además la realización y visualización de diferentes tipos de reportes que permitan conocer la interacción de los usuarios con los recursos que ofrece la plataforma.

Se plantea como ***problema de la investigación:*** ¿Cómo contribuir a la toma de decisiones basado en el uso de los recursos de la Plataforma Educativa ZERA?

Una vez identificado el problema el ***objeto de la investigación*** se centra en los almacenes de datos. Se especifica como ***campo de acción*** los almacenes de datos operacionales que contribuyen a la toma de decisiones en la Plataforma Educativa ZERA.

Para dar solución al problema de la investigación se define como ***objetivo general*** desarrollar un almacén de datos operacional que contribuya a la toma de decisiones basado en el uso de los recursos de la Plataforma Educativa ZERA.

Los ***objetivos específicos*** son:

- ✓ Efectuar el estudio del estado del arte para la creación de almacenes de datos operacionales.
- ✓ Seleccionar y estudiar las herramientas y metodologías para la creación de un almacén de datos operacional.
- ✓ Realizar el análisis del almacén de datos operacional.
- ✓ Diseñar el almacén de datos operacional.
- ✓ Implementar el almacén de datos operacional.
- ✓ Validar la solución propuesta.

Para guiar la investigación se plantea la siguiente ***idea a defender*****:**

Si se desarrolla un almacén de datos operacional basado en el uso de los recursos de la Plataforma Educativa ZERA se contribuirá la toma de decisiones.

Para dar cumplimiento a los objetivos específicos se planifican las siguientes ***tareas de la investigación:***

1. Elaboración de los fundamentos teóricos sobre el desarrollo de almacenes de datos operacionales.

2. Caracterización de los procesos de análisis, diseño e implementación de almacenes de datos en sistemas que manejan grandes volúmenes de datos.

3. Selección de la metodología a utilizar en el desarrollo del almacén de datos operacional.

4. Estudio y selección de las herramientas que brinden los servicios y las funcionalidades necesarias para construir un almacén de datos operacional.

5. Análisis y diseño del almacén de datos operacional.

6. Implementación y prueba del almacén de datos operacional.

7. Validación de la solución propuesta.

Para el desarrollo de la investigación fueron utilizados métodos científicos, dentro de los que se encuentran los métodos teóricos que permiten estudiar las características que no son observables del modelo de investigación, y los métodos empíricos que facilitan la descripción de las características fenomenológicas del objeto.

Métodos teóricos:

- ✓ El método Histórico-lógico permitió la realización del estudio acerca de los almacenes de datos, las herramientas, metodologías y arquitectura de estos.
- ✓ El método Analítico-sintético posibilitó el análisis, estudio y extracción de conceptos y definiciones relacionadas con el tema.

Métodos empíricos:

- ✓ La observación se utilizó para la recopilación de información importante referente al objeto de estudio y el campo de acción de la investigación.
- ✓ La entrevista se utilizó para conocer si los reportes generados por el almacén de datos operacional permiten obtener información relevante para apoyar la toma de decisiones.

- ✓ La encuesta se empleó con el objetivo de realizar una comparación entre los reportes que se generan actualmente en el módulo de Reportes de la Plataforma Educativa ZERA y los que permite realizar el almacén de datos, dirigida a profesionales que trabajan con la Plataforma Educativa ZERA.

Además de los métodos científicos anteriormente expuestos también se utilizó:

- ✓ El criterio de expertos que permitió obtener opiniones entre diferentes expertos para verificar que los reportes generados por el almacén de datos operacional basado en el uso de los recursos de la Plataforma Educativa ZERA permiten apoyar la toma de decisiones.
- ✓ El estadístico matemático para el manejo y análisis de los datos cualitativos y cuantitativos de la investigación.

Los ***resultados esperados*** son:

Con el desarrollo del presente trabajo de diploma se pretende obtener:

- ✓ Un almacén de datos operacional que permita contribuir a la toma de decisiones basado en el uso de los recursos de la Plataforma Educativa ZERA.
- ✓ Documentación asociada a la investigación y desarrollo de la herramienta.

Estructura capitular:

Capítulo 1: Fundamentación teórica de la investigación

En el capítulo 1 se exponen los elementos teóricos que sustentan el problema científico y los objetivos del trabajo. En este se realiza un estudio de las soluciones similares y se analizan las metodologías y herramientas que se ajustan al desarrollo de la investigación, justificando la selección y utilización de cada una de ellas.

Capítulo 2: Análisis y diseño del almacén de datos operacional

En el capítulo 2 se define el tema de análisis, detallando el funcionamiento interno de la base de datos. Se describe además la definición del negocio centrado en la Plataforma Educativa ZERA, identificando los requisitos funcionales, no funcionales y los requisitos de información. Se diseña la arquitectura que tendrá el almacén de datos operacional. Se describen los casos de uso, así como las tablas de hechos, las medidas y las dimensiones que estructuran el modelo dimensional de la solución. También se realiza la Matriz Bus, el diseño datos dimensional y físico para obtener una mejor visión de las relaciones entre el hecho y las dimensiones.

Capítulo 3: Implementación y prueba del almacén de datos operacional

En el capítulo 3 se describe todo el proceso de implementación y posteriormente las pruebas realizadas al almacén de datos operacional. Se representaran las estructuras de datos definidas así como la implementación del subsistema de integración de datos y el de visualización. Además se explica todo el proceso de extracción, transformación y carga de datos que se realiza. Además se explica cómo se efectuaron las pruebas que consistieron en la realización de una lista de chequeo y los correspondientes casos de pruebas realizados a los casos de uso identificados. También se comprueba que el almacén de datos operacional contribuya a la toma de decisiones, para ello se analiza el criterio de varios expertos en el tema.

Capítulo 1: Fundamentación teórica de la investigación

1.1 Introducción

Este capítulo está dirigido a plantear todos los elementos teóricos que sustentan el objeto de estudio y el objetivo de la investigación. Se llevará a cabo el estudio de sistemas y soluciones similares que logren un eficiente manejo de la información en sistemas con grandes volúmenes de datos. Se relacionan todos los conceptos que desde el punto de vista teórico permiten un mejor entendimiento de lo que se plantea en la situación problemática. También se estudiarán las tecnologías y las herramientas que se emplearán para dar cumplimiento al objetivo general de la investigación así como las metodologías utilizadas para el desarrollo de almacenes de datos, seleccionando la que más se adecúe a las características del almacén de datos que se desea implementar.

1.2 Sistemas similares

En este epígrafe se realizará un estudio del creciente desarrollo de la información y su papel cada vez más determinante en la toma de decisiones de compañías. Se analizarán sistemas que al igual que la Plataforma Educativa ZERA manejan grandes volúmenes de datos y cuáles son las soluciones actuales que estos utilizan para la gestión de la información.

Las empresas que operan dentro de la actual economía digital en la que la información es vasta, está interconectada y automatizada, deben transformar continuamente sus modelos de negocios para mantenerse competitivas en un mundo en el que los eventos y las condiciones económicas globales varían a gran velocidad. Estas por sus procesos de negocio generan un gran cúmulo de información y se ven en la necesidad de buscar soluciones que permitan, la obtención de los datos significativos entre los grandes volúmenes de información que generan utilizando las nuevas tecnologías que surgen para la obtención y manejo de los datos almacenados. Algunas de las instituciones que se pueden citar son: (5)

- ✓ Empresas de telecomunicaciones: Disponen de datos de millones de clientes, llamadas, acciones de marketing, facturas, servicios, etc. Telefónica móviles es un

claro ejemplo de este tipo de compañías además de Jazztel, Vodafone y France Telecom.

- ✓ Empresas de transporte: Aerolíneas, Transporte de Cargas y Transporte de Pasajeros. Entre ellas British Airways, Union Pacific y Air France.
- ✓ Turismo: Centrales de Reservas, Cadenas Hoteleras y Agencias de Viajes.
- ✓ Empresas de fabricación de bienes de consumo masivo: Entre ellas Coca-Cola, Adidas, Nike, 3M, Bosh Siemens y prácticamente todas las empresas de fabricación de automóviles.
- ✓ Entidades Financieras: BBVA, Caja Madrid y Caja Extremadura.
- ✓ Empresas Aseguradoras.
- ✓ Compañía WalMart.
- ✓ Compañía Twentieth Century Fox.
- ✓ Empresa CIMEX.

La necesidad de interactuar con la información de manera óptima, ha conllevado a que aparezcan procesos y tecnologías nuevas que buscan suplir las necesidades de manejo de información existentes. Esto ha posibilitado que se desarrolle el uso de almacenes de datos ya que estos presentan toda la información coherente, organizada y normalizada. Comprenden además una vista de los datos, los cuales pueden ser publicados para que accedan a ellos los usuarios y la información es organizada y presentada al usuario en una forma que le permita fácilmente formular sus propias preguntas.

La mayoría de las instituciones anteriormente mencionadas utilizan este concepto, al igual que la compañía WalMart. Esta cuenta con uno de los almacenes de datos más voluminoso y poblado del mundo el cual usa para tomar decisiones acerca de todos los procesos que realizan en el mercado internacional, elevar su economía y mantenerse en competencia respecto a otras compañías.

Igualmente, Twentieth Century Fox utiliza la información relacionada con las películas que se proyectan en distintos lugares de los Estados Unidos para predecir qué actores, argumentos y filmes serán más populares, con el objetivo de ganar audiencia en sus producciones.

El uso de los almacenes de datos no sólo se limita a los sectores económicos sino que es aplicable al 100% de las áreas fuera de este. Se evidencia su utilización en países como

Venezuela, aplicado al tema de la seguridad ciudadana; así como en hospitales de Perú para la sectorización de pacientes en el consumo de medicamentos.

Cuba no se encuentra indiferente a la aplicación de esta herramienta para la toma de decisiones. La empresa CIMEX, destacada por el crecimiento constante y la estabilidad financiera, tanto dentro como fuera del país, utiliza un almacén de datos para la gestión de inventarios. Además, en la Universidad de las Ciencias Informáticas se ha desarrollado un almacén de datos para la toma de decisiones en cuanto al consumo energético. (6)

La plataforma educativa MudRi ha implementado el uso de almacenes de datos para solucionar los problemas de manejo y análisis de la información. Especialmente en relación con el seguimiento y análisis del uso de la plataforma, incluyendo estadísticas de su usabilidad, nivel de adopción a la plataforma por parte de los estudiantes y profesores, actividades, entre otros. (7)

Por todo lo expuesto anteriormente se ha determinado la utilización de un almacén de datos operacional basado en el uso de los recursos de la Plataforma Educativa ZERA para contribuir a la toma de decisiones en la misma.

1.3 Almacén de Datos

En este epígrafe se definirá que es un Almacén de Datos y cuáles son sus principales características. Se especifican conceptos asociados a este, así como las ventajas asociadas a su uso.

1.3.1 Definición

1. Un Almacén de Datos o Depósito de Datos es una colección de datos orientado a temas, integrado, no volátil, de tiempo variante, que se usa para el soporte del proceso de toma de decisiones gerenciales. (8)

2. Un Almacén de Datos es un conjunto de datos integrados, orientados a un material que varían con el tiempo y que no son transitorios, los cuales soportan el proceso de toma de decisiones de una administración. (9)

3. Ralph Kimball conocido autor en el tema de almacenes de datos lo define como: *"El Almacén de Datos es una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis; es la unión de todos los Data Marts de una entidad".* (10)

4. Bill Inmon: *"Un Almacén de Datos es una colección de datos orientados al tema, integrados, no volátiles e historiados, organizados para el apoyo de un proceso de ayuda a la decisión".* (11)

Después de haber estudiado las definiciones anteriores de almacenes de datos se arriba a la conclusión de que los autores concuerdan en que su principal función es contribuir a la toma de decisiones basado en la información histórica de las empresas, además de ser una colección de datos orientado al tema, integrado, variable en el tiempo y no volátil.

1.3.2 Características principales

Orientados al tema: 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: Los datos deben de ser consistentes siempre dentro del almacén de datos e integrados de distintas fuentes de datos operacionales. Algunos ejemplos de integración de los datos son:

- ✓ **Medida de atributos**: Los diseñadores de aplicaciones miden las unidades de medida en una variedad de formas. Un diseñador almacena los datos de longitud en centímetros, otros en pulgadas, otros en metros y otros en kilómetros. Al dar medidas a los atributos, la transformación traduce las diversas unidades de medida usadas para transformarlas en una medida estándar común.
- ✓ **Fuentes Múltiples**: El mismo elemento puede derivarse desde fuentes múltiples. En este caso, el proceso de transformación debe asegurar que la fuente apropiada sea usada, documentada y movida al almacén de datos.

No volátil: La manipulación de datos en el almacén de datos es mucho más simple que en el ambiente operacional. Hay dos únicos tipos de operaciones: la carga inicial de datos

y el acceso a los mismos. Los datos almacenados no se modifican ni se eliminan nunca, solo se añaden nuevos datos.

De tiempo variante: Toda la información del almacén de datos es requerida en algún momento. Esta característica básica de los almacenes de datos, es muy diferente de la información encontrada en el ambiente operacional. En éstos, cuando se accede a una unidad de información, se espera que los valores requeridos se obtengan a partir del momento de acceso. Como la información en el almacén de datos es solicitada en cualquier momento, los datos encontrados en el depósito se llaman de tiempo variante.

Las variantes del tiempo se pueden notar de tres formas:

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

1.3.3 Ventajas del uso de los almacenes de datos

En este epígrafe de la investigación se presentaran las principales ventajas de los almacenes de datos:

- ✓ Transforma datos orientados a las aplicaciones en información orientada a la toma de decisiones.
- ✓ Integra y consolida diferentes fuentes de datos en una única plataforma sólida y centralizada.

- ✓ Provee la capacidad de analizar y explotar toda la información que posee.
- ✓ Permite reaccionar rápidamente a los cambios del mercado.
- ✓ Aumenta la competitividad en el mercado.
- ✓ Mejora la entrega de información, es decir, información completa, correcta, consistente, oportuna y accesible.
- ✓ Facilita la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén; obteniendo un valor añadido para el negocio de dicha información.
- ✓ Los usuarios pueden tener a su disposición una gran cantidad de información multidimensional, presentada coherentemente como fuente única, confiable y disponible en sus estaciones de trabajo.
- ✓ Proporciona la capacidad de aprender de los datos del pasado y predecir situaciones futuras en diversos escenarios.

1.3.4 Almacén de Datos Operacional

Según Ralph Kimball, un almacén de datos operacional es un almacén de información detallada orientado a temas, integrado, aumentado con frecuencia, dentro del Almacén de Datos de una empresa. (10)

Bill Inmon plantea**,** un almacén de datos operacional es una colección de datos orientada a temas, integrada, volátil, actualizada, sólo detallada, que sustenta las necesidades de información reciente, operacional, integrada y colectiva de la organización. (11)

Después de haber estudiado las diferentes definiciones de almacenes de datos operacionales se llega a la conclusión de que un almacén de datos operacional es una colección de datos integrados, detallados y actuales para soportar la toma de decisiones tácticas.

1.3.5 Conceptos principales asociados a los almacenes de datos

Con el objetivo de proporcionar al lector un mejor entendimiento de los temas que serán abordados en la investigación, se describen a continuación una serie de conceptos asociados a los almacenes de datos.

Modelo Multidimensional:

Realizar el Modelo Dimensional es una técnica que permite modelar bases de datos simples y entendibles al usuario final. La idea fundamental es que el usuario visualice fácilmente la relación que existe entre las distintas componentes del modelo.

El modelo multidimensional le permite a los analistas y diseñadores más flexibilidad en el diseño, para lograr un mayor desempeño y optimizar la recuperación de la información, desde un punto de vista más cercano al usuario. Es una disciplina de diseño que se sustenta en el modelo entidad-relación (MER) y en datos numéricos.

Modela las particularidades de los procesos que ocurren en una organización, dividiéndolos en mediciones y entorno. Las medidas son en su mayoría, medidas numéricas, y se les denomina hechos. Alrededor de estos hechos existe un contexto que describe en qué condiciones y en qué momento se registró este hecho.

En contraste con las bases de datos relacionales, las multidimensionales están compuestas por dimensiones que son atributos estructurales de un cubo, organizadas con jerarquías de categorías que describen los datos en tablas.

De manera general, un modelo multidimensional provee dos conceptos principales: medida y dimensión. Las dimensiones son fundamentalmente textos descriptivos mientras que las medidas son en su mayoría, medidas numéricas, y se les denomina hechos. (12)

En general, la estructura básica de un almacén de datos para el Modelo Multidimensional está definida por dos elementos, esquemas y tablas.

- ✓ Tablas de los almacenes de datos: como cualquier base de datos relacional, un almacén de datos se compone de tablas. Hay dos tipos básicos de tablas en el Modelo Multidimensional:

- Tablas Fact: contienen los valores de las medidas de negocios, por ejemplo: ventas promedio en pesos, número de unidades vendidas, etc.

- Tablas Lock_up: contienen el detalle de los valores que se encuentran asociados a la tabla Fact.

✓ Esquemas de los almacenes de datos: la colección de tablas en el almacén de datos se conoce como Esquema. Los esquemas caen dentro de dos categorías básicas: esquemas estrellas y esquemas copo de nieve.

OLTP (Online Transaction Proccesing):

Un OLTP, representa toda aquella información transaccional que genera la empresa en su accionar diario, además, de las fuentes externas con las que puede llegar a disponer. Estas fuentes de información, son de características muy disímiles entre sí, en formato, procedencia, función, etc. (13)

Entre los OLTP más habituales que pueden existir en cualquier organización se encuentran:

✓ Archivos de textos.
✓ Hipertextos.
✓ Hojas de cálculos.
✓ Informes semanales, mensuales, anuales, etc.
✓ Bases de datos transaccionales.

OLAP (Online Analitical Proccesing):

El término OLAP, define a una tecnología que se basa en el análisis multidimensional de los datos y que le permite al usuario tener una visión más rápida e interactiva de los mismos.

Este análisis, también conocido como análisis del hipercubo, organiza la información según los parámetros que se consulten, de manera tal que a partir de estructuras multidimensionales que contienen los datos resumidos de sistemas transaccionales, conocidos como OLTP o de grandes bases de datos, se obtendrá la información requerida. (14)

Data Mart:

Un Data Mart es un almacén de datos limitado a un área concreta de la organización. Muchos expertos definen el almacén de datos como un almacén centralizado que contiene una serie de Data Marts. Es además un modelo multidimensional basado en tecnología OLAP que representa a un área específica de la empresa, incluyendo las variables claves y los indicadores para el proceso de toma de decisiones. Su enfoque es el cumplimento de los requerimientos específicos de un determinado grupo de usuarios en términos de análisis, contenido, presentación y facilidad de uso. Los usuarios de un Data Mart pueden tener datos que se presentan en términos que le son familiares.

También se puede decir que un Data Mart es una base de datos departamental, que se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento. Es un subconjunto del almacén de datos usado normalmente para el análisis parcial de los datos. El objetivo de subdividir está dado por la complejidad computacional del análisis global de todas las dimensiones del almacén de datos y por la necesidad de rapidez. (15)

Área de almacenamiento temporal:

Es un área temporal donde se recogen los datos que se necesitan de los sistemas origen. Se recogen los datos estrictamente necesarios para las cargas, y se aplica el mínimo de transformaciones a los mismos. No se aplican restricciones de integridad ni se utilizan claves, los datos se tratan como si las tablas fueran ficheros planos. De esta manera se minimiza la afectación a los sistemas origen, la carga es lo más rápida posible para minimizar la ventana horaria necesaria, y se reduce también al mínimo la posibilidad de error. Una vez que los datos están traspasados, el almacén de datos se independiza de los sistemas origen hasta la siguiente carga. Lo único que se suele añadir es algún campo que almacene la fecha de la carga.

Obviamente estos datos no van a dar servicio a ninguna aplicación de reportes, son datos temporales que una vez hayan cumplido su función serán eliminados, de hecho en el esquema lógico de la arquitectura muchas veces no aparece, ya que su función es meramente operativa. (16)

1.4 Herramientas

1.4.1 Herramientas para BI

Pentaho BI: Pentaho BI ofrece una amplia gama de herramientas orientadas a la integración de información y al análisis inteligente de los datos de una organización. Cuenta con potentes capacidades para la gestión de procesos de extracción, transformación y carga (***ETL,*** por sus siglas en ingles), informes interactivos, análisis multidimensionales de información o minería de datos. Todos estos servicios están integrados en una plataforma web, en la que los usuarios pueden consultar la información de una manera fácil e intuitiva.

Los módulos incluidos por Pentaho BI, pueden utilizarse de manera conjunta o de forma separada según las necesidades de la organización. Las soluciones de Pentaho están escritas en Java y tienen un ambiente de implementación también basado en este lenguaje. Esto hace que Pentaho sea una solución muy flexible para cubrir una amplia gama de necesidades empresariales tanto las típicas como las sofisticadas y especificas al negocio. (17)

Los módulos de la plataforma Pentaho BI son:

- ✓ **Reporting:** *Pentaho Reporting* es una solución basada en el proyecto JFreeReport y permite generar informes ágil y de gran capacidad. Permite la distribución de los resultados del análisis en múltiples formatos. Todos los informes incluyen la opción de imprimir o exportar a formato PDF, XLS, HTML y texto. Los reportes de Pentaho permiten también programación de tareas y ejecución automática de informes con una determinada periodicidad.

- ✓ **Analysis:** *Pentaho Analysis* suministra a los usuarios un sistema avanzado de análisis de información. Con el uso de las tablas dinámicas, el usuario puede navegar por los datos, ajustando la visión de estos, los filtros de visualización, añadiendo o quitando los campos de agregación. Los datos pueden ser representados en forma de SVG (gráficos de vector escalables), *Flash*, *dashboard widget*, o también integrados con los sistemas de minería de datos y los portales web.

- ✓ **Dashboard:** todos los componentes del módulo *Pentaho Reporting* y *Pentaho Analysis* pueden formar parte de un *Dashboard*. En *Pentaho Dashboard* es muy fácil incorporar una gran variedad en tipos de gráficos, tablas y velocímetros e integrarlos con los portales web, en donde se podrá visualizar informes, gráficos y análisis OLAP.

- ✓ **Data Mining:** Mediante *Pentaho Data Mining s*e podrá descubrir patrones de comportamiento e indicadores ocultos en la información de una organización. Prevenir eventos futuros basados en patrones históricos para así apoyar las tareas de análisis predictivo.

- ✓ **Integración de Datos:** se realiza con la herramienta para ETL *Pentaho Data Integration (**PDI**, por sus siglas en inglés*) que permite implementar los procesos de extracción, transformación y carga de la información.

Requisitos mínimos de Pentaho BI:

- ✓ Memoria RAM: 1Gb.
- ✓ Espacio en disco duro: 1Gb.
- ✓ Procesador: Celeron 2.0 GHz.
- ✓ Necesita un JDK de java instalado con anticipación, se recomienda el JDK de Sun 1.5 o superior.
- ✓ Se necesita también los drivers JDBC de la base de datos relacional que se utilizará como fuente de datos.

Spago BI: Se trata de una aplicación BI de tipo OLAP construida para acceso web y que permite acceder a datos de SQL Server y Mondrian. Es una plataforma ya que cubre y satisface todos los requisitos de BI, tanto en términos de análisis, de gestión de datos, administración y seguridad.

En el mundo analítico ofrece soluciones para la presentación de informes, análisis multidimensional, minería de datos, tableros de mando y consultas ad-hoc. Añade módulos originales para la gestión de procesos de colaboración. Cuenta con herramientas para ETL y apoya al administrador en el mantenimiento de los documentos analíticos, la gestión para el control de versiones y la aprobación del flujo de trabajo. Permite generar

informes perfectamente estructurados y exportarlos a multitud de formatos (HTML, PDF, XLS, XML, TXT, CSV y RTF) además es multiplataforma y tiene licencia GNU LGPL. (18)

Estructura modular: (19)

- ✓ Spago BI Server: núcleo central de Spago BI que integra la funcionalidad de los diferentes motores.
- ✓ Spago BI Studio: entorno de desarrollo único e integrado.
- ✓ Spago BI Meta: entorno enfocado a la capa de metadatos.
- ✓ Spago BI SDK: nivel de integración para utilizar Spago BI con aplicaciones externas.
- ✓ Spago BI Applications: para mantener los modelos verticales de análisis desarrollados con Spago BI.

Requisitos mínimos de Spago BI:

- ✓ Memoria RAM: 512Mb.
- ✓ Servidor de aplicaciones J2EE como *Tomcat, JBoss, WebSphere*.

1.4.2 Herramientas para el proceso de ETL

PDI: Desarrollado íntegramente en Java, posee licencia LGPL. Se utiliza para la integración de datos, carga de almacenes de datos y Data Marts, limpieza de datos, análisis perfilado de datos, migración de datos entre base de datos y exportar datos de bases de datos a archivos planos. Transforma e integra datos entre sistemas de información existentes y los Data Marts que compondrán el sistema BI. Posee como principales características: (20)

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

Incluye cuatro herramientas:

- ✓ SPOON: para diseñar transformaciones ETL usando un entorno gráfico.
- ✓ PAN: para ejecutar transformaciones diseñadas con Spoon.
- ✓ CHEF: permite diseñar la carga de datos incluyendo un control de estado de los trabajos.
- ✓ KITCHEN: permite ejecutar los trabajos *batch* diseñados con *CHEF*.

Soporta diferentes fuentes de información como son: Excel, PostgreSQL, MySql, Informix, dBaseIII, IVo5, FirebirdSQL, IBMDB2, MSSQLServer, MSAccess, Oracle, SAPERPSystem, Teradata, LucidDB, Hypersonic y ApacheDerby.

Talend Open Studio: Talend Open Studio tiene como principal ventaja que está implementado en Java, por lo tanto dispone de un entorno de desarrollo multiplataforma. Además, puede usar Java como lenguaje de apoyo en las tareas de transformación de datos, y se pueden crear nuevos componentes usando esté lenguaje. Por último, todas las operaciones se hacen de forma visual; Talend las transforma en código Java, que compila y entrega en forma de un archivo .jar y un script .sh o .bat; para poder ejecutarlo desde Linux, Windows o Mac. Permite también de forma visual conectar las fuentes de datos con el sistema de destino, transformando los datos mediante componentes ya creados en la aplicación.

Talend cuenta con una gran cantidad de componentes, y con una comunidad que trabaja añadiendo nuevas opciones. En cuanto a bases de datos, se pueden encontrar desde las más generales como: MySQL, SQL Server, Oracle o PostgreSQL, a aquellas con aplicaciones más específicas como Grenplum, ParAccel o eXists. También dispone de componentes para adquirir o volcar datos utilizando ficheros de diferentes tipos: XML, Excel, delimitados (csv, tsv), JSON; e incluso la posibilidad de capturar las filas mediante expresiones regulares. (21)

Scriptella: Es una herramienta ETL basada en Java y una herramienta de secuencias de comandos de ejecución. El lenguaje de programación principal es una antigua llanura de SQL ejecutadas por el puente JDBC. Al mismo tiempo, otros proveedores de JDBC no pueden ser fácilmente añadidos a la mezcla que permite secuencias de comandos SQL con otros lenguajes de *scripting*. (22)

Ventajas:

- ✓ Soporta diferentes tipos de bases de datos con el conector JDBC que incluye el software.
- ✓ Dado el tipo de lenguaje es fácil de entender.
- ✓ Integra Java EE, Ajax, JNDI, Java Mail, Spring Framework (MVC).
- ✓ Permite la ejecución de *scripts* tales como JavaScript, JEXL, Java.
- ✓ Multiplataforma.

Jitterbit: Es una plataforma que facilita la integración de aplicaciones, datos y sistemas empresariales, disponible bajo licencia Open Source y, alternativamente, bajo licencia comercial. Ofrece una herramienta gráfica que permite definir fácilmente mapeos de datos, lo que reduce al mínimo la necesidad de programación. Desde el punto de vista técnico, Jitterbit cuenta con una arquitectura y rendimiento avanzado y está diseñado para poder escalar en caso de que el proyecto así lo requiera. (23)

Algunos ejemplos de integración que pueden realizarse con esta herramienta son:

- ✓ Integrar aplicaciones empresariales (CRM[1], ERP[2]) con bases de datos corporativas.
- ✓ Compartir datos transacciones que provienen de un ERP con aplicaciones corporativas propietarias.
- ✓ Creación de procesos automáticos que implican la participación de procesos heterogéneos a lo largo de la compañía.
- ✓ Integrar servicios web externos con los sistemas *back-end* de la empresa.
- ✓ Consolidar datos corporativos para ofrecerlos mediante un nuevo servicio web.
- ✓ Integrar datos de *partners* y terceras partes dentro de los sistemas existentes en la empresa
- ✓ Consolidación y refinamiento de datos provenientes de múltiples fuentes

Desde un punto de vista técnico, la herramienta ofrece las siguientes características:

1. Soporta los protocolos HTTP, HTTPS, FTP, SFTP, SOAP, ODBC y Windows File Share

[1] CRM (*Customer Relationship Management*, en español Software para la administración de la relación con los clientes)
[2] ERP (*Enterprise Resource Planning*, en español Sistemas de planificación de recursos empresariales).

2. Soporta los formatos XML Schema, DTD, WSDL, ficheros planos, ficheros jerárquicos.

3. Permite integrar llamadas a bases de datos, tanto de consulta como de manipulación de datos.

4. Herramienta gráfica de integración (mapeo entre sistemas, transformaciones gráficas de datos, planificación y ejecución de procesos y tareas, notificaciones y transacciones de integración).

5. Jitterpacks (documentos XML portables que definen todo lo necesario para completar una integración estándar).

1.4.3 Comparación de las herramientas para el proceso de ETL

Las herramientas evaluadas intentan satisfacer o cumplir los parámetros necesarios para la creación de almacenes de datos. Todas ofrecen amplia conectividad con varias fuentes, brindan un abundante conjunto de transformaciones. Permiten también, que el usuario pueda definir sus propias funciones y brindan medios con los que se pueda corregir los errores y registrar eventos durante el flujo de datos. Para un mejor análisis se comparan las dos herramientas más potentes y reconocidas para el proceso de ETL, para así poder elegir la herramienta a utilizar y respaldar el porqué de la selección.

Tabla 1: Comparación entre PDI y Talend Open Studio. (24)

Parámetros	PDI	Talend Open Studio
Facilidad de uso	X	
Extensible	X	X
Reusabilidad	X	X
Ficheros planos	X	X
Ficheros XML	X	X
Ficheros EXCEL	X	X
Bases de datos Microsoft Access	X	
Bases de datos DB2	X	X
Bases de datos Oracle	X	X
Bases de datos SQL Server	X	X

Sistemas ERP (SAP)	X	X
Motor OLAP		
Gestión de dimensiones	X	X
Gestión de la calidad de datos	X	X
Gestión de la sustitución de claves	X	
Perfil de datos		
Manipulación de errores	X	X
Registro de eventos	X	X
Planificación de tareas	X	X
Ejecución automáticas de tareas	X	X
Gestión de metadatos	X	X
Soporte para ejecución paralela de tareas	X	X

Luego de analizar las herramientas asociadas al concepto de almacenes de datos se seleccionan Pentaho BI y Pentaho Data Integration porque son fáciles de utilizar, gestionan además la sustitución de claves. PDI está desarrollada bajo el prisma de las problemáticas de procesos ETL y transformación de datos y son las que más se ajustan a las necesidades del proyecto, además de ser herramientas empleadas por la universidad.

1.5 Metodologías para el diseño e implementación de un almacén de datos

1.5.1 Metodología Hefesto

Hefesto es una metodología cuya propuesta está fundamentada en una amplia investigación, comparación de metodologías existentes y experiencias propias en procesos de desarrollo de almacenes de datos. Está orientada a la construcción de almacenes de datos para análisis dimensional. (25)

Características

Las principales características de la metodología Hefesto son: (26)

- ✓ Los objetivos y resultados esperados en cada fase se distinguen fácilmente y son sencillos de comprender.
- ✓ Se basa en los requisitos de los usuarios, 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 a los usuarios finales en cada etapa para que tome decisiones respecto al comportamiento y funciones del almacén de datos.
- ✓ Utiliza modelos conceptuales y lógicos, los cuales son sencillos de interpretar y analizar.
- ✓ Es independiente del tipo de ciclo de vida que se emplee para contener la metodología.
- ✓ Es independiente de las herramientas que se utilicen para su implementación.
- ✓ Es independiente de las estructuras físicas que contenga el almacén de datos y de su respectiva distribución.
- ✓ Cuando se culmina con una fase, los resultados obtenidos se convierten en el punto de partida para llevar a cabo el paso siguiente.
- ✓ Se aplica tanto para almacenes de datos como para Data Mart.

Comprende las siguientes fases: (27)

1. Análisis de requerimientos.
   - ✓ Identificar preguntas.
   - ✓ Identificar indicadores y perspectivas.
   - ✓ Modelo conceptual.

2. Análisis de los OLTP.
   - ✓ Conformar indicadores.
   - ✓ Establecer correspondencias.
   - ✓ Nivel de granularidad.
   - ✓ Modelo conceptual ampliado.

3. Modelo lógico del almacén de datos.
   - ✓ Tipo de modelo lógico del almacén de datos.
   - ✓ Tablas de dimensiones.
   - ✓ Tablas de hechos.

- ✓ Uniones.

4. Integración de datos.
   - ✓ Carga inicial.
   - ✓ Actualización

Análisis de requerimientos

Lo primero que se hará es identificar los requerimientos de los usuarios a través de preguntas que expliquen los objetivos de su organización. Luego, se analizarán estas preguntas a fin de identificar cuáles serán los indicadores y perspectivas que serán tomadas en cuenta para la construcción del almacén de datos. Finalmente se confeccionará un modelo conceptual en donde se podrá visualizar el resultado obtenido en este primer paso.

Análisis de los OLTP

Se analizarán las fuentes OLTP para determinar cómo serán calculados los indicadores y para establecer las respectivas correspondencias entre el modelo conceptual creado en el paso anterior y las fuentes de datos. Luego, se definirán qué campos se incluirán en cada perspectiva. Finalmente, se ampliará el modelo conceptual con la información obtenida en este paso.

Modelo lógico del Almacén de Datos

Se confecciona el modelo lógico de la estructura del almacén de datos, teniendo como base el modelo conceptual que ya ha sido creado. Para ello, primero se definirá el tipo de modelo que se utilizará y luego se llevarán a cabo las acciones propias al caso, para diseñar las tablas de dimensiones y de hechos. Finalmente, se realizarán las uniones pertinentes entre estas tablas.

Integración de datos

Una vez construido el modelo lógico, se deberá proceder a poblarlo con datos, utilizando técnicas de limpieza y calidad de datos así como los procesos ETL; luego se definirán las

reglas y políticas para su respectiva actualización, así como también los procesos que la llevarán a cabo.

1.5.2 Metodología de Kimball

La metodología se basa en lo que Kimball denomina Ciclo de Vida Dimensional del Negocio. Este ciclo de vida del proyecto del almacén de datos, está basado en cuatro principios básicos: (28)

- ✓ Centrarse en el negocio: hay que concentrarse en la identificación de los requisitos del negocio y su valor asociado, y usar estos 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: diseñar una base de información única, integrada, fácil de usar, de alto rendimiento donde se reflejará la amplia gama de requisitos de negocio identificados en la empresa.

- ✓ Realizar entregas en incrementos significativos: crear 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: proporcionar todos los elementos necesarios para entregar valor a los usuarios de negocios. Se necesita tener un almacén de datos sólido, bien diseñado, con calidad probada, y accesible. También se deberá entregar herramientas de consulta, aplicaciones para informes y análisis avanzado, capacitación, soporte, sitio web y documentación.

La construcción de una solución de almacén de datos 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:

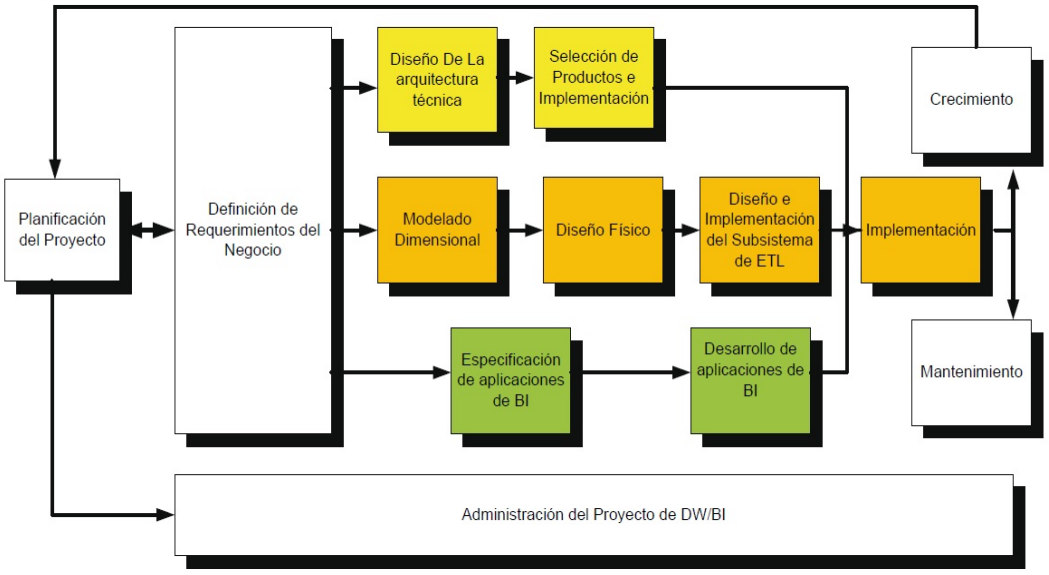


Figura 1: Metodología de Kimball.

Planificación del Proyecto

La planificación busca identificar la definición y el alcance del proyecto de almacén de datos, incluyendo justificaciones del negocio y evaluaciones de factibilidad. La planificación del proyecto se focaliza sobre recursos, perfiles, tareas, duraciones y secuencialidad.

Definición de los Requerimientos del Negocio

Los usuarios finales y sus requisitos impactan siempre en las implementaciones realizadas de un almacén de datos. Según la perspectiva de Kimball, los requisitos del negocio se posicionan en el centro del universo del almacén de datos. Los requisitos del negocio deben determinar el alcance del almacén de datos (qué datos debe contener, cómo debe estar organizado, cada cuánto debe actualizarse, quiénes y desde dónde accederán).

Modelado Dimensional

La definición de los requerimientos del negocio determina los datos necesarios para cumplir los requerimientos analíticos de los usuarios. Diseñar los modelos de datos para

soportar estos análisis requiere un enfoque diferente al usado en los sistemas operacionales.

Diseño Físico

El diseño físico de las base de datos se focaliza sobre la selección de las estructuras necesarias para soportar el diseño lógico. Se intentará contestar las siguientes preguntas:

- ✓ ¿Cómo puede determinar cuán grande será el sistema del almacén de datos?
- ✓ ¿Cuáles son los factores de uso que llevarán a una configuración más grande y más compleja?
- ✓ ¿Cómo se debe configurar el sistema?
- ✓ ¿Cuánta memoria y servidores se necesitan?
- ✓ ¿Qué tipo de almacenamiento y procesadores?
- ✓ ¿Cómo instalar el software en los servidores de desarrollo, prueba y producción?
- ✓ ¿Qué necesitan instalar los diferentes miembros del equipo de almacén de datos en sus estaciones de trabajo?
- ✓ ¿Cómo convertir el modelo de datos lógico en un modelo de datos físicos en la base de datos relacional?
- ✓ ¿Cómo conseguir un plan de indexación inicial?
- ✓ ¿Debe usarse la partición en las tablas relacionales?

Diseño e implementación del subsistema de ETL

El sistema de ETL es la base sobre la cual se alimenta el almacén de datos. Si el sistema ETL se diseña adecuadamente, puede extraer los datos de los sistemas de origen de datos, aplicar diferentes reglas para aumentar la calidad y consistencia de los mismos, consolidar la información proveniente de distintos sistemas, y finalmente cargar la información en el almacén de datos en un formato acorde para la utilización por parte de las herramientas de análisis.

Implementación

La implementación representa la convergencia de la tecnología, los datos y las aplicaciones de usuarios finales accesibles a los usuarios del negocio.

1.5.3 Rapid Warehousing Methodology

Rapid Warehousing Methodology es una metodología iterativa basada en el desarrollo incremental del proyecto de almacén de datos. Establece que son siete las fases para el desarrollo de un almacén de datos, estas son: (29)

1. **Evaluación y definición de objetivos.**

Esta fase es crucial para determinar si la empresa está lista para emprender el proyecto de almacén de datos y los alcances del proyecto, los cuales deben ser realistas enfocados a lo que la empresa desea con respecto al almacén de datos.

2. **Requerimientos.**

El levantamiento de los requerimientos en esta fase va desde un alto nivel, hasta el análisis de los documentos y fuentes que existen en la empresa para conocer de dónde proviene la información y qué áreas de la empresa son las que lo producen. También en esta fase se reconocen los requerimientos del negocio y los técnicos.

3. **Diseño y modelización.**

Se identifican las fuentes de los datos (sistema operacional, fuentes externas) y las transformaciones necesarias para que, a partir de dichas fuentes, se obtenga el modelo lógico de datos del almacén de datos. El modelo lógico se traduce posteriormente en el modelo físico de datos que se almacenará en el almacén de datos.

4. **Construcción.**

Durante esta fase, el equipo de desarrollo inserta los datos el almacén de datos, por medio de la extracción y transformación, que provienen de las diversas fuentes. Se establece la carga de los datos y con qué periodicidad se realiza para la actualización del almacén. Se utilizan o generan aplicaciones para la explotación de los datos para que pasen por un proceso de revisión que permite determinar si es correcto lo establecido en el modelado y diseño del almacén de datos; entrando en esta fase en un proceso de retroalimentación interactivo para la mejora del proyecto.

5. **Prueba Final.**

Un equipo independiente de aseguramiento de la calidad prueba el sistema antes de ser entregado al cliente. Esto permite verificar que los requerimientos funcionales establecidos dentro del ambiente del proyecto del almacén de datos son alcanzados.

6. **Explotación y Despliegue.**

Esta fase asegura que los usuarios estén bien entrenados, que las aplicaciones y los datos son realmente accesibles para que ayuden a promover por completo la aceptación del proyecto total del almacén de datos.

7. **Revisión.**

Después de la fase de construcción y para asegurar el proceso de implementación, se aprende de los aciertos obtenidos y de los fracasos para que se entre en un ciclo de mejora que va desde la primera construcción del almacén de datos a la fase de explotación y despliegue para verificar que los datos sigan siendo disponibles y útiles para los usuarios y en caso de errores, fallas o mejoras se vuelva a las primeras etapas de desarrollo del almacén de datos.

1.5.4 Metodología de desarrollo para proyectos de almacenes de datos

La Metodología de Desarrollo de Proyectos de Almacenes de Datos toma como base la metodología de Kimball para definir los aspectos específicos del desarrollo de almacenes de datos. Para incorporar los principios básicos que permiten una adecuada gestión del proyecto, utiliza la Guía para los Fundamentos de la Dirección de Proyectos. Los temas asociados a CMMI se incorporan a partir del Programa de Mejora por lo tanto hereda algunos de sus enfoques, artefactos y actividades (30). La figura siguiente muestra el ciclo de vida de la metodología.

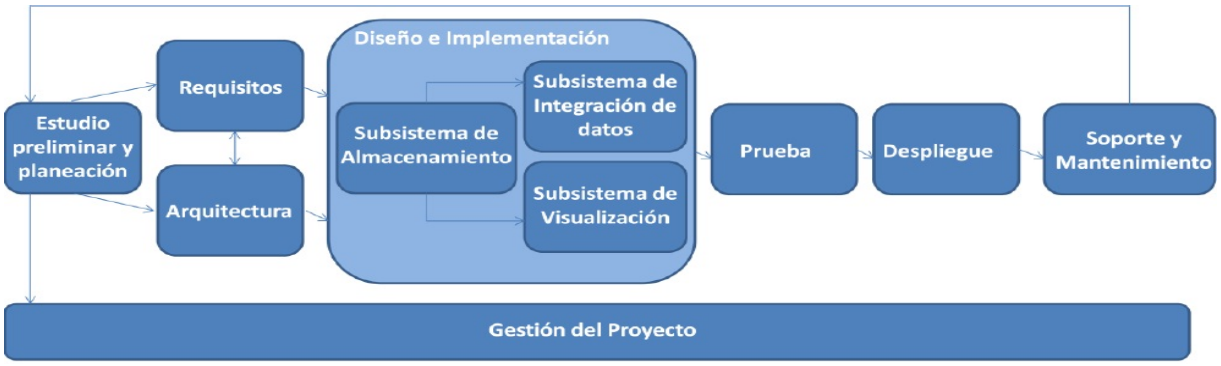


Figura 2: Ciclo de vida de la metodología.

Gestión del Proyecto

Constituye un flujo de trabajo que se ejecuta a lo largo de todo el ciclo de vida del proyecto. Está compuesto por un grupo de procesos que se encargan de mantener la adecuada gestión del proyecto a partir de la aplicación de conocimientos, habilidades, herramientas y técnicas.

Estudio preliminar y planeación

La fase se compone por dos procesos, el estudio preliminar y la planeación inicial del proyecto. El estudio preliminar consiste en hacer un diagnóstico integral de la organización dividido en tres áreas, diagnóstico del negocio, diagnóstico de los datos y diagnóstico de la infraestructura tecnológica. Con los resultados del diagnóstico se hace un estudio de factibilidad que permita estimar los costos de desarrollo, con el fin de establecer el monto del presupuesto que se necesita para desarrollar el proyecto. Además estos resultados son de vital importancia para las fases de Requisitos y Arquitectura, ya que establecen los aspectos iníciales que se deben tener en cuenta. Además se realizan las tareas de planeación inicial del proyecto, para ello se definen un grupo de aspectos importantes relacionados con la gestión de proyectos como son: alcance del proyecto, riesgos, calidad del producto, recursos humanos, adquisiciones, cronograma, entregables, costos y presupuesto.

Requisitos

Esta fase se divide en dos procesos claves, el levantamiento de requisitos, donde se identifican todos los requerimientos de la solución y el análisis de los requisitos definidos,

que permite identificar las estructuras bases del modelo lógico dimensional. El levantamiento de requisitos consiste en identificar las necesidades de información de la organización, las características y cualidades que debe poseer el sistema. En las soluciones de almacenes de datos se identifican tres tipos de requisitos, requisitos de información, requisitos funcionales y requisitos no funcionales.

Arquitectura

En esta fase se define los aspectos arquitectónicos de la solución. Para complementar esta fase se recomienda utilizar los principios contenidos en la Guía base para la especificación de arquitecturas de software, con sus nueve vistas de la arquitectura, agrupadas en tres áreas principales, vistas de arquitectura de sistema, vistas de arquitectura de tecnología, vistas de arquitectura de infraestructura.

Diseño e Implementación

En esta etapa se obtiene el producto de software, se diseñan e implementan los tres subsistemas que conforman el almacén de datos. Cada subsistema puede verse como un componente de software que se desarrolla de forma independiente, para luego ser integrados conformando el producto final. A pesar de esto existen restricciones de dependencia entre los componentes que definen el orden en que deben ser desarrollados. Para poder diseñar e implementar los subsistemas de integración de datos y visualización de información es necesario que esté implementado el subsistema del repositorio de datos. Cumplida esa condición pueden desarrollarse de forma paralela los otros dos subsistemas mencionados. Durante la realización de esta fase es muy común encontrar inconsistencias o problemas en los requisitos identificados que pueden provocar cambios o desviaciones en el proyecto. Esto debe ser analizado cuidadosamente por todo el equipo de desarrollo para definir las acciones a seguir y evitar un impacto negativo en el desarrollo del proyecto.

Prueba

En esta fase se realizan las pruebas necesarias para validar la calidad del software una vez implementado el mismo. Este no es el único momento donde se realizan pruebas al sistema, durante su desarrollo la metodología propone que se utilice el modelo V. Este

modelo relaciona las etapas de análisis y diseño con la ejecución de pruebas, definiendo en qué etapa del desarrollo se deben definir las pruebas que permiten validar el resultado de la misma. Esto sirve de gran ayuda a los desarrolladores que durante la ejecución de las pruebas sabrán exactamente qué fase hay que revisar para corregir el error detectado. Las pruebas que se realizan son, pruebas unitarias, pruebas de integración y pruebas de sistema (validación).

1.5.5 Selección de la metodología

Al realizar un análisis de las metodologías para el desarrollo de almacenes de datos se determina utilizar la Metodología de Desarrollo de Proyectos de Almacenes de Datos definida por especialistas del centro DATEC de la Facultad 6 de la UCI para el desarrollo del almacén de datos operacional de la Plataforma Educativa ZERA. Pues la misma define claramente los procesos y actividades que deben realizarse para el adecuado desarrollo de un almacén de datos, incluyendo las tareas asociadas a la gestión del proyecto y el aseguramiento de la calidad. Los procesos propuestos permiten la gestión integral del desarrollo del proyecto, se adaptan a las características de este tipo de soluciones y sirven de guía a los líderes de proyectos que cuentan con escaza experiencia en el tema. Los procesos y actividades propuestas permiten implementar las buenas prácticas definidas por CMMI en su nivel 2. El ciclo de vida propuesto por la metodología es flexible y puede ser adaptado al ambiente de cualquier organización que desarrolle almacenes de datos. (30)

Existen muchos aspectos asociados a esta metodología que abarcan la gestión de un proyecto de mucha más envergadura que el que se realiza en la presente investigación, por lo que se decidió incorporar solo aquellos procesos que aportan al desarrollo del almacén y no los asociados a la gestión de proyectos.

1.6 Sistema Gestor de Base de Datos

Un Sistema Gestor de Bases de Datos (SGBD) es una colección de programas cuyo objetivo es servir de interfaz entre la base de datos, el usuario y las aplicaciones. Se compone de un lenguaje de definición de datos, de un lenguaje de manipulación de datos y de un lenguaje de consulta. Los SGBD permiten definir los datos a distintos niveles de

abstracción y manipular dichos datos, garantizando la seguridad e integridad de los mismos.

Se decide utilizar PostgreSQL 9.1 pues es un SGBD objeto-relacional, distribuido bajo licencia BSD y con su código fuente disponible libremente. Es el SGDB de código abierto más potente del mercado.

PostgreSQL 9.1 utiliza un modelo cliente/servidor y usa multiprocesos en vez de multihilos para garantizar la estabilidad del sistema. Un fallo en uno de los procesos no afectará el resto y el sistema continuará funcionando. Funciona muy bien con grandes cantidades de datos y una alta concurrencia de usuarios accediendo a la vez al sistema.

A continuación se exponen algunas de las características más importantes y soportadas por PostgreSQL 9.1: (31)

Generales:

- ✓ Replicación sincrónica/asincrónica.
- ✓ Copias de seguridad.
- ✓ Unicode.
- ✓ Juegos de caracteres internacionales.
- ✓ Regionalización por columna.
- ✓ *Multi-Version Concurrency Control* (MVCC).
- ✓ Múltiples métodos de autentificación.
- ✓ Acceso encriptado vía SSL.
- ✓ Disponible para Linux y UNIX en todas sus variantes (AIX, BSD, HP-UX, SGI IRIX, Mac OS X, Solaris, Tru64) y Windows 32/64bit.

Programación y desarrollo:

- ✓ Funciones y procedimientos almacenados en numerosos lenguajes de programación, entre otros PL/pgSQL, PL/Perl, PL/Python y PL/Tcl.
- ✓ Bloques anónimos de código de procedimientos.
- ✓ Numerosos tipos de datos y posibilidad de definir nuevos tipos. Además de los tipos estándares en cualquier base de datos, están disponibles, entre otros, tipos geométricos, de direcciones de red, de cadenas binarias, UUID, XML y matrices.

- ✓ Soporta el almacenamiento de objetos binarios grandes (gráficos, videos, sonido, etc.).

SQL:

- ✓ SQL92, SQL99, SQL2003, SQL2008.
- ✓ Llaves primarias y foráneas.
- ✓ Columnas auto-incrementales.
- ✓ Índices compuestos, únicos, parciales y funcionales en cualquiera de los métodos de almacenamiento disponibles, B-tree, R-tree, *hash* ó *GiST*.
- ✓ Consultas recursivas.
- ✓ Funciones ventanas.
- ✓ Joins.
- ✓ Vistas.
- ✓ Disparadores comunes, por columna, condicionales.
- ✓ Reglas.
- ✓ Herencia de tablas.

Algunos de los límites de PostgreSQL 9.1 son:

Tabla 2: Límites de PostgreSQL 9.1.

Límite	Valor
Máximo tamaño de base de datos	Ilimitado (Depende de tu sistema de almacenamiento)
Máximo tamaño de tabla	32 TB
Máximo tamaño de fila	1.6 TB
Máximo tamaño de campo	1 GB
Máximo número de filas por tabla	Ilimitado
Máximo número de columnas por tabla	250 - 1600 (dependiendo del tipo)
Máximo número de índices por tabla	Ilimitado

1.7 PgAdmin III

PgAdmin III es una aplicación gráfica para administrar el SGBD PostgreSQL, siendo la más completa y popular con licencia Open Source. Es capaz de gestionar versiones a

partir de la PostgreSQL 7.3 ejecutándose en cualquier plataforma. Está diseñado para responder a las necesidades de todos los usuarios, desde escribir consultas sql simples hasta desarrollar bases de datos complejas. (32)

Entre sus principales características se tienen:

- ✓ Multiplataforma.
- ✓ Amplia documentación.
- ✓ Acceso a los datos.
- ✓ Acceso a todos los objetos de PostgreSQL.
- ✓ Diseñado para múltiples versiones de PostgreSQL.

1.8 Lenguaje de modelado unificado

El Lenguaje Unificado de Modelado (***UML,*** por sus siglas en inglés) es un lenguaje que permite modelar, construir y documentar los elementos que forman un producto de software que responde a un enfoque orientado a objetos. Se ha convertido en el estándar internacional para definir organizar y visualizar los elementos que configuran la arquitectura de una aplicación orientada a objetos. Con este lenguaje, se pretende unificar las experiencias acumuladas sobre técnicas de modelado e incorporar las mejores prácticas actuales en un acercamiento estándar. (33)

1.9 Herramienta de modelado con UML

Las herramientas de modelado con UML permiten aplicar la metodología de análisis y diseño orientado a objetos, así como abstraerse del código fuente, en un nivel donde la arquitectura y el diseño se tornan más obvios, más fáciles de entender y modificar. Se decide utilizar la herramienta Visual Paradigm para UML y para el modelado del proceso de ETL, además de encontrarse dentro del marco tecnológico del proyecto.

1.9.1 Visual Paradigm

Visual Paradigm es una herramienta UML profesional que soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue. El software de modelado UML ayuda a una más rápida

construcción de aplicaciones de calidad, mejores y a un menor coste. Permite dibujar todos los tipos de diagramas de clases, código inverso, generar código desde diagramas y generar documentación. La herramienta UML también proporciona abundantes tutoriales de UML, demostraciones interactivas de UML y proyectos UML. Esta herramienta de modelado ofrece: (34)

- ✓ Entorno de creación de diagramas para UML 2.1.
- ✓ Diseño centrado en casos de uso y enfocado al negocio que genera un software de mayor calidad.
- ✓ Uso de un lenguaje estándar común a todo el equipo de desarrollo que facilita la comunicación.
- ✓ Disponibilidad de múltiples versiones, para cada necesidad.
- ✓ Disponibilidad en múltiples plataformas.

Conclusiones del capítulo

Los almacenes de datos están en el centro de atención de las grandes compañías actualmente, pues brindan una herramienta que permite hacer uso efectivo de la información así como dar soporte al proceso de toma de decisiones. Estos reflejan la lógica del negocio y deben ser construidos para adaptarse perfectamente a esta lógica.

Para su construcción se cuentan con diversas herramientas, dentro de estas se encuentran Pentaho BI y PDI que fueron seleccionadas para ser utilizadas, pues apoyan el desarrollo de soluciones de inteligencia del negocio reduciendo y optimizando el ciclo de vida de las aplicaciones, además permite de forma paralela el proceso ETL, el modelado y visualización de datos, que a su vez ayuda a reducir costos, mejorar la productividad y acortar el tiempo necesario para obtener resultados concretos.

Como se demuestra los almacenes de datos constituyen una importante solución a los problemas de manejo de información facilitando la aplicación de técnicas estadísticas de análisis y modelación para encontrar relaciones ocultas entre los datos; obteniendo un valor añadido para el negocio. Proporcionan además la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.

Capítulo 2: Análisis y diseño del almacén de datos operacional

2.1 Introducción

En este capítulo se abordará el análisis y diseño del almacén de datos operacional. Se describe la aplicación de la metodología seleccionada, el desarrollo de los procesos definidos por esta, los artefactos generados así como un análisis de los procesos de negocio, la identificación de los requisitos de información, los casos de uso, la especificación de las medidas, hechos, dimensiones, los modelos de datos dimensional y físico, proporcionando una base empírica y metodológica adecuada para las implementaciones de este tipo de aplicaciones.

2.2 Análisis

2.2.1 Definición del negocio

ZERA es una plataforma para la gestión del aprendizaje que tiene sus orígenes en la concepción pedagógica denominada “hiperentornos de aprendizaje“, propuesta y desarrollada por pedagogos y especialistas del Ministerio de Educación de Cuba (MINED) y de la Universidad de las Ciencias Informáticas (UCI).

Su estructura es la siguiente: se divide en seis subsistemas, los cuales se encuentran estrechamente interrelacionados, además de dos simuladores desarrollados con el fin de apoyar el aprendizaje de las asignaturas de Física y Matemática.

Actualmente existen más de 200 tablas, 5 funciones, 600 funciones *Trigger* y 10 vistas. Tiene un peso o volumen mayor a los 90 MB y cuenta con varios mecanismos de indexación para facilitar el trabajo con los campos más consultados en la base de datos, un ejemplo es el efectuado para consultar el campo deleted_at de la tabla tb_matter.

La plataforma facilita la gestión de recursos para apoyar el proceso de enseñanza aprendizaje en cualquier institución que la utilice, brindando la oportunidad de publicar, compartir e interactuar con estos.

2.2.2 Tema de análisis

Un tema de análisis no es más que la división o categorización de la información de una organización según las diferentes temáticas que contenga. Se elige un tema de análisis basándose en los objetivos que se persigan. La identificación de los temas de análisis es de suma importancia para el desarrollo del almacén de datos, estos permiten la factibilidad, utilidad, y éxito de las estructuras que se están diseñando.

Se define entonces como tema de análisis:

- ✓ Los recursos de la Plataforma Educativa ZERA.

2.2.3 Roles y permisos

Para el acceso a la información contenida dentro del almacén de datos que se implementará se definieron los siguientes roles:

- ✓ Administrador: administra el almacén y se encarga de todas las actividades referentes al proceso de ETL.
- ✓ Analista: es el encargado de analizar la información y visualizar los reportes.

Tabla 3: Roles y permisos.

Roles	Permisos	
	Lectura	Escritura
Administrador	X	X
Analista	X	

2.2.4 Reglas del negocio

Las reglas del negocio describen las políticas, normas, operaciones, definiciones y restricciones presentes en una organización y son de vital importancia para alcanzar los objetivos propuestos. Las reglas del negocio deben ser expresadas en lenguaje natural y orientadas al negocio, a continuación se mencionan las reglas del negocio asociadas al almacén de datos operacional de la Plataforma Educativa ZERA.

RN 1. El tiempo de visitas a un recurso será igual a la diferencia entre la fecha inicial en la que un usuario entra a un recurso y la fecha fin en que sale del mismo.

RN 2. El tiempo total de visitas a un recurso será guardado en segundos.

RN 3. Una vez cargados los datos en el almacén no pueden existir valores nulos.

RN 4. No se podrá insertar un par usuario y tipo de usuario que no exista en la tabla que guarda la relación entre ellos.

2.2.5 Necesidades de usuario

Después de un análisis del negocio se determina que los administradores de la Plataforma Educativa ZERA, como principales usuarios del almacén de datos, necesitan conocer la interacción de sus usuarios con los recursos que se brindan en la plataforma.

La situación actual muestra que no es posible obtener por parte de los directivos información sobre aspectos asociados al uso de los recursos. Partiendo de esto se describen los requisitos de información, funcionales y no funcionales del sistema.

2.2.6 Requisitos de información

Los requisitos de información describen la información y los datos que el sistema debe proveer o debe acceder. Estos se definen a partir de las necesidades de información identificadas en el negocio, que permitan el análisis del comportamiento de los indicadores a medir según los objetivos y metas de la organización. (35)

A continuación se muestran los requisitos de información definidos:

RI 1. Tiempo de visualización y cantidad de visitas a los recursos comprendido en un período de tiempo.

RI 2. Tiempo de visualización y cantidad de visitas a los recursos comprendido en un período de tiempo agrupado por usuarios.

RI 3. Tiempo de visualización y cantidad de visitas a los recursos comprendido en un período de tiempo agrupado por escuelas.

RI 4. Tiempo de visualización y cantidad de visitas a los recursos comprendido en un período de tiempo agrupado por programas de estudio.

RI 5. Tiempo de visualización y cantidad de visitas a los recursos comprendido en un período de tiempo agrupado por usuarios y escuelas.

RI 6. Tiempo de visualización y cantidad de visitas a los recursos comprendido en un período de tiempo agrupado por usuarios y programas de estudio.

RI 7. Tiempo de visualización y cantidad de visitas a los recursos comprendido en un período de tiempo agrupado por escuelas y programas de estudio.

RI 8. Tiempo de visualización y cantidad de visitas a los recursos comprendido en un período de tiempo agrupado por usuarios, escuelas y programas de estudio.

2.2.7 Requisitos funcionales

Los requisitos funcionales describen lo que el sistema debe hacer. Estos dependen del tipo de software que se desarrolle, de los posibles usuarios del software y del enfoque general tomado por la organización. (36)

Los requisitos funcionales definidos para el almacén de datos son los siguientes:

RF 1. Extraer los datos de la Plataforma Educativa ZERA.

Se sustraen los datos de las disímiles fuentes de origen, estos pueden presentar formatos o estructuras diferentes definidos por cada una de las fuentes primarias.

RF 2.Realizar la transformación de los datos extraído.

Al presentar diferentes estructuras los datos obtenidos de las fuentes de información se deben transformar para convertirse en aptos para la carga.

RF 3. Cargar datos.

Una vez extraídos los datos y transformados correctamente son insertados en el sistema de destino.

RF 3. Analizar la información.

Se debe analizar la información guardada en el almacén de datos operacional para la generación de reportes que apoyen el proceso de toma de decisiones.

2.2.8 Requisitos no funcionales

Los requisitos no funcionales son aquellos que no se refieren directamente a las funciones específicas que proporciona el sistema, sino a las propiedades emergentes de éste como la fiabilidad, el tiempo de respuesta y la capacidad de almacenamiento. De forma alternativa definen las restricciones del sistema como la capacidad de los dispositivos de entrada/salida y las representaciones de datos que se utilizan en las interfaces del sistema. (36)

Los requisitos no funcionales definidos para el almacén de datos son los siguientes:

RNF 1. Software:

- ✓ Sistema Operativo Linux.
- ✓ Herramienta PDI en su versión 4.2.1.
- ✓ El SGBD para implementar el almacén de datos es el PostgreSQL en su versión 9.1. Como interfaz de administración del SGBD se usará el PgAdmin III.

RNF 2. Hardware:

- ✓ Servidor de BD: Procesador Quad Core o superior, 8Gb de RAM para garantizar el correcto funcionamiento del sistema al ser accedido por varios usuarios y 500 Gb de capacidad en disco duro para el almacenamiento de la información.

RNF 3. Fiabilidad:

- ✓ La cantidad de errores en el proceso de integración define la calidad de los datos que se están almacenando, es por ello que es crítico, definiéndose así 0 errores/puntos de función.
- ✓ El tiempo medio de reparación depende fundamentalmente de la magnitud del fallo pero se estima que como promedio sea de 24 horas.

RNF 4. Eficiencia:

- ✓ El tiempo de respuesta de los reportes deberá estar comprendido entre 5 segundos y hasta un día.

RNF 5: Capacidad:

- ✓ En el proceso de integración solo tendrá conectado un usuario, el administrador, que tendrá la tarea de monitorizar el proceso de integración de datos.

RNF 6. Restricciones del diseño:

- ✓ El lenguaje de programación del proceso de integración de la base de datos será SQL, desarrollado en PostgreSQL 9.1.
- ✓ Lograr que los elementos definidos en el almacén tengan una estructura homogénea. Las estructuras del almacén de datos operacional se nombrarán de una manera estándar teniendo en cuenta el tipo de estructura que se maneje.

2.2.9 Casos de uso

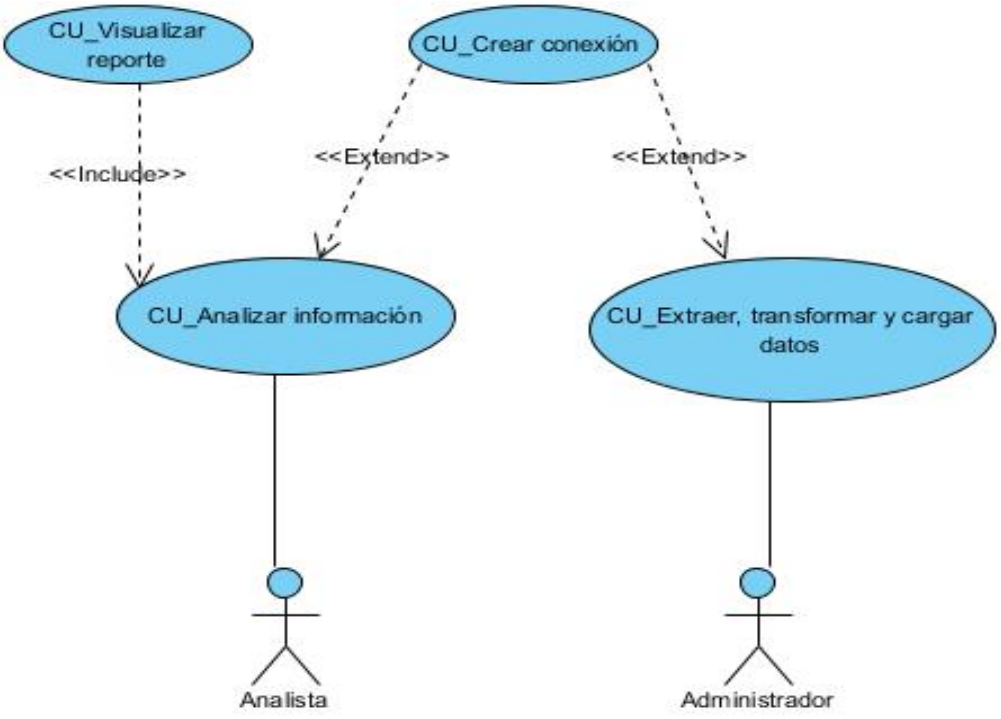


Figura 3: Diagrama de casos de uso.

2.2.10 Descripción de los casos de uso

A continuación se muestra una tabla con los cuatro casos de uso definidos así como una breve descripción de cada uno, para conocer la descripción detallada de estos (ver *Anexo 1. Descripción de los casos de uso*)

Tabla 4: Descripción de CU.

Caso de uso	Descripción
Extraer, transformar y cargar datos	El caso de uso inicia cuando el administrador selecciona la opción ejecutar trabajo. El sistema realiza todas las transformaciones necesarias. El CU finaliza cuando todos los datos de la fuente son cargados en el almacén.
Analizar información	El CU inicia cuando el actor desea consultar la información guardada en el almacén de datos. Luego de configurar las diferentes opciones se visualiza el reporte deseado.
Crear conexión	El CU inicia cuando el analista desea crear una nueva conexión. El sistema le solicita los datos necesarios y este los introduce. Finaliza el CU cuando se crea la conexión.
Visualizar reporte	El CU inicia cuando el analista desea hacer algún reporte sobre la información guardada en el almacén de datos. Selecciona la fuente de datos y genera el reporte deseado. El CU finaliza cuando el sistema muestra el reporte.

2.2.11 Arquitectura del almacén de datos

La arquitectura del almacén de datos quedó estructurada de manera tal que se compone por la fuente de datos (base de datos de la Plataforma educativa ZERA) y tres subsistemas bases, los cuales son:

- ✓ Subsistema de integración: encargado de la extracción e integración de la información para su posterior carga al almacén de datos.
- ✓ Subsistema de almacenamiento: encargado de almacenar toda la información del almacén de datos en las diferentes tablas de hechos y dimensiones definidas.

- ✓ Subsistema de visualización: encargado de consultar los datos guardados en el almacén, con el objetivo de mostrarlos a los usuarios finales en los distintos reportes, contribuyendo a la toma de decisiones.

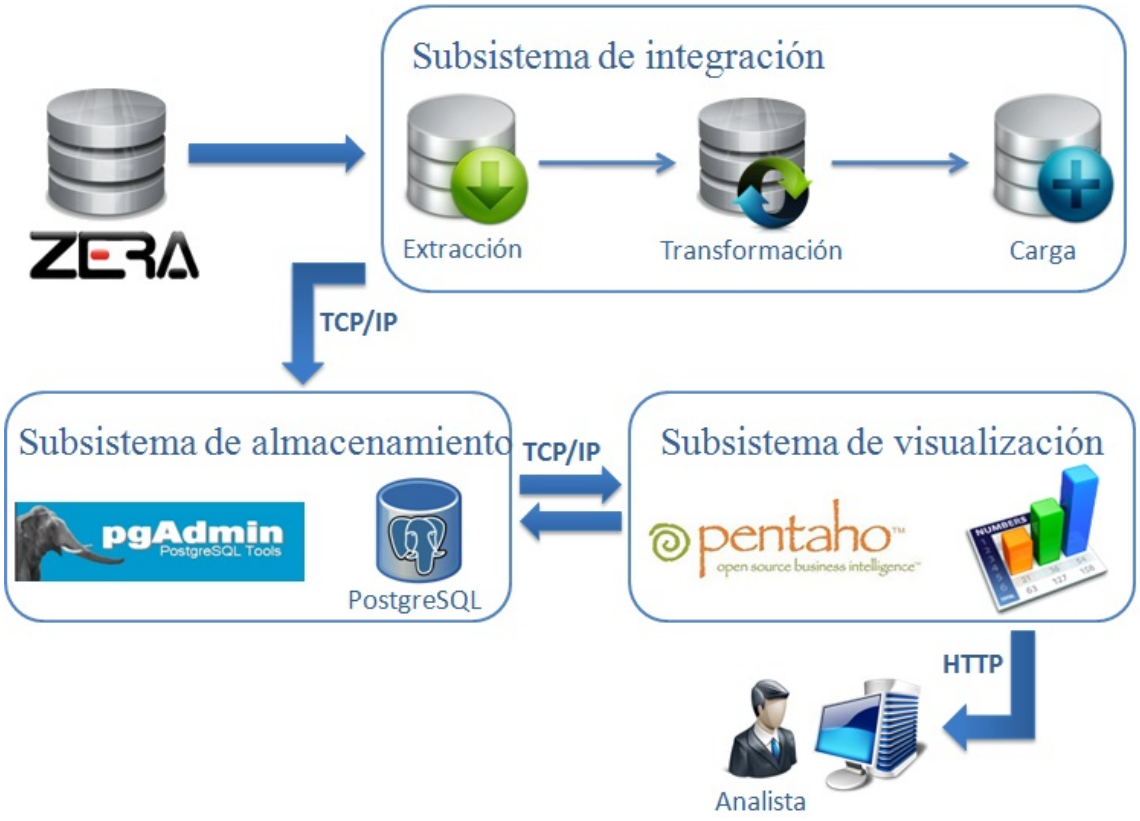


Figura 4: Arquitectura del almacén de datos operacional.

2.3 Diseño

2.3.1 Dimensiones, hechos y medidas

En este epígrafe se describirán cada uno de los hechos, dimensiones y medidas que conformarán el almacén de datos operacional así como una descripción de los mismos.

Tablas de hechos: son medidas numéricas que pueden calcularse con la suma de varias cantidades de la tabla, cada una de las mediciones es tomada como la intersección de todas las dimensiones. En consecuencia, por lo general los hechos a almacenar en una tabla de hechos van a ser casi siempre valores numéricos, enteros o reales.

Tablas dimensionales: son aquellas donde las descripciones textuales de las dimensiones del negocio son almacenadas. Cada una de ellas ayuda a describir un miembro de la dimensión respectiva. Se utilizan para restringir y agrupar los datos almacenados en una tabla de hechos cuando se realizan consultas sobre dicho datos.

El hecho definido en la investigación se describe en la siguiente tabla:

Tabla 5: Descripción del hecho.

Hecho	Descripción
hech_visitas_usuario_recurso	Contiene el indicador del hecho y todas las dimensiones y medidas definidas en el modelo.

Los atributos que componen al hecho anteriormente expuesto se describen en la siguiente tabla:

Tabla 6: Descripción de los atributos que componen el hecho.

Atributos del hecho	Descripción
dim_usuarioid	Almacena la llave primaria de la tabla dim_usuario.
dim_escuelaid	Almacena la llave primaria de la tabla dim_escuela.
dim_tipo_usuarioid	Almacena la llave primaria de la tabla dim_tipo_usuario.
dim_recursoid	Almacena la llave primaria de la tabla dim_recurso.
dim_programa_estudioid	Almacena la llave primaria de la tabla dim_programa_estudio.

Las dimensiones definidas en la investigación se describen en la siguiente tabla:

Tabla 7: Descripción de dimensiones.

Dimensiones	Descripción
dim_usuario	Contiene el identificador de la dimensión y el nombre los usuarios.
dim_tipo_usuario	Contiene el identificador de la dimensión y el tipo de usuario.
dim_escuela	Contiene el identificador de la dimensión y el nombre de las escuelas.
dim_recurso	Contiene el identificador de la dimensión, el nombre del recurso y la llave de la tabla dim_tipo_recurso.

dim_programa_estudio	Contiene el identificador de la dimensión y el nombre del programa de estudio.
dim_tipo_recurso	Contiene el identificador de la dimensión y el tipo de recurso.

En cuanto a las medidas asociadas al hecho, se muestran en la siguiente tabla sus aspectos principales:

Tabla 8: Descripción de medidas.

Medidas	Descripción	Calculable	Hecho al que pertenece
fecha_inicio	Fecha en que el usuario comienza a visitar un recurso.	No	hech_visitas_usuario_recurso
fecha_fin	Fecha en que el usuario termina de visitar un recurso.	No	hech_visitas_usuario_recurso
tiempo	Tiempo total de un usuario en un recurso.	Si	hech_visitas_usuario_recurso

2.3.2 Matriz Bus o Matriz Dimensional

En este epígrafe se realizará la Matriz Bus o Dimensional la cual representa las relaciones existentes entre los hechos y las dimensiones del almacén de datos operacional. Conocido el hecho y las dimensiones la matriz queda de la siguiente forma:

Tabla 9: Matriz Bus.

Dimensión / Hecho	hech_visita_usuario_recurso
dim_usuario	X
dim_tipo_usuario	X
dim_escuela	X
dim_recurso	X
dim_programa_estudio	X
dim_tipo_recurso	

2.3.4 Modelo de datos dimensional

Una vez descritas las tablas de hecho y las dimensiones y confeccionada la matriz bus para definir las relaciones entre estas, se procede a diseñar el modelo de datos dimensional. Este modelo en cuanto a tipología de esquema es un copo de nieve, pues contiene una relación de mucho-mucho (*-*) y de uno a mucho (1-*) entre dos dimensiones.

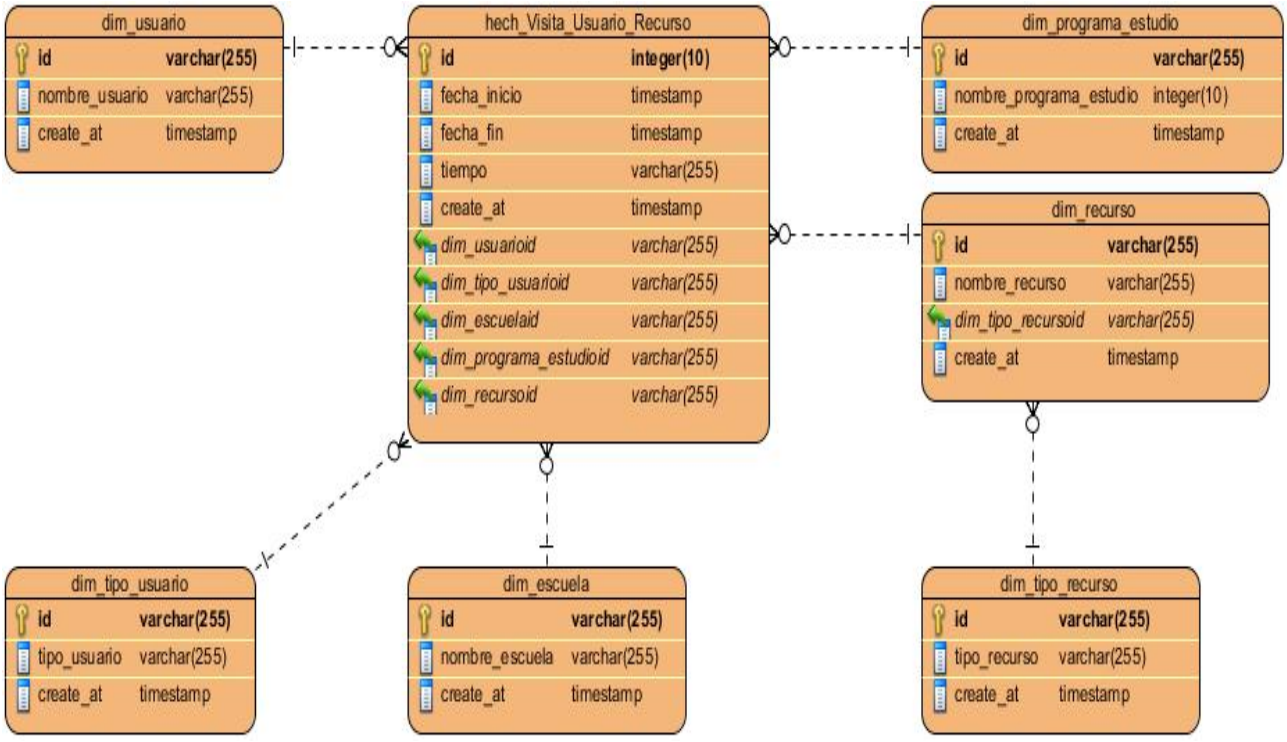


Figura 5: Modelo de datos dimensional.

2.3.5 Modelo de datos físico

El modelo físico se obtuvo a partir del modelo de datos dimensional, en dicho modelo aparece la tabla cuyo nombre es dim_usuario_dim_tipo_usuario debido a la existencia de la relación de mucho-mucho entre las tablas dim_usuario y dim_tipo_usuario.

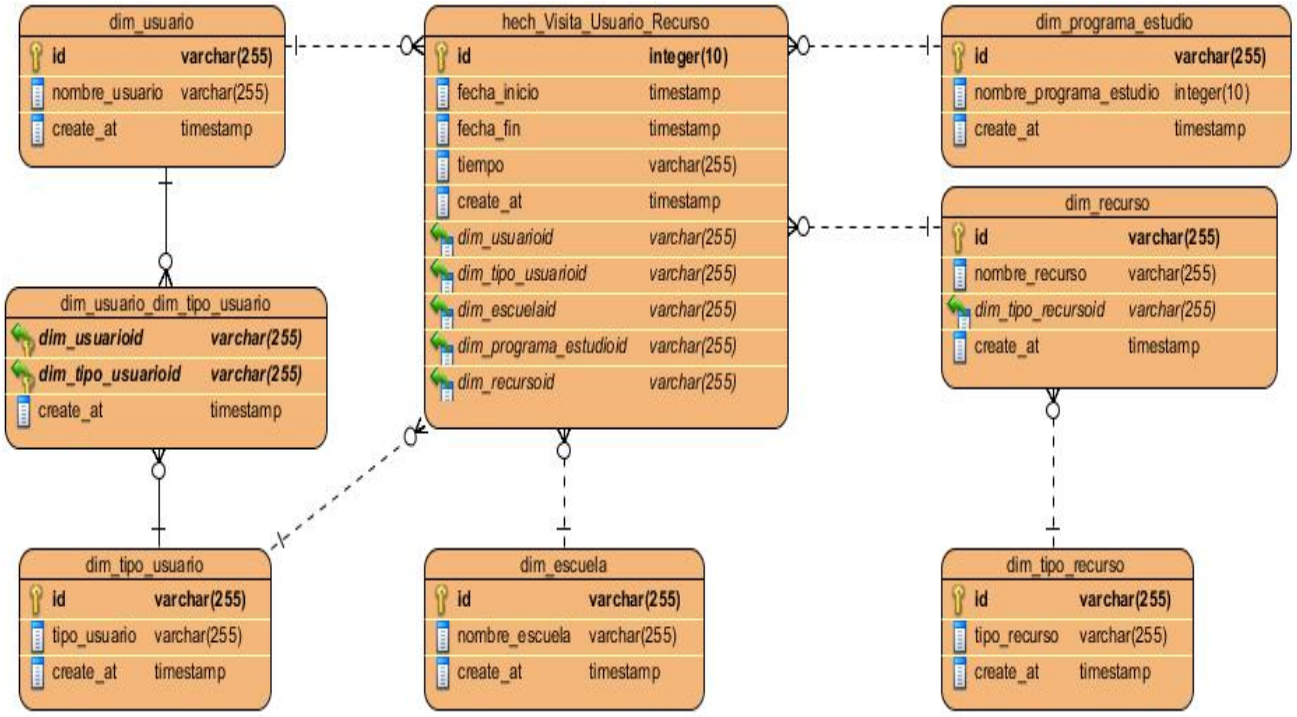


Figura 6: Modelo de datos físico.

Conclusiones del capítulo

El análisis y diseño de sistemas es una guía que permite estructurar todo el proceso de desarrollo. Aún más cuando se trata del diseño de sistemas que manejan grandes volúmenes de información, en estos casos se busca analizar sistemáticamente la entrada de datos, la transformación de estos, el almacenamiento y la salida de información. También se facilita el analizar, diseñar e implementar mejoras que puedan incorporarse al sistema.

El proceso de análisis y diseño del almacén de datos para apoyar la toma de decisiones en la Plataforma Educativa ZERA permitió identificar cuatro reglas del negocio. Además se realizó la especificación de requisitos, donde se definieron ocho requisitos de información, tres funcionales y seis no funcionales. Se identificaron los actores y casos de uso del sistema. También se facilitó la especificación de los modelos de datos físico y dimensional con el objetivo de tener una mejor visión de las relaciones existentes en la base de datos y que se incorporarán al almacén de datos operacional.

Capítulo 3: Implementación y prueba del almacén de datos operacional

3.1 Introducción

En el presente capítulo se describirá la estructura de los datos del almacén, se realizará la implementación de los subsistemas de integración y el subsistema de visualización de datos definiéndose la arquitectura de integración y explicándose los procesos de extracción, transformación y carga de los datos, así como la implementación de los trabajos y la generación de reportes utilizando el almacén de datos. Además se realizarán pruebas que parten de la aplicación de listas de chequeo y casos de pruebas, con el objetivo verificar el correcto funcionamiento del almacén de datos así como la validación de la propuesta de solución para contribuir a la toma de decisiones.

3.2 Implementación del subsistema de almacenamiento

En la implementación del subsistema de almacenamiento se realiza el desarrollo de la estructura física del almacén de datos, además se definen todos los estándares de codificación que van a poseer las estructuras del almacén de datos, para facilitar la comprensión por parte del cliente.

3.2.1. Estándares de codificación

Con el objetivo de organizar la estructura del almacén de datos, se formaliza un modelo, norma, patrón o estándar de codificación. Esta acción permite a los desarrolladores entender cada una de las estructuras. En la siguiente tabla se muestran como quedaron definidos estos estándares.

Tabla 10: Estándares de codificación.

Estructura	Descripción	Ejemplo
Tabla de hecho	La tabla de hecho tendrá una cadena que demuestra que es un hecho y el concepto que describe.	hech_<concepto>
Tablas de dimensiones	Todas las tablas de dimensiones tendrán una cadena que demuestra que son dimensiones y el	dim_<concepto>

	concepto que describen.	
Llaves primarias	Todas las llaves primarias de cada tabla se nombrarán id	id

3.2.2 Estructuras de datos

Las estructuras de datos son una forma de organizar un conjunto de datos con el objetivo de facilitar su manipulación. En un almacén de datos se deben crear estructuras lógicas que faciliten y optimicen el tratamiento de la información, para que el manejo de los datos se realice de manera correcta.

3.2.3 Esquemas y tablas

Los esquemas son una forma de organizar los datos. En su interior pueden contener tablas, operadores, funciones, tipos de datos, a los que el usuario puede acceder siempre y cuando tenga los permisos necesarios. El esquema creado fue:

- ✓ recurso

Las tablas definidas en la base de datos son:

- ✓ dim_escuela
- ✓ dim_programa_estudio
- ✓ dim_recurso
- ✓ dim_tipo_recurso
- ✓ dim_tipo_usuario
- ✓ dim_usuario
- ✓ dim_usuario_dim_tipo_usuario
- ✓ hech_visita_usuario_recurso

3.2.4 Estructura física de la base de datos

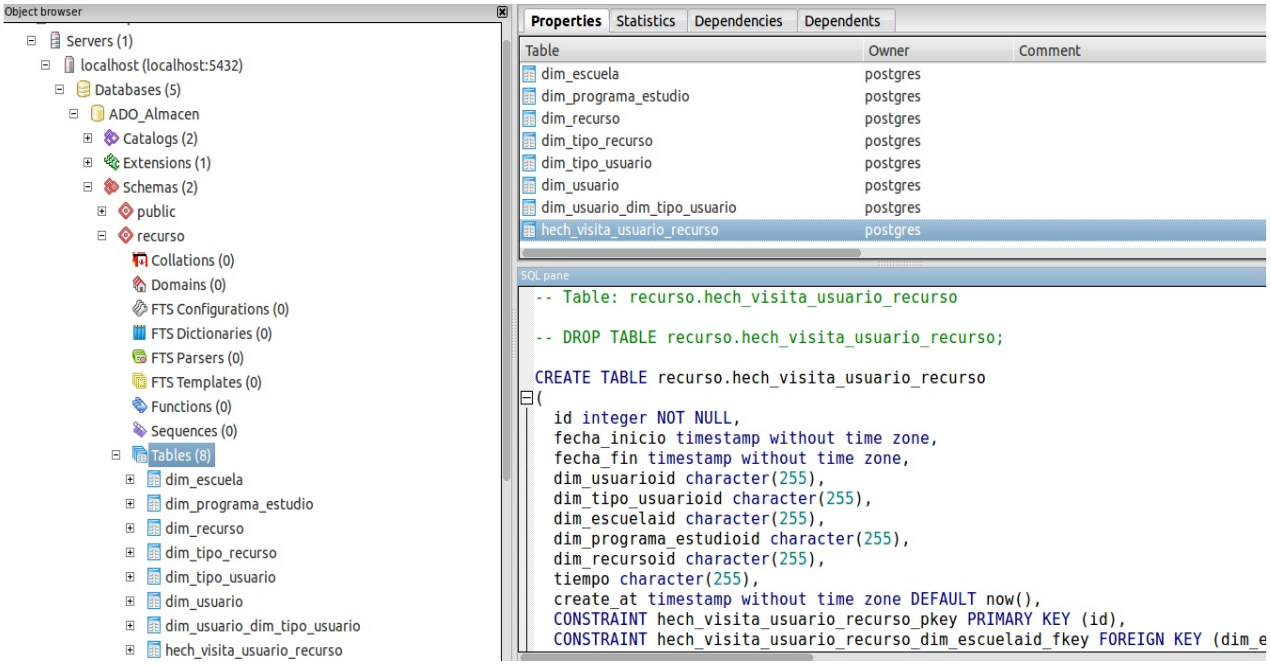


Figura 7: Vista física de la base de datos.

3.3 Implementación del subsistema de integración de datos

Seguidamente se describen los elementos asociados a la implementación del subsistema de integración de datos y se realiza un análisis de los principales componentes de este.

3.3.1 Arquitectura de integración

En el proceso de desarrollo de un software, la arquitectura es el diseño más importante en la estructura del sistema, debido a que permite guiar la construcción del mismo. Para un correcto desarrollo del almacén es recomendable que se analice rigurosamente todo el proceso de integración de los datos. En este proceso la arquitectura está fundamentada por algunos elementos necesarios para la correcta implementación del sistema que se describen a continuación:

- ✓ La base de datos de la Plataforma Educativa ZERA que representa la fuente de datos.

- ✓ El área de almacenamiento temporal; es el punto intermedio entre la fuente y el almacén, es donde se realiza la integración y transformación de los datos.
- ✓ El almacén de datos el cual constituye el destino a donde se integrarán los datos a cargar.

Figura 8: Arquitectura de integración.

3.3.2 Implementación de los procesos ETL

Los procesos de ETL en la herramienta PDI se efectúan a través de transformaciones y trabajos. A continuación se describen estos procesos:

- ✓ Extracción

El primer proceso de ETL consiste en extraer los datos desde los sistemas de origen, que pueden ser provenientes de diferentes fuentes. A través de las conexiones a las fuentes se establece desde dónde se extraerán los datos para analizarlos. En este caso particular los datos son extraídos de la base de datos de la Plataforma Educativa ZERA. La extracción se realizó a través del componente Entrada de Tabla.

- ✓ Transformación

La transformación es el proceso básico de ETL, se compone de pasos que están enlazados a través de saltos. Los pasos son los elementos más pequeños dentro de las transformaciones. Los saltos son el medio por donde fluye la información entre los diferentes pasos. Después de realizada la extracción de los datos el sistema se encuentra listo para la etapa de transformación. Durante el proceso se llevaron a cabo tareas tales como: unión por clave, validación de campos nulos; así como asignación de llaves para relacionar la información de los hechos con las dimensiones.

- ✓ Carga

La carga es el último subproceso dentro de los procesos de ETL, el cual consiste en cargar todos los datos que ya han sido transformados satisfactoriamente en el almacén de datos. La carga de los datos se realizó a través de la opción Insertar/Actualizar.

3.3.3 Implementación de las transformaciones

Las trasformaciones realizadas a los datos obtenidos de la base de datos de la Plataforma Educativa ZERA fueron las siguientes:

Para la transformación de las dimensiones se utilizaron tres pasos:

- ✓ Entrada de tabla.
- ✓ Seleccionar/Renombrar valores.
- ✓ Insertar/actualizar.

Todas las transformaciones para crear cada una de las dimensiones fueron realizadas de igual forma. Mediante el componente Entrada de tabla se extraen los datos necesarios desde la base de datos de la Plataforma Educativa ZERA. Luego se renombran los valores de la tabla para que concuerden con los del modelo físico del almacén y por ultimo último se cargan los datos mediante la opción Insertar/Actualizar. En la Figura 9 se muestra la transformación realizada para extraer, transformar y cargar los datos de los usuarios.

Figura 9: Transformación Usuario realizada con la herramienta PDI.

Para la transformación del hecho se utilizaron los siguientes pasos:

1. Entrada de tabla.
2. Seleccionar/Renombrar valores.
3. Añadir secuencia.
4. Data validator.

5. Dummy.
6. Insertar/actualizar.

En la transformación del hecho se utilizó el paso 1 para extraer los datos. Luego se renombraron los valores de la tabla para que concuerden con los del modelo físico del almacén utilizando el paso 2. Se añadió una secuencia para crear un id auto incremental para la tabla en el paso 3. Para evitar que existieran campos nulos se utilizó el pasos 4 y el 5 para desechar los campos nulos encontrados. Finalmente para la carga de los datos se utilizó el paso 6.

En la Figura 10 se muestra la transformación realizada para extraer, transformar y cargar los datos que conforman el hecho.

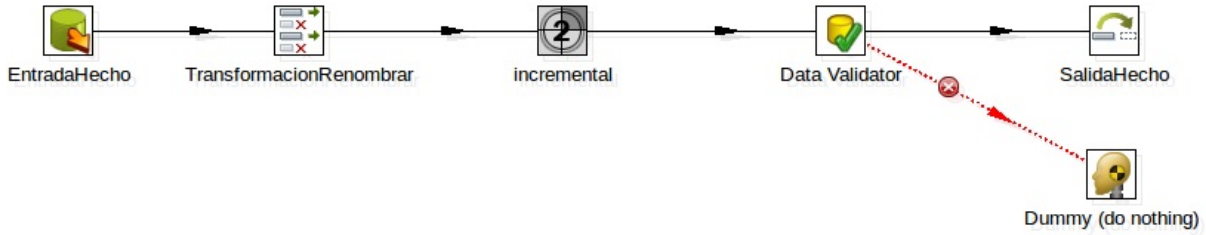


Figura 10: Transformación del hecho realizada con la herramienta PDI.

3.3.4 Implementación de los trabajos.

Un trabajo es un conjunto de tareas con el objetivo de realizar una acción determinada. En los trabajos se utilizan pasos específicos que son diferentes a los disponibles en las transformaciones. Permite ejecutar una o varias transformaciones de las diseñadas siguiendo una secuencia de ejecución. Después que se realizaron las transformaciones a los datos se organizó la carga de las tablas al almacén de datos. A continuación será descrito este proceso:

Primero ser realiza un trabajo para cargar y actualizar el área de almacenamiento temporal lugar desde donde se poblará finalmente el almacén de datos. En él se cargan cada una de las transformaciones en un orden lógico donde se respetan las relaciones entre las tablas. En la Figura 11 se muestra el trabajo realizado para poblar el área temporal.

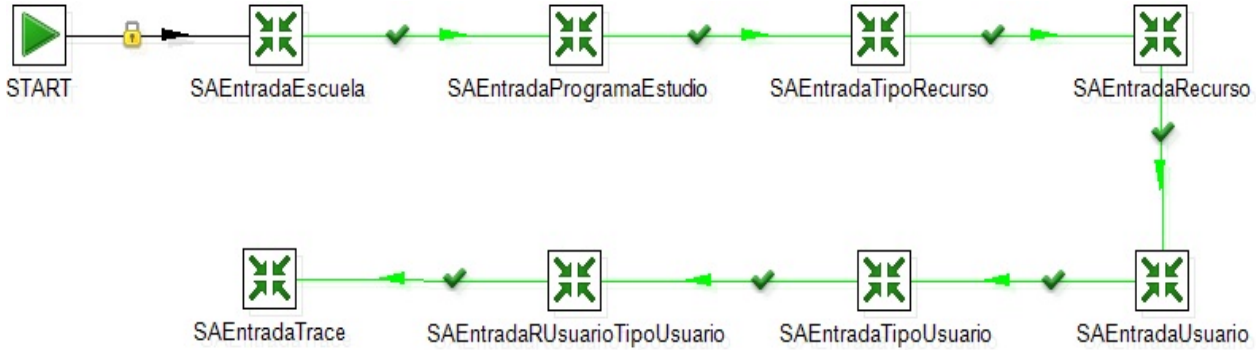


Figura 11: Trabajo para realizar la carga del área de almacenamiento temporal.

Finalmente se realiza el trabajo para cargar y actualizar el almacén. En él se carga el trabajo del área temporal y luego cada una de las transformaciones previamente realizadas para cargar cada uno de las dimensiones y el hecho. En el trabajo se realiza una validación para comprobar las conexiones a las diferentes bases de datos mediante el paso *Check Db connections* y en caso de encontrarse un error se le notificará al administrador mediante correo electrónico, para ello se utilizó el paso Mail. Para la actualización del almacén una vez que esté poblado el mismo, se definió que se realice en un intervalo de 24 horas (diariamente), aunque si los administradores desean modificar el tiempo de actualización solo deben reconfigurar el paso START. En la Figura 12 se muestra el trabajo realizado para poblar el almacén de datos operacional.

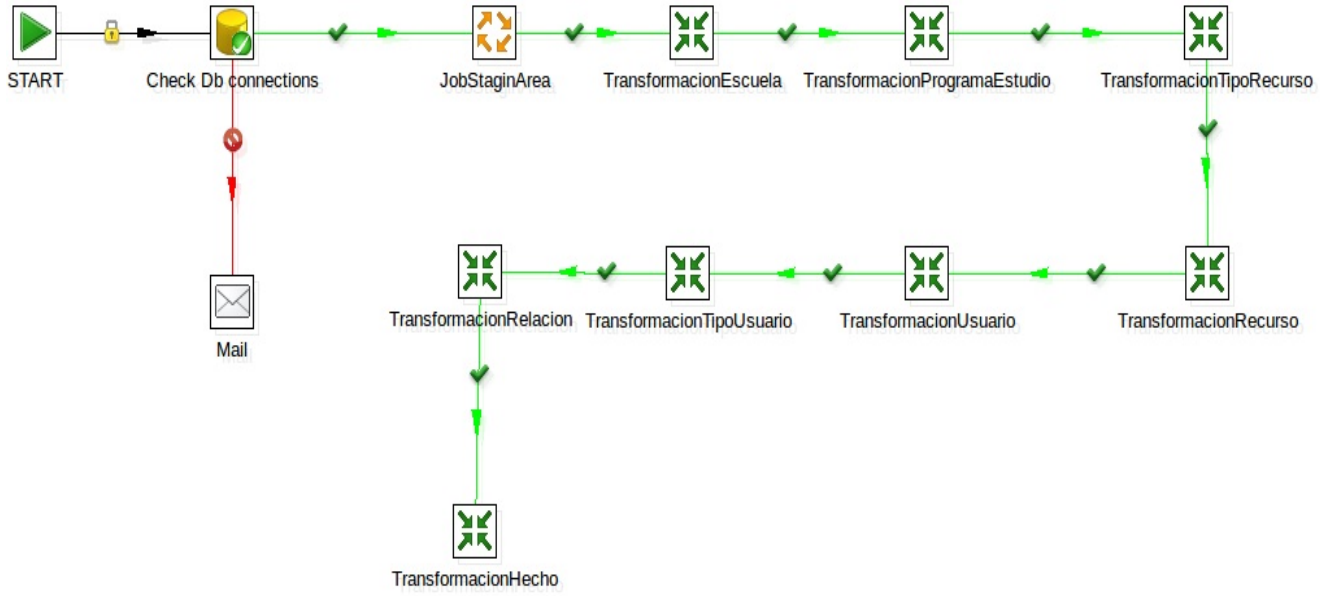


Figura 12: Trabajo para realizar la carga del almacén.

3.4 Implementación del subsistema de visualización de datos

En este epígrafe se describen los elementos asociados a la creación de reportes para la visualización y análisis de la información.

3.4.1 Reportes

Una vez realizado el almacén de datos se podrán realizar reportes con el objetivo de obtener información relevante de los datos que se encuentran almacenados para contribuir a la toma de decisiones. Para la implementación de la solución se pueden generar más de 100 tipos de reportes diferentes, teniendo en cuenta la información que se maneja en el almacén de datos con respecto a los recursos, así como los usuarios, programas de estudio y escuelas que están asociados a ellos. Los mismos se pueden mostrar en formato HTML, PDF, XML y CSV. A continuación se muestran algunos ejemplos de los reportes que se pueden generar:

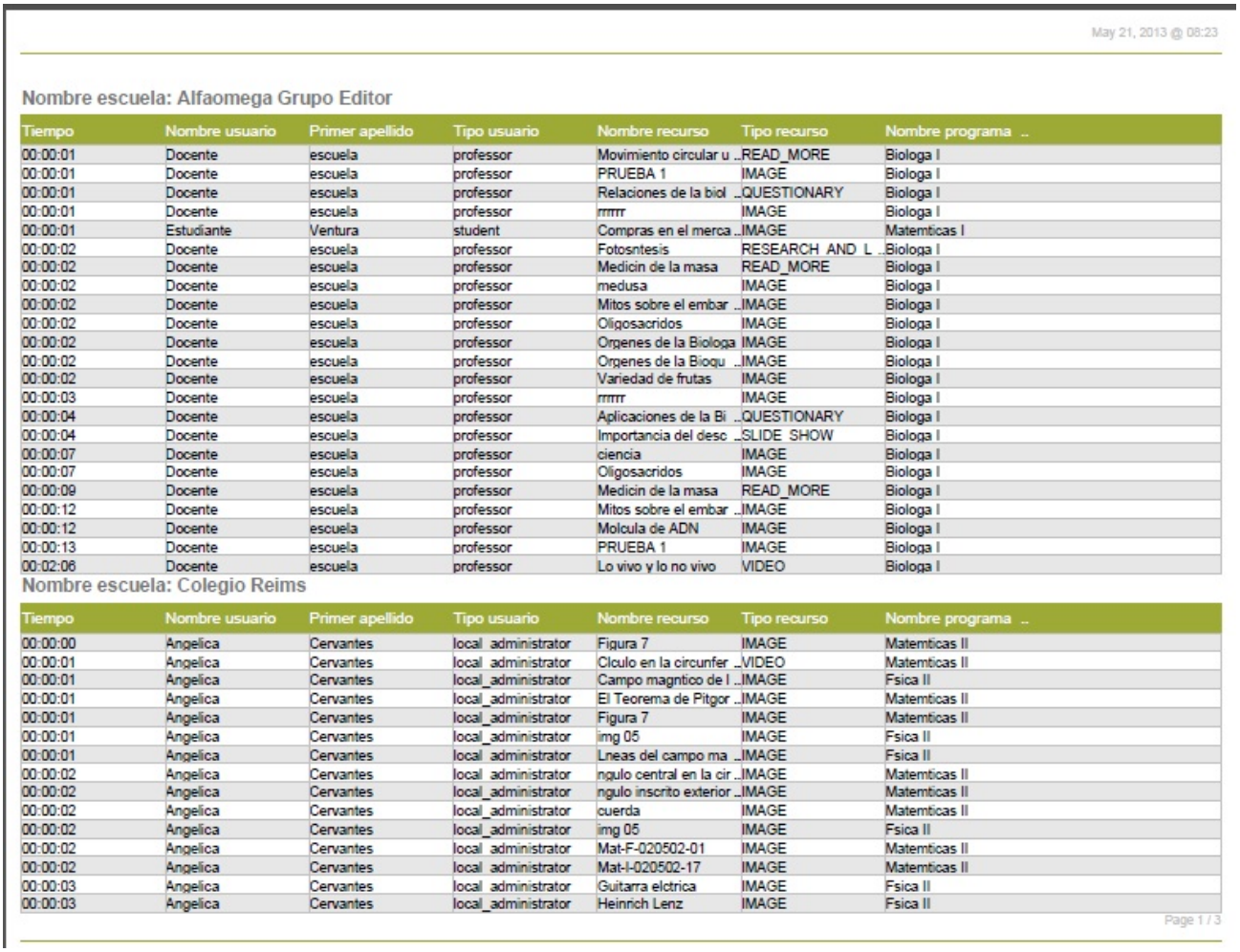
Nombre usuario: Angelica

Nombre programa estudio	Tiempo	Nombre recurso
Física II	00:00:01	Campo magnético de la tierra
Física II	00:00:01	img 05
Física II	00:00:01	Líneas del campo magnético de un imán recto
Física II	00:00:02	img 05
Física II	00:00:03	Guitarra eléctrica
Física II	00:00:03	Heinrich Lenz
Física II	00:00:03	Sentido de la corriente eléctrica
Física II	00:00:07	img
Matemáticas II	00:00:00	Figura 7
Matemáticas II	00:00:01	Cálculo en la circunferencia y en el círculo.
Matemáticas II	00:00:01	El Teorema de Pitágoras
Matemáticas II	00:00:01	Figura 7
Matemáticas II	00:00:02	Ángulo central en la circunferencia.
Matemáticas II	00:00:02	Ángulo inscrito exterior al centro de la circunferencia
Matemáticas II	00:00:02	cuerda
Matemáticas II	00:00:02	Mat-F-020502-01
Matemáticas II	00:00:02	Mat-I-020502-17
Matemáticas II	00:00:03	Mat-I-020502-07
Matemáticas II	00:00:04	Cálculo en la circunferencia y en el círculo
Matemáticas II	00:00:06	Circunferencia
Matemáticas II	00:00:23	Ángulo inscrito exterior al centro de la circunferencia
Matemáticas II	00:01:12	Mesa con sombrilla

Nombre usuario: ARIAN DAVID

Nombre programa estudio	Tiempo	Nombre recurso
Matemáticas II	00:00:01	Ala Delta
Matemáticas II	00:00:01	Ángulos adyacentes

Figura 13: Reporte por usuario en formato HTML.



May 21, 2013 @ 08:23

Nombre escuela: Alfaomega Grupo Editor

Tiempo	Nombre usuario	Primer apellido	Tipo usuario	Nombre recurso	Tipo recurso	Nombre programa ..
00:00:01	Docente	escuela	professor	Movimiento circular u ..	READ_MORE	Biologa I
00:00:01	Docente	escuela	professor	PRUEBA 1	IMAGE	Biologa I
00:00:01	Docente	escuela	professor	Relaciones de la biol ..	QUESTIONARY	Biologa I
00:00:01	Docente	escuela	professor	rrrrrrr	IMAGE	Biologa I
00:00:01	Estudiante	Ventura	student	Compras en el merca ..	IMAGE	Matemticas I
00:00:02	Docente	escuela	professor	Fotosntesis	RESEARCH_AND_L ..	Biologa I
00:00:02	Docente	escuela	professor	Medicin de la masa	READ_MORE	Biologa I
00:00:02	Docente	escuela	professor	medusa	IMAGE	Biologa I
00:00:02	Docente	escuela	professor	Mitos sobre el embar ..	IMAGE	Biologa I
00:00:02	Docente	escuela	professor	Oligosacridos	IMAGE	Biologa I
00:00:02	Docente	escuela	professor	Orgenes de la Biologa	IMAGE	Biologa I
00:00:02	Docente	escuela	professor	Orgenes de la Bioqu ..	IMAGE	Biologa I
00:00:02	Docente	escuela	professor	Variedad de frutas	IMAGE	Biologa I
00:00:03	Docente	escuela	professor	rrrrrrr	IMAGE	Biologa I
00:00:04	Docente	escuela	professor	Aplicaciones de la Bi ..	QUESTIONARY	Biologa I
00:00:04	Docente	escuela	professor	Importancia del desc ..	SLIDE_SHOW	Biologa I
00:00:07	Docente	escuela	professor	ciencia	IMAGE	Biologa I
00:00:07	Docente	escuela	professor	Oligosacridos	IMAGE	Biologa I
00:00:09	Docente	escuela	professor	Medicin de la masa	READ_MORE	Biologa I
00:00:12	Docente	escuela	professor	Mitos sobre el embar ..	IMAGE	Biologa I
00:00:12	Docente	escuela	professor	Molcula de ADN	IMAGE	Biologa I
00:00:13	Docente	escuela	professor	PRUEBA 1	IMAGE	Biologa I
00:02:06	Docente	escuela	professor	Lo vivo y lo no vivo	VIDEO	Biologa I

Nombre escuela: Colegio Reims

Tiempo	Nombre usuario	Primer apellido	Tipo usuario	Nombre recurso	Tipo recurso	Nombre programa ..
00:00:00	Angelica	Cervantes	local_administrator	Figura 7	IMAGE	Matemticas II
00:00:01	Angelica	Cervantes	local_administrator	Clculo en la circunfer ..	VIDEO	Matemticas II
00:00:01	Angelica	Cervantes	local_administrator	Campo magntico de l ..	IMAGE	Fsica II
00:00:01	Angelica	Cervantes	local_administrator	El Teorema de Pitgor ..	IMAGE	Matemticas II
00:00:01	Angelica	Cervantes	local_administrator	Figura 7	IMAGE	Matemticas II
00:00:01	Angelica	Cervantes	local_administrator	img 05	IMAGE	Fsica II
00:00:01	Angelica	Cervantes	local_administrator	Lneas del campo ma ..	IMAGE	Fsica II
00:00:02	Angelica	Cervantes	local_administrator	ngulo central en la cir ..	IMAGE	Matemticas II
00:00:02	Angelica	Cervantes	local_administrator	ngulo inscrito exterior ..	IMAGE	Matemticas II
00:00:02	Angelica	Cervantes	local_administrator	cuerda	IMAGE	Matemticas II
00:00:02	Angelica	Cervantes	local_administrator	img 05	IMAGE	Fsica II
00:00:02	Angelica	Cervantes	local_administrator	Mat-F-020502-01	IMAGE	Matemticas II
00:00:02	Angelica	Cervantes	local_administrator	Mat-I-020502-17	IMAGE	Matemticas II
00:00:03	Angelica	Cervantes	local_administrator	Guitarra elctrica	IMAGE	Fsica II
00:00:03	Angelica	Cervantes	local_administrator	Heinrich Lenz	IMAGE	Fsica II

Page 1 / 3

Figura 14: Reporte por escuela en formato PDF.

3.5 Guía de implantación

La guía de implantación contiene los pasos necesarios para la implantación de cualquier sistema informático. Antes de abordar los pasos necesarios para lograr esta implantación, es preciso conocer los requisitos del sistema los cuales constituyen requisitos no funcionales que este debe cumplir.

Requisitos del sistema:

- ✓ 1 Servidor Intel Quad Core a 2.4 GHz, con 8 GB de memoria RAM, 500 GB de capacidad de almacenamiento y Red a 100 Mbps o más.

Pasos de implantación de la solución:

- ✓ Instalar sistema operativo Ubuntu.
- ✓ Instalar máquina virtual de java (java-6-openjdk o superior).
- ✓ Configurar la variable de entorno.
- ✓ Instalar el SGBD PostgreSQL 9.1.
- ✓ Instalar una herramienta de administración de base de datos.
- ✓ Se debe crear la nueva base de datos utilizando la herramienta de administración, para el área de almacenamiento temporal y el almacén.
- ✓ Copiar las herramientas PDI y Pentaho be-server y configurarlas para su utilización.

3.6 Pruebas

Una vez que se ha dado por concluida la implementación, se debe dar paso a una de las etapas más importantes en el ciclo de desarrollo de un software: las pruebas. Estas garantizan que se haya cumplido con las especificaciones iniciales que fueron definidas para el almacén de datos.

La prueba es el proceso de ejecución de un programa con el fin de encontrar deficiencias. Una prueba tiene éxito si descubre errores que no han sido detectados hasta entonces.

Las pruebas deben centrarse en dos objetivos:

1. Probar si el software no hace lo que debe hacer.
2. Probar si el software hace lo que no debe hacer. (37)

Existen diferentes tipos de pruebas, cada uno aplicable en un entorno diferente de acuerdo a los objetivos que se persigan en su realización. Para verificar el correcto funcionamiento del almacén de datos se realizaron las siguientes pruebas.

3.6.1 Lista de chequeo

La lista de chequeo es un documento que tiene un conjunto de parámetros a medir sobre un aspecto determinado, dígase documentación o aplicación. Es un instrumento de medición y evaluación que consiste básicamente en un formulario de preguntas referentes al atributo de calidad que se está probando y de las características del documento en el

caso de la documentación. Cada pregunta tiene asociada una evaluación en una escala que da una medida del grado de cumplimiento y disponibilidad de la propiedad evaluada, de esta manera se determina la evaluación del elemento probado.

La lista de chequeo contiene diferentes indicadores a evaluar los cuales se encuentran distribuidos en dos secciones fundamentales:

- ✓ Estructura del documento: se comprueban todos los aspectos definidos por el expediente de proyecto.
- ✓ Elementos definidos por el modelo de desarrollo: abarca todos los indicadores a evaluar durante la etapa de desarrollo del mercado según el modelo de desarrollo.

Los elementos que forman parte de la estructura de la lista de chequeo son:

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

Los resultados obtenidos de la aplicación de la lista de chequeo, fueron satisfactorios en vista a la construcción del almacén de datos operacional, pues en el mismo solo se detectaron errores en los entregables. (Ver A*nexo 5. Resultados de la lista de chequeo*).

3.6.2 Casos de prueba

Los casos de prueba son un conjunto de condiciones o variables bajo las cuales se podrá determinar si el requisito de una aplicación es parcial o completamente satisfactorio. Con

el propósito de verificar los requisitos del almacén de datos se diseñaron cuatro casos de pruebas basados en los cuatro casos de uso definidos en la etapa de Análisis y diseño. Los casos de prueba realizados fueron:

Tabla 11: Descripción de los casos de pruebas.

Casos de prueba	Descripción
Extraer, transformar y cargar datos	El administrador selecciona la opción ejecutar trabajo. El sistema realiza todas las transformaciones necesarias. Todos los datos de la fuente son cargados en el almacén.
Analizar información	El analista entra en la herramienta Pentaho be-server. Luego de configurar las diferentes opciones visualiza el reporte deseado.
Crear conexión	El analista desea crear una nueva conexión. El sistema le solicita los datos necesarios. Se insertan todos los datos correctamente y se crea la conexión.
Visualizar reporte	El analista desea hacer algún reporte sobre la información guardada en el almacén de datos. Selecciona la fuente de datos y genera el reporte deseado. El sistema muestra el reporte.

Anteriormente se muestran los cuatro casos de pruebas definidos así como una breve descripción de cada uno, para conocer la descripción detallada de estos (ver *Anexo 2. Descripción de los casos de prueba*).

Luego de realizar los casos de prueba, se procede a verificar las no conformidades encontradas en las pruebas de calidad realizadas al almacén de datos operacional. Para el control del proceso de corrección de las no conformidades se realizará una tabla, la misma contará con el requisito funcional, la no conformidad encontrada y su estado.

Tabla 12: No conformidades detectadas durante las pruebas.

No. NC	Requisito funcional	No Conformidad	Estado con respecto a la solución
1	CP_Extraer, transformar y cargar información	Cuando hay fallo en la conexión a la base de datos, el sistema no envía un correo de notificación al administrador.	Resuelta

2	CP_Extraer, transformar y cargar información	La tabla dim_recursos carga una mayor cantidad de recursos de los que realmente son visitados.	Resuelta

3.7 Validación de la propuesta de solución para apoyar la toma de decisiones

3.7.1 Introducción

En este epígrafe se realiza la validación de la propuesta de almacén de datos operacional basado en el uso de los recursos de la Plataforma Educativa ZERA en cuanto al apoyo que brinda este para contribuir a la toma de decisiones. Para ellos se utiliza el método de validación de expertos.

3.7.2 Elección de los expertos

Para llevar a cabo este proceso se entiende como experto, no solo a aquel que es un especialista en su campo, sino, a aquellos que puedan realizar contribuciones válidas sobre el tema en cuestión, dado que poseen conocimientos basados en la práctica y la experiencia.

Fueron seleccionados para la validación de la propuesta especialistas capaces de ejercer criterios concluyentes del trabajo, de realizar recomendaciones acertadas, de ayudar para el enriquecimiento del mismo y estar dispuestos a participar. Para ello se tomaron en cuenta los siguientes criterios de selección:

- ✓ Calificación profesional.
- ✓ Ser graduado de nivel superior.
- ✓ Tener conocimientos en el ámbito de los procesos educativos.
- ✓ Tener conocimientos en el trabajo con la Plataforma Educativa ZERA.
- ✓ Tener más de un año de experiencia en proyectos productivos.
- ✓ Tener disposición a participar en la encuesta.

Finalmente se escogieron 8 expertos entre los que se encuentran especialistas en áreas educativas así como profesionales que trabajan con la Plataforma Educativa ZERA.

3.7.3 Elaboración de las encuestas

Para obtener la información referente a la contribución del almacén de datos operacional basado en el uso de los recursos de la Plataforma Educativa ZERA a la toma de decisiones, se confeccionó una encuesta que establece una comparación entre los reportes que son generados actualmente por el módulo de Reportes de la Plataforma Educativa ZERA y los que permite generar el almacén de datos operacional. Los expertos deben responder tres preguntas asociadas a si son capaces de tomar decisiones teniendo en cuenta cada uno de estos reportes. (Ver *Anexo 6. Encuesta para validar que el almacén de datos contribuya a la toma de decisiones*.)

3.7.4 Análisis de los resultados

Para realizar un análisis de los resultados se procedió a efectuar el cálculo del nivel de concordancia pues para que la propuesta tenga un mayor nivel de validez es preciso que exista un acuerdo favorable entre los expertos entrevistados. La concordancia entre los expertos se considera aceptable con respecto a un determinado valor, generalmente cuando Cc ≥ 60% (38).

Para ello se calcula el coeficiente de concordancia entre las respuestas dadas a través de la fórmula siguiente: $Cc = (1 – Vn/Vt) * 100$

Donde:

- ✓ Cc: Coeficiente de concordancia expresado en porcentaje.
- ✓ Vn: Cantidad de expertos en contra del criterio predominante.
- ✓ Vt: Cantidad total de expertos. (38)

En la siguiente tabla se muestra un resumen de las respuestas dadas por el panel de expertos así como el coeficiente de concordancia, expresado en por ciento, para cada una de las preguntas formuladas en el encuesta. Las celdas marcadas con “x” representan las respuestas contrarias al criterio predominante en el panel.

Tabla 13: Cálculo del coeficiente de concordancia.

	Preguntas correspondientes a la encuesta		
Expertos	**1**	**2**	**3**
1			
2			
3			
4	X		
5			
6		X	X
7			X
8			
Cc	87,5%	87,5%	75%

La siguiente figura muestra los resultados obtenidos en la tabla anterior:

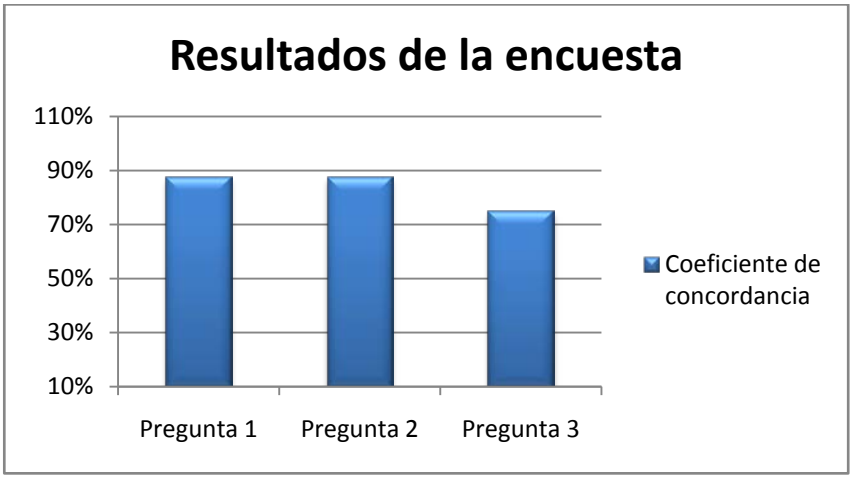


Figura 15: Resultados de la encuesta.

Luego de analizar los resultados de la figura anterior se concluye que un 100% de las preguntas quedaron con un valor aceptable al ser mayor que 60 %; demostrándose que los expertos determinaron que el almacén de datos operacional posibilita a aquellos que analicen los reportes que se generan contribuir a la toma de decisiones basado en el uso de los recursos de la Plataforma Educativa ZERA.

Conclusiones del capítulo

Durante este capítulo se realizó exitosamente la implementación y prueba del almacén de datos para contribuir a la toma de decisiones basado en el uso de los recursos de la Plataforma Educativa ZERA. Se implementaron los subsistemas de integración de datos y de visualización que permitieron obtener la información a almacenar y poder realizar los reportes para el análisis de la información respectivamente.

Se describió la implementación de los procesos ETL que constituye uno de los pasos fundamentales en el desarrollo de un almacén de datos así como la realización de los trabajos y mediante la construcción de una guía de implantación, se logran detallar los pasos para la instalación del almacén de datos, así como los requerimientos necesarios para ello.

Se realizó también la validación del almacén de datos para comprobar mediante el método de criterio de expertos que este permite contribuir a la toma de decisiones basado en el uso de los recursos de la Plataforma Educativa ZERA.

Conclusiones generales

La investigación tuvo como propósito desarrollar un almacén de datos que permitiera apoyar el proceso de toma de decisiones basado en el uso de los recursos de la Plataforma Educativa ZERA, de esta se arriban a las siguientes conclusiones:

- ✓ Los almacenes de datos representan actualmente una solución para el manejo y análisis de grandes volúmenes de información en diversos ámbitos, pues permiten organizar los datos obteniendo información valiosa para apoyar el proceso de toma de decisiones.

- ✓ La metodología utilizada, desarrollada por especialistas de centro DATEC, constituye una guía referencial para organizar el proceso de desarrollo de almacenes de datos, basada en la metodología de Kimball e incorpora aspectos asociados al desarrollo de software como artefactos y pruebas.

- ✓ La suite de código abierto Pentaho BI ofrece potentes herramientas para la realización de los procesos ETL y la generación de reportes asociada al desarrollo de almacenes de datos.

- ✓ Con la culminación de este trabajo se espera aportar una herramienta para contribuir a la toma de decisiones basado en el uso de los recursos de la Plataforma Educativa ZERA.

- ✓ Las pruebas realizadas al almacén de datos permitieron validar la solución por lo que se le dio cumplimiento al objetivo general de esta investigación.

- ✓ Mediante el método de criterio de expertos se constató que el almacén de datos operacional desarrollado durante la investigación permite contribuir a la toma de decisiones basado en el uso de los recursos de la Plataforma Educativa ZERA.

Recomendaciones

Luego de la investigación realizada para el desarrollo del presente trabajo y teniendo en cuenta las ideas que surgieron durante el progreso de la misma, se recomienda:

- ✓ Integrar la visualización de los reportes a la Plataforma Educativa ZERA.
- ✓ Crear nuevos almacenes de datos que permitan analizar otras variables.
- ✓ Mantener actualizada la información del almacén de datos, incorporando progresivamente una mayor cantidad de datos para que se realice un análisis más detallado de los mismos.

Referencias bibliográficas

1. **Albarracín, P.** AeTecno. *América económica Tecno.* [En línea] 2011. [Citado el: 8 de diciembre de 2012.] http://tecno.americaeconomia.com/noticias/el-desafio-del-big-data-mas-que-solo-grandes-volumenes-de-datos.

2. **Ae-Tecno.** América economía. [En línea] 2012. [Citado el: 2 de diciembre de 2012.] http://tecno.americaeconomia.com/noticias/el-desafio-del-big-data-mas-que-solo-grandes-volumenes-de-datos.

3. **Rodríguez, J. L. y Sáenz, O.** *Tecnología Educativa y Nuevas tecnologías aplicadas a la educación.* Madrid : Marfil Alcoy, 1995.

4. **Eduardo, A.** Redescubriendo el aprendizaje. El uso de las TIC. [En línea] 2012. [Citado el: 2 de diciembre de 2012.] http://aprendizajepordentro.blogspot.com/2012/06/analiticas-del-aprendizaje.html.

5. **Creadores de Soluciones de Informáticas, S.L.** Navactiva. [En línea] 2005. [Citado el: 5 de diciembre de 2012.] http://www.navactiva.com/es/asesoria/empresas-que-utilizan-data-ware-house_21378.

6. **León, A. y Sotolongo, R.** *Modelo de descripción de arquitectura de almacenes de datos para ensayos clínicos del Centro de Inmunología Molecular.* La Habana : s.n., 2010.

7. **ZORAN, N y VILJAN, M.** *Data warehouse for an e-learning platform.* Vukovarska : s.n., 2010.

8. **Cesares, C.** *Data Warehousing.* México : Instituto Tecnológico de Chihuahua.

9. **Mora, R.** Adictos al trabajo. [En línea] 2012. [Citado el: 13 de diciembre de 2012.] http://www.adictosaltrabajo.com/tutoriales/tutoriales.php?pagina=datawarehouse.

10. **Kimball, R.** *The Data warehouse Lifecycle Toolkit.* New York : WILEY, 2006.

11. **Inmon, B.** *Building the Data Warehouse. .* s.l. : Wiley, 1992.

12. **Wrembel, R. y Concilia, C.** *DATA WAREHOUSES AND OLAP Concepts, Architectures and Solutions.* s.l. : IDEA GROUP PUBLISHING, 2007.

13. **Bernabeu, R. Dario.** DATAPRIX. [En línea] 2009. [Citado el: 14 de enero de 2013.] http://www.dataprix.com/32-oltp.

14. **Lanzillotta, A.** MasterMagazine. [En línea] 2009. [Citado el: 14 de enero de 2013.] http://www.mastermagazine.info/termino/6841.php.

15. **Synerplus.** Synerplus. La Plataforma Selenne ERP. [En línea] 2011. [Citado el: 10 de enero de 2013.] http://www.synerplus.es/Informacion-Tecnica/Data-Mart/309.html.

16. **Fernández, C.** DataPrix. [En línea] 2009. [Citado el: 15 de enero de 2013.] http://www.dataprix.com/arquitectura-data-warehouse-areas-datos-nuestro-almacen-corporativo.

17. **Pentaho.** BIDW Directory.Business Intelligence And Data Warehousing Directory. [En línea] 2009. [Citado el: 16 de enero de 2013.] http://www.bi-dw.info/pentaho.htm.

18. **Stratebi .** Stratebi Open business intelligence. [En línea] 2011. [Citado el: 18 de enero de 2013.] http://www.stratebi.com/spagobi.

19. **Díaz, J.** BeyeNETWORK Global coverage of the business intelligence ecosystem. [En línea] 2009. [Citado el: 18 de enero de 2013.] http://www.beyenetwork.es/view/10428.

20. **Pierri, M.** slideshare. [En línea] 2013. [Citado el: 20 de enero de 2013.] http://www.slideshare.net/mpierri/manipulacion-de-datos-con-kettle.

21. **Pérez, A.** Brujuleo. Tecnología, eComerce y otras cosas. [En línea] 2012. [Citado el: 20 de enero de 2013.] http://www.brujuleo.es/introduccion-a-talend.

22. **scriptella.** Scriptella ETL. [En línea] 2011. [Citado el: 19 de enero de 2013.] http://scriptella.javaforge.com/.

23. **Jitterbit.** Jitterbit. [En línea] 2013. [Citado el: 19 de enero de 2013.] http://www.jitterbit.com/solutions/etl-data-integration.

24. **Hartman, Y. y Ramón, D.** *Implementación del proceso de extracción, transformación y carga en un almacén de datos operacional para CIMEX.* La Habana : s.n., 2009.

25. **Bernabeu, D.** DATAPRIX. [En línea] 2009. [Citado el: 21 de enero de 2013.] http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/-metodologia-hefesto/51-introduccion.

26. —. DATAPRIX. [En línea] 2009. [Citado el: 21 de enero de 2013.] http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/-metodologia-hefesto/53-caracteristicas.

27. **Espinosa, R.** DATAPRIX. [En línea] 2010. [Citado el: 21 de enero de 2013.] http://www.dataprix.com/blogs/respinosamilla/fases-implantaci-n-sistema-dw-metodologia-para-construcci-n-dw-0.

28. **Rivadera, G.** UCASAL. Universidad católica de Salta. [En línea] 2010. [Citado el: 25 de enero de 2013.] http://www.ucasal.net/templates/unid-academicas/ingenieria/apps/5-p56-rivadera-formateado.pdf.

29. **Ruiz, O.** *Data Mart para la gestión de reportes y apoyo a la toma de decisiones del departamento de RR.HH. de la empresa de agua S.A.* Bolivia : UPSA, 2010.

30. **Hernández, Y.** *Metodología de desarrollo para proyectos de almacenes de datos.* La Habana : s.n., 2012.

31. **Martinez, R.** PostgreSQL. [En línea] 2011. [Citado el: 29 de enero de 2013.] http://www.postgresql.org.es/sobre_postgresql.

32. **Ubuntu.** GUÍA DOCUMENTADA PARA UBUNTU. [En línea] 2013. [Citado el: 10 de mayo de 2013.] http://www.guia-ubuntu.com/index.php?title=PgAdmin_III.

33. **Carrillo, A.** *Herramienta Multimedia de apoyo a la Enseñanza de la Metodología RUP de Ingeniería del Software.* Edición electrónica gratuita. 2009.

34. **Targetware.** Targetware Informática S.A.C. [En línea] 2013. [Citado el: 5 de febrero de 2013.] http://www.software.com.ar/visual-paradigm-para-uml.html.

35. **Schiefer, J. y Brunckner, M.** *A HOLISTIC APPROACH FOR MANAGING REQUIREMENTS OF DATA WAREHOUSE SYSTEMS.* Vienna : Vienna University of Technology, 2002.

36. **Sommerville, I.** *Ingeniería del software.* Madrid : Pearson Educación, 2005.

37. **Pressman, R.** *Ingeniería del Software. Un enfoque práctico.* Madrid : Ma Graw Hill, 2005.

38. **Sánchez, A. y Fernández, M.** *Fuentes de Información para la Inteligencia Competitiva en I+D.* La Habana : CETRA, 2004.

Bibliografía consultada

ETL-Tools.Info. ETL-Tools.Info. Business Intelligence - Almacenes de Datos - ETL. [En línea] 2006. [Citado el: 15 de diciembre de 2012.] http://etl-tools.info/es/bi/proceso_etl.htm.

García, M. *Implementación de un datawarehouse para el soporte de toma de decisiones.* Guatemala : Universidad Francisco Marroquín, 2001.

Inmon, B. Bussiness Inteligence 2007. [En línea] 2007. [Citado el: 8 de diciembre de 2012.] http://www.kimballgroup.com.

Inmon, B. *DW 2.0 - Architecture for the Next Generation of Data Warehousing. .* s.l. : Elsevier Press, 2008.

Kendrik, T. *Identifying and Managing Project Risk.* New York, USA : AMACON, 2003.

LIST, B., BRUCKNER, K. y SCHIEFER, J. *A Comparison of Data Warehouse Development Methodologies Case Study of the Process Warehouse.* Berlin, Alemania : Berlin Heidelberg, 2002.

Martínez, R. PostgreSQL. 2010. Tinysofa Copyright-grupo de desarrollo global de postgreSQL. [En línea] 2010. [Citado el: 4 de febrero de 2013.] http://www.postgresql.org.

MAZÓN, J. *Designing Data Warehouses: From Business Requirement Analysis to Multidimensional Modeling. .* Paris, Francia : In Proceedings of the 1st Int. Workshop on Requirements Engineering for Business Need and IT Alignment, 2005.

Méndez, E. y Senso, J. A. SEDIC Asocioación española de documentación e información. [En línea] 2004. [Citado el: 13 de enero de 2013.] http://www.sedic.es/autoformacion/metadatos/tema1.htm.

Mohedano, F. El método Delphi, prospectiva en Ciencias Sociales a través del análisis de un caso práctico. Colombia : s.n., 2008.

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

Orallo, E. El lenguaje Unificado de Modelado (UML). [En línea] [Citado el: 13 de eneno de 2013.] http://www.disca.upv.es/enheror/pdf/ActaUML.PDF.

Pentaho. Pentaho Open Source Business Intelligence: Kettle.Project. [En línea] 2005. [Citado el: 21 de enero de 2013.] http://kettle.pentaho.org.

Repinosa, M. Kimball vs Inmon. [En línea] 2010. [Citado el: 10 de enero de 2013.] http://churriwifi.wordpress.com/2010/04/19/15-2-ampliacion-conceptos-del-modelado-dmensional.

Rubia, J. M. *Introducción a los almacenes de datos.* Madrid, España : Instituto Católico de Artes e Industrias (ICAI), 2009.

Sánchez, A. M. *Fuentes de Información para la Inteligencia Competitiva en I+D.* La Habana : CETRA, 2004.

Sanz., M. *Análisis y Diseño de un Data Mart para el Seguimiento Academico de Alumnos en un Entorno Universitario.* Madrid : Escuela Politécnica Superior Ingeniería en Informática , 2010.

Anexos

Anexo 1: Descripción de los Casos de Uso (CU)

- ✓ **CU_Extraer, transformar y cargar información**

Objetivo	Realizar la extracción, transformación y carga de los datos de la Plataforma Educativa ZERA hacia el almacén.	
Actores	Administrador	
Resumen	El caso de uso inicia cuando el administrador selecciona la opción ejecutar trabajo. El sistema realiza todas las transformaciones necesarias. El CU finaliza cuando todos los datos de la fuente son cargados en el almacén.	
Complejidad	Alta	
Prioridad	Crítico	
Precondiciones	Se debe haber iniciado la herramienta Pentaho Data integration.	
Postcondiciones	El almacén de datos operacional quedará poblado de información	
Flujo de eventos		
Flujo básico <Nombre del flujo básico>		
	Actor	**Sistema**
1.	Selecciona la opción ejecutar trabajo	
2.		Comienza el trabajo con la transformación STAR
3.		Verifica la conexión a las diferentes bases de datos
4.		Ejecuta el trabajo para extraer, transformar y cargar los datos hacia el área de almacenamiento temporal.
5.		Ejecuta cada una de las transformaciones encontradas en la secuencia.
6.		Finaliza el trabajo y se llena el almacén de datos.
7.		El caso de uso termina
Flujos alternos		
3.a Fallo en la conexión a las base de datos		
	Actor	**Sistema**
1.		Verifica la conexión a las bases de datos
2.		Encuentra un fallo en la conexión con alguna base de datos
3.		Envía un correo electrónico notificando el error al administrador
4.		Finaliza el caso de uso
Relaciones	**CU Incluidos**	
	CU	

	Extendidos	
Requisitos no funcionales		
Asuntos pendientes		

- ✓ **CU_Analizar Información**

Objetivo	Analizar la información existente en almacén de datos
Actores	Analista
Resumen	El CU inicia cuando el actor desea consultar la información guardada en el almacén de datos. Luego de configurar las diferentes opciones se visualiza el reporte deseado.
Complejidad	Alta
Prioridad	Crítico
Precondiciones	Debe haberse iniciado la herramienta Pentaho bi-server. El almacén de datos debe estar lleno de datos.
Postcondiciones	Se generan los reportes para el análisis de la información.

Flujo de eventos		
Flujo básico <Nombre del flujo básico>		
	Actor	**Sistema**
1.	Selecciona la opción crear fuente de datos	
2.		Muestra una ventana para insertar los datos: • Tipo de fuente • Nombre de fuente
3.	Inserta los datos	
4.		Permite seleccionar una conexión. Permite: • Siguiente • Cancelar
5.	Selecciona la conexión la opción siguiente	
6.		Muestra una ventana para seleccionar las tablas que serán necesaria en el análisis Permite: • Atrás • Siguiente • Cancelar
7.	Selecciona por cada esquema las tablas necesarias y selecciona la opción siguiente	
8.		Muestra una ventana para crear los JOIN entre las tablas. Permite: • Atrás • Finalizar • Cancelar

9.	Realiza los JOIN entre las tablas y selecciona la opción finalizar	
10.		Crea la nueva fuente de datos y muestra la página principal
11.	Selecciona la opción crear reporte (Ver CU Visualizar reporte)	
12.		Muestra el reporte
13.		El caso de uso termina
Flujos alternos		
6.a El actor selecciona la opción atrás		
	Actor	**Sistema**
1.		Vuelve a la página anterior
2.		Regresa al paso 4 del flujo básico
Flujos alternos		
8.a El actor selecciona la opción atrás		
	Actor	**Sistema**
1.		Vuelve a la página anterior
2.		Regresa al paso 6 del flujo básico
Flujos alternos		
5.a El actor selecciona la opción crear conexión		
	Actor	**Sistema**
1.0	(Ver CU Crear conexión)	
		Regresa al paso 5 del flujo básico
Flujos alternos		
*.a El actor selecciona la opción cancelar		
	Actor	**Sistema**
1.		Cierra todo
2.		El caso de uso termina

Relaciones	**CU Incluidos**	CU Visualizar reporte
	CU Extendidos	CU Crear conexión
Requisitos no funcionales	**RI 1, RI 2, RI 3, RI 4, RI 5, RI 6, RI 7, RI 8**	
Asuntos pendientes		

- ✓ **CU_Crear conexión**

Objetivo	Descripción detallada de cómo realizar la conexión entre las herramientas de la suite de Pentaho y las fuentes de datos (Plataforma Educativa ZERA, Almacén de datos, etc.).
Actores	Administrador, Analista.
Resumen	El CU inicia cuando el administrador o analista desean crear una nueva conexión. El sistema le solicita los datos necesarios y este los introduce. Finaliza el CU cuando se crea la conexión.
Complejidad	Alta

Prioridad	Critico	
Precondiciones	Debe haberse iniciado alguna de las herramientas de la suite de Pentaho.	
Postcondiciones	Quedará creada la conexión.	
Flujo de eventos		
Flujo básico <Nombre del flujo básico>		
	Actor	**Sistema**
1.	Selecciona la opción crear conexión	
2.		Muestra una ventana para insertar los datos necesarios: • Nombre de la conexión • Tipo de conexión • Acceso • Nombre del host • Nombre de la Base de datos • Numero de puerto • Nombre de usuario • Contraseña Permite: • Probar • Ok • Cancelar
3.	Inserta los datos necesarios y selecciona la opción probar • Nombre de la conexión • Tipo de conexión • Acceso • Nombre del host • Nombre de la Base de datos • Numero de puerto • Nombre de usuario • Contraseña	
4.		Muestra una ventana de información. Permite: • OK
5.	Selecciona la opción Ok	
6.		Cierra la ventana del mensaje y vuelve a la ventana anterior. Permite: • Aceptar • Cancelar
7.	Selecciona la opción Aceptar	
8.		Crea la conexión. El caso de uso termina
Flujos alternos		
*.a El actor selecciona la opción cancelar		
	Actor	**Sistema**
1.		Cancela todo y cierra la ventana

2.		El caso de uso termina.
Flujos alternos		
3.a El actor inserta mal los datos o deja campos nulos		
1.		Muestra una ventana de error Permite: • OK
2.	Selecciona la opción OK	
3.		Regresa al paso 3 del flujo básico

Relaciones	**CU Incluidos**	
	CU Extendidos	
Requisitos no funcionales		
Asuntos pendientes		

- ✓ **CU_Visualizar reporte**

Objetivo	Visualización de reportes.
Actores	Analista
Resumen	El CU inicia cuando el analista desea hacer algún reporte sobre la información guardada en el almacén de datos. Selecciona la fuente de datos y genera el reporte deseado. El CU finaliza cuando el sistema muestra el reporte.
Complejidad	Alta
Prioridad	Crítico
Precondiciones	Debe haberse iniciado la herramienta Pentaho bi-server. El almacén de datos debe estar lleno.
Postcondiciones	Se creará y visualizará el reporte
Flujo de eventos	
Flujo básico <Nombre del flujo básico>	

	Actor	**Sistema**
1.	Selecciona la opción crear nuevo reporte	
2.		Muestra una ventana para insertar los datos necesarios: • Fuente de datos • Estilo del reporte • Formato del reporte Permite: • Siguiente • Cancelar
3.	Selecciona los datos necesarios y la opción siguiente	
4.		Muestra una ventana para seleccionar los campos que se desean mostrar en el reporte

		Permite • Atrás • Siguiente • Finalizar • Cancelar
5.	Selecciona los campos deseados y la opción finalizar	
6.		Muestra el reporte con los campos seleccionados
7.		El caso de uso termina
Flujos alternos		
*.a Selecciona la opción cancelar		
	Actor	**Sistema**
1.		Cierra todas las ventanas
2.		El caso de uso termina
Flujos alternos		
4.a Selecciona la opción atrás		
	Actor	**Sistema**
1.		Muestra la página anterior
2.		Regresa al paso 2 del flujo básico
Flujos alternos		
4.b Selecciona la opción siguiente		
	Actor	**Sistema**
1.		Muestra una ventana para configurar el reporte Permite: • Finalizar • Atrás • Cancelar
2.	Configura el reporte y selecciona la opción finalizar	
3.		Muestra el reporte
4.		El caso de uso termina

Relaciones	**CU Incluidos**	
	CU Extendidos	
Requisitos no funcionales		
Asuntos pendientes		

Anexo 2: Descripción de los Casos de Prueba (CP)

- ✓ **CP_Extraer, transformar y cargar información**

SC Extraer, transformar y cargar datos

Escenario	Descripción	Respuesta del sistema	Flujo central
EC 1.1 Ejecutar	Selecciona la opción ejecutar trabajo	Comienza el trabajo con la transformación STAR	Herramienta Pentaho/ opción ejecutar trabajo
		Verifica la conexión a las diferentes bases de datos	Herramienta Pentaho
		Ejecuta el trabajo para extraer, transformar y cargar los datos hacia el área de almacenamiento temporal.	Herramienta Pentaho
		Ejecuta cada una de las transformaciones encontradas en la secuencia.	Herramienta Pentaho
		Finaliza el trabajo y se llena el almacén de datos.	Herramienta Pentaho

SC Fallo en la conexión a las base de datos

Escenario	Descripción	Respuesta del sistema	Flujo central
EC 1.1 Fallo en la conexión	Se valida la conexión con las bases de datos	Verifica la conexión a las bases de datos	Herramienta Pentaho
		Encuentra un fallo en la conexión con alguna base de datos	Herramienta Pentaho
		Envía un correo electrónico notificando el error al administrador	Herramienta Pentaho

- ✓ **CP_Analizar información**

SC Analizar información

Escenario	Descripción	Nombre de fuente	Tipo de fuente	Conexión	Respuesta del sistema	Flujo central
EC 1.1 Crear fuente de datos	Selecciona la opción crear fuente de datos	NA	NA		Muestra una ventana para insertar los datos	Herramienta Pentaho/ botón Create New
EC 1.2 Insertar datos	Inserta los datos	V	V		Permite seleccionar una conexión	Herramienta Pentaho/ Formulario Crear nueva fuente de datos
EC 1.3 Conexión	Selecciona la conexión la opción siguiente	V	V	V	Muestra una ventana para seleccionar las tablas que serán necesaria en el análisis	Herramienta Pentaho/ Formulario Crear nueva fuente de datos/ botón siguiente
EC 1.4 Seleccionar tablas	Selecciona por cada esquema las tablas necesarias y selecciona la opción siguiente				Muestra una ventana para crear los JOIN entre las tablas	Herramienta Pentaho/ Formulario Crear nueva fuente de datos/ botón siguiente
Ec 1.5 Realizar JOIN	Realiza los JOIN entre las tablas y selecciona la opción finalizar				Crea la nueva fuente de datos y muestra la página principal	Herramienta Pentaho/ Formulario Crear nueva fuente de datos/ botón finalizar
EC 1.6 Crear reporte	Selecciona la opción crear reporte				Muestra el reporte	Herramienta Pentaho/ botón New Report

SC Datos Incorrectos o campos nulos

Escenario	Descripción	Nombre de fuente	Tipo de fuente	Conexión	Respuesta del sistema	Flujo central
EC 1.1 Insertar datos	Inserta los datos incorrectamente	I	I		No muestra las posibles conexiones	Herramienta Pentaho/ Formulario Crear nueva

						fuente de datos
EC 1.2 Datos nulos	Deja campos incompletos	I	I	I	Muestra un mensaje de error	Herramienta Pentaho/ Formulario Crear nueva fuente de datos/ botón siguiente

SC Atrás

Escenario	Descripción	Nombre de fuente	Tipo de fuente	Conexión	Respuesta del sistema	Flujo central
EC 1.1 Atrás	Selecciona la opción atrás	NA	NA	NA	Vuelve a la vista anterior	Herramienta Pentaho/ Formulario Crear nueva fuente de datos/ botón atrás

SC Cancelar

Escenario	Descripción	Nombre de fuente	Tipo de fuente	Conexión	Respuesta del sistema	Flujo central
EC 1.1 Cancelar	Selecciona la opción Cancelar	NA	NA	NA	Cierra todo y sale a la página principal de la herramienta	Herramienta Pentaho/ Formulario Crear nueva fuente de datos/ botón cancelar

Descripción de las variables.

No	Nombre de campo	Clasificación	Valor Nulo	Descripción
1	Nombre de fuente	Campo de texto	No	Campo de texto donde se introduce un nombre para la fuente.
2	Conexión	Campo de selección	No	Se selecciona la conexión para conectarse al almacén de datos.
3	Tipo de fuente	Campo de selección	No	Campo de selección donde se especifica el tipo de fuente. Debe ser Postgres

- ✓ **CP_Crear conexión**

SC Crear conexión

Escenario	Descripción	Nombre de la conexión	Tipo de conexión	Acceso	Nombre del host	Nombre de la Base de datos	Numero de puerto	Nombre de usuario	Contraseña	Respuesta del sistema	Flujo central
EC 1.1 Crear conexión	Selecciona la opción crear conexión	NA	NA	NA	NA	NA	NA	NA	NA	Muestra una ventana para insertar los datos necesarios	Herramienta Pentaho
EC 1.2 Insertar datos	Inserta los datos necesarios y selecciona la opción probar	V	V	V	V	V	V	V	V	Muestra una ventana de información.	Herramienta Pentaho/ Formulario Crear conexión/ botón Test
EC 1.3 Ok	Selecciona la opción Ok	V	V	V	V	V	V	V	V	Cierra la ventana del mensaje y vuelve a la ventana anterior.	Herramienta Pentaho/ Formulario Crear conexión/ ventana mensaje/ botón ok
EC 1.4 Aceptar	Selecciona la opción Aceptar	V	V	V	V	V	V	V	V	Crea la conexión	Herramienta Pentaho/ Formulario Crear conexión/ botón aceptar

SC Cancelar

Escenario	Descripción	Nombre de la conexión	Tipo de conexión	Acceso	Nombre del host	Nombre de la Base de datos	Numero de puerto	Nombre de usuario	Contraseña	Respuesta del sistema	Flujo central

EC 1.1 Cancelar	Selecciona la opción cancelar	NA	NA	NA	NA	NA	NA	NA	NA	Cierra la ventada de crear la conexión.	Herramienta Pentaho/ Formulario Crear conexión/ botón cancelar

SC Datos incorrectos o campos incompletos

Escenario	Descripción	Nombre de la conexión	Tipo de conexión	Acceso	Nombre del host	Nombre de la Base de datos	Numero de puerto	Nombre de usuario	Contraseña	Respuesta del sistema	Flujo central
EC 1.1 Datos incorrectos	Inserta los datos incorrectos	I	I	I	I	I	I	I	I	Muestra un mensaje de error en los datos	Herramienta Pentaho/ Formulario Crear conexión/ botón aceptar
EC 1.2 Datos nulos	Deja campos incompletos	I	I	I	I	I	I	I	I	Muestra un mensaje de error en los datos	Herramienta Pentaho/ Formulario Crear conexión/ botón aceptar

Descripción de las variables.

No	Nombre de campo	Clasificación	Valor Nulo	Descripción
1	**Nombre de la conexión**	Campo de texto	No	Campo de texto donde se introduce un nombre para la conexión.
2	**Tipo de conexión**	Campo de selección	No	Campo de selección para especificar el tipo de conexión. Debe ser PostgreSQL.

3	**Acceso**	Campo de selección	No	Campo de selección para especificar por donde se va a acceder a la conexión. Debe ser Native(JDBC).
4	**Nombre del host**	Campo de texto	No	Campo de texto para especificar el nombre del host. Debe ser localhost.
5	**Nombre de la Base de datos**	Campo de texto	No	Campo de texto donde se introduce el nombre de la base de datos a la cual se le está realizando la conexión. Debe ser el nombre de la base de dato de ZERA.
6	**Numero de puerto**	Campo de texto	No	Campo de texto donde se introduce el número del puerto de la base de datos.
7	**Nombre de usuario**	Campo de texto	No	campo de texto donde se introduce el usuario de la base de datos. Debe ser postgres.
8	**Contraseña**	Campo de texto	No	Campo de texto donde se introduce la contraseña del usuario de la base de datos. Debe ser postgres.

- ✓ **CP_Visualizar reporte**

SC Visualizar reporte

Escenario	Descripción	Fuente de datos	Estilo del reporte	Formato del reporte	Respuesta del sistema	Flujo central
EC 1.1 Crear reporte	Selecciona la opción crear nuevo reporte	NA	NA	NA	Muestra una ventana para insertar los datos necesarios	Herramienta Pentaho/ botón Create New Report
EC 1.2 Seleccionar datos	Selecciona los datos necesarios y la opción siguiente	V	V	V	Muestra una ventana para seleccionar los campos que se desean mostrar en el reporte	Herramienta Pentaho/ vista crear reporte/ botón siguiente

EC 1.3 Seleccionar campos	Selecciona los campos deseados y la opción finalizar	V	V	V	Muestra el reporte con los campos seleccionados	Herramienta Pentaho/ vista crear reporte/ botón finalizar

SC Atrás

Escenario	Descripción	Fuente de datos	Estilo del reporte	Formato del reporte	Respuesta del sistema	Flujo central
EC 1.1 Atrás	Selecciona la opción atrás	NA	NA	NA	Vuelve a la vista anterior	Herramienta Pentaho/ vista crear reporte/ botón atrás

Descripción de las variables.

No	Nombre de campo	Clasificación	Valor Nulo	Descripción
1	Fuente de datos	Campo de selección	No	Nombre de la fuente de datos que hace la relación al almacén de datos.
2	Estilo del reporte	Campo de selección	No	Diseño de cómo será visto el reporte.
3	Formato del reporte	Lista desplegable	No	Formato en el que se verá el reporte. Puede ser: HTML, PDF, Excel, etc.

Anexo 3: Sentencias SQL realizadas para cargar el área de almacenamiento temporal

- ✓ **Sentencia SQL para extraer los datos de las escuela:**

  Select distinct (s.id), s.name from tb_school s

- ✓ **Sentencia SQL para extraer los datos de los programa de estudio:**

  Select distinct (sp.id), sp.name from tb_study_program sp

- ✓ **Sentencia SQL para extraer los datos de los recursos:**

```
Select distinct (sco.id), gl.title, nn.id from tb_sco sco
inner join nom_nomenclator nn on (nn.id = sco.sco_type_id )
inner join tb_general gl on (sco.id = gl.id)
inner join nom_nomenclator_type nnt on (nnt.id = nn.nomenclator_type_id and
nnt.name = 'NOM_SCO_TYPE')
```

- ✓ **Sentencia SQL para extraer los datos de los usuarios:**

```
Select distinct(u.id), u.first_name, u.first_last_name, u.second_last_name
from sf_guard_user u
```

- ✓ **Sentencia SQL para extraer los datos de los tipos de usuarios:**

```
Select distinct (g.id), g.name from sf_guard_group g
```

- ✓ **Sentencia SQL para extraer los datos de los tipos de recursos:**

```
Select nn.id, nn.type from nom_nomenclator nn
```

- ✓ **Sentencia SQL para extraer los datos de la relación usuario con tipo de usuario:**

```
Select distinct (u.id), g.id from sf_guard_user u
inner join sf_guard_user_group ug on (ug.user_id = u.id)
inner join sf_guard_group g on (g.id = ug.group_id)
```

- ✓ **Sentencia SQL para extraer los datos de las trazas:**

```
Select distinct t.begin_date, t.end_date, t.end_date - t.begin_date,
        u.id as user_id, s.id as school_id, sco.id as sco_id, nn.id as sco_type_id,
       sp.id as study_program_id from tb_school s
inner join sf_guard_user u on (u.school_id = s.id)
inner join sf_guard_user_group ug on (ug.user_id = u.id)
inner join sf_guard_group g on (g.id = ug.group_id)
inner join tb_session se on (u.id = se.user_id)
inner join r_trace_app ta on (ta.session_id = se.id)
inner join r_trace_module tm on (tm.trace_app_id = ta.id)
inner join tb_trace t on (t.trace_module_id = tm.id)
inner join tb_study_program sp on (t.message ilike '%study_program_id:'||sp.id||'%')
```

```
inner join tb_sco sco on (t.message ilike '%item_id:'||sco.id||'%')
inner join tb_general tg on (tg.id = sco.id)
inner join nom_nomenclator nn on (sco.sco_type_id = nn.id)
inner join nom_nomenclator_type nnt on (nnt.id = nn.nomenclator_type_id and
nnt.name = 'NOM_SCO_TYPE')
```

Anexo 4: Sentencia SQL realizadas para cargar el almacén de datos operacional

- ✓ **Sentencia SQL para cargar los datos de la dimensión escuela:**

```
Select distinct (s.id), s.nombre from escuela s
inner join trace on (s.id = trace.escuelaid)
```

- ✓ **Sentencia SQL para cargar los datos de la dimensión programa de estudio:**

```
Select distinct (sp.id), sp.nombre from programa_estudio sp
inner join trace on (sp.id = trace.programa_estudioid)
```

- ✓ **Sentencia SQL para cargar los datos de la dimensión recurso:**

```
Select distinct(r.id), r.nombre, tr.id from recurso r
inner join trace on (r.id = trace.recursoid)
inner join tipo_recurso tr on (tr.id = trace.tipo_recursoid)
```

- ✓ **Sentencia SQL para cargar los datos de la dimensión usuario:**

```
Select distinct(u.id), u.nombre, u.apellido, u.sapellido from usuario u
inner join trace on (u.id = trace.usuarioid)
```

- ✓ **Sentencia SQL para cargar los datos de la dimensión tipo_usuario:**

```
Select distinct(tu.id), tu.tipo from tipo_usuario tu
inner join r_usuario_tipo_usuario rutu on (tu.id = rutu.tipo_usuarioid)
inner join usuario u on(rutu.usuarioid = u.id)
inner join trace on (u.id = trace.usuarioid)
```

- ✓ **Sentencia SQL para cargar los datos de la dimensión tipo_recurso:**

```
Select distinct(nn.id), nn.tipo from tipo_recurso nn
inner join trace on (nn.id = trace.tipo_recursoid)
```

- ✓ **Sentencia SQL para cargar los datos de la dimensión usuario_tipo_usuario:**

```
Select distinct (u.id), tu.id from tipo_usuario tu
inner join r_usuario_tipo_usuario rutu on(tu.id = rutu.tipo_usuarioid)
inner join usuario u on (rutu.usuarioid = u.id)
inner join trace on (u.id = trace.usuarioid)
```

- ✓ **Sentencia SQL para cargar los datos del hecho:**

```
select distinct tc.inicio, tc.fin, tc.time, tc.usuarioid, tu.id, tc.escuelaid,
tc.programa_estudioid, r.id from recurso r
inner join tipo_recurso tr on (r.tipo_recursoid = tr.id)
inner join trace tc on (r.id = tc.recursoid)
inner join usuario u on (tc.usuarioid = u.id)
inner join r_usuario_tipo_usuario utu on (u.id = utu.usuarioid)
inner join tipo_usuario tu on (utu.tipo_usuarioid = tu.id)
```

Anexo 5: Resultados de la lista de chequeo

Estructura del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿Los entregables contienen las secciones obligatorias de la plantilla estándar definidas para un expediente de proyecto? (portada, control de versiones, reglas de confidencialidad, tabla de contenidos y contenido) (ver expediente de proyecto)	0			
crítico	1. ¿Se han identificado errores ortográficos en los entregables?	1		6	En las primeras revisiones de los

					entregables se encontraron algunas faltas de ortografía.
Indicadores definidos en el desarrollo					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
	1. ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?	0			
crítico	2. ¿Se creó el modelo físico a partir del modelo lógico?	0			
	3. ¿Se utilizó el menor número de transformaciones posibles?	0			
crítico	4. ¿Cumple la implementación del proceso de ETL con la arquitectura definida?	0			
	5. ¿Se realiza una limpieza de los datos antes de realizar la carga de los mismos?	0			
crítico	6. ¿Los reportes son configurables a través de la interfaz del sistema?	0			
crítico	7. ¿Los usuarios son capaces de manipular los resultados de manera que se ajusten a sus necesidades, conformando nuevos reportes?	0			
	8. ¿El sistema responde de una forma rápida a la información que	0			

	le sea solicitada por el usuario?				
	9. ¿Los datos e información derivados del proceso de análisis realizado mediante la aplicación, apoyan la toma de decisiones en la Institución?	0			
crítico	10. ¿Los cambios en los datos se reflejan automáticamente en los reportes de forma instantánea?	0			

Anexo 6. Encuesta para validar que el almacén de datos contribuya a la toma de decisiones.

Estimado(a) compañero(a):

Con el fin de validar los resultados de la investigación *"Almacén de datos operacional para contribuir a la toma de decisiones basado en el uso de los recursos de la Plataforma Educativa ZERA"* que tiene como objetivo brindar una herramienta que contribuya a la toma de decisiones sobre los recursos que brinda la plataforma; se solicita su participación en la siguiente encuesta.

Solución 1:

Permite generar distintos reportes, permitiendo al usuario configurar los mismos, así, podrían generarse uno o varios reportes que manejen la siguiente información:

- Recursos visitados y cantidad de visitas realizadas a un recurso específico.
- Recursos visitados por un usuario en específico.
- Listado de usuarios, mostrando los recursos que han visitados y el tiempo de visita a los mismos.

Solución 2:

- Muestra en un mismo reporte todas las acciones que realiza un usuario sobre la plataforma: Datos asociados a la materia, programa de estudio, acción, descripción de la acción, fecha de consulta y tiempo de visita a cada elemento de la Plataforma.

Considerando las características de las soluciones antes expuestas, analice las mismas y determine cuál brinda la información suficiente para tomar las siguientes decisiones:

- ✓ Se necesita eliminar los recursos menos utilizados.

Solución 1____ Solución 2 _____ Ambas _____

- ✓ Se necesita asignar los recursos no visitados a un estudiante.

Solución 1____ Solución 2 _____ Ambas _____

- ✓ Se necesita, tomando como referencia el tiempo de consulta de los estudiantes a un recurso, asignar el mismo a estudiantes con características similares.

Solución 1____ Solución 2 _____ Ambas _____

Institución:	**Categoría docente:**
	Categoría Científica:
Años de experiencia en temas relacionados con la educación a distancia:	
Labor que realiza:	

Glosario de términos

Data Mart: mercado de datos.

ETL: proceso de extracción, transformación y carga.

BI: inteligencia del negocio.

Staging area: es un área de almacenamiento temporal donde se realizan un conjunto de procesos comúnmente conocidos como extracción, transformación y carga.

UML: lenguaje visual para especificar, construir y documentar un sistema de software. Sus siglas vienen dadas por su nombre en inglés Unified Modeling Language.

XML: estándar de información cuyas siglas vienen dadas por su nombre en inglés Extensible Markup Language.

DB2: gestor de bases de datos relacional.

SQL: lenguaje de consulta estructurado o SQL (por sus siglas en inglés Structured Query Language) es un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones en ellas.

JDBC: es el acrónimo de Java Database Connectivity, una API que permite la ejecución de operaciones sobre bases de datos desde el lenguaje de programación Java, independientemente del sistema operativo donde se ejecute o de la base de datos a la cual se accede utilizando el dialecto SQL del modelo de base de datos que se utilice.

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

DATEC: Centro de Tecnologías de Gestión de Datos.

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