Universidad de las Ciencias Informáticas

Facultad 3

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

Mercado de datos para el análisis de los medios de prensa digitales en el Departamento de Operaciones Web y Análisis de Información.

Autor:

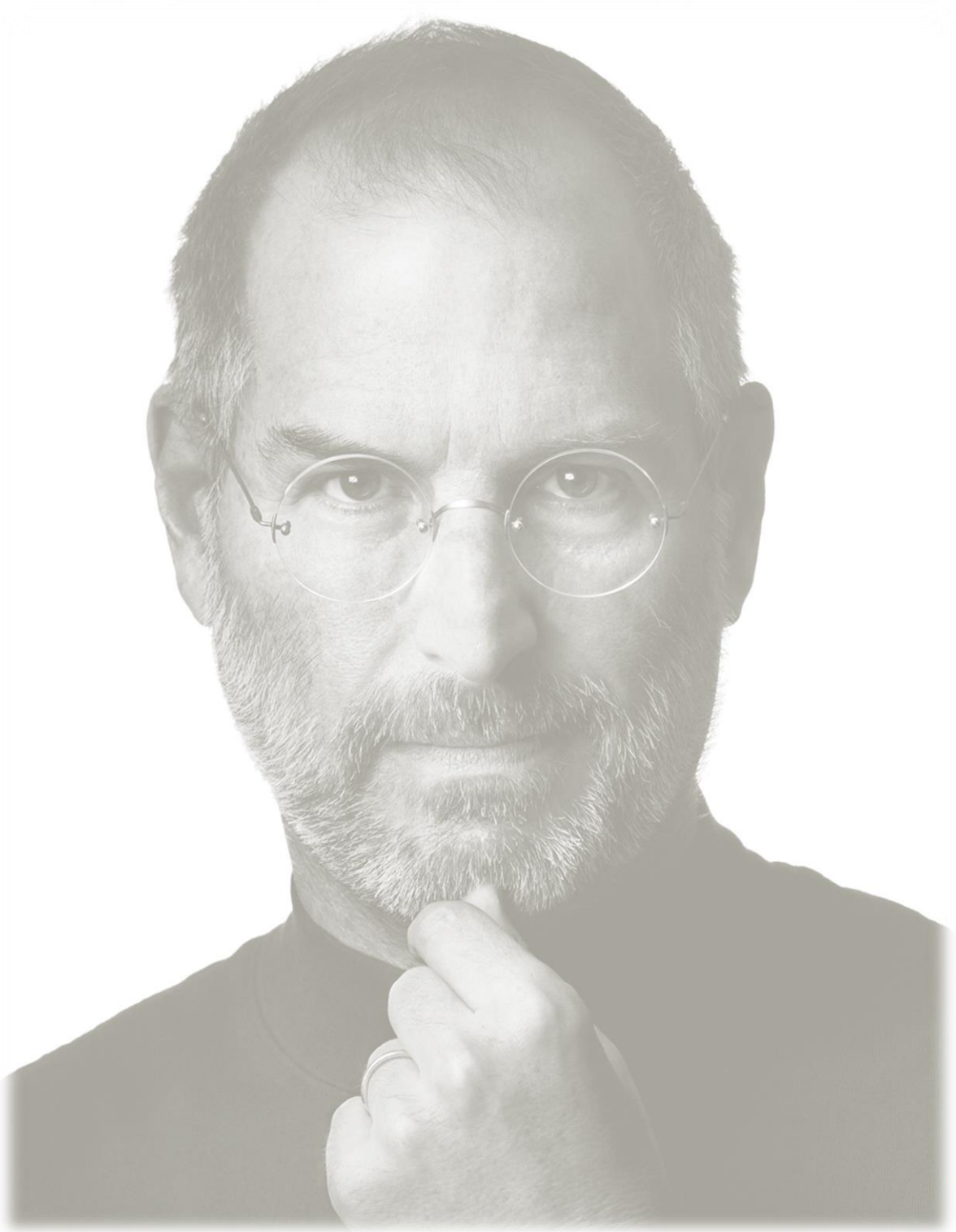
Robin Rodríguez Rodríguez

Tutor:

Ing. Darlon Antonio Santana Carvajal

Ciudad de la Habana, Cuba, junio 2015

¨ Año 57 de la Revolución ¨



La innovación distingue a los líderes de los seguidores.

Steve Jobs

Declaro ser autor del presente trabajo “Mercado de datos para el análisis de los medios de prensa digitales en el Departamento de Operaciones Web y Análisis de Información” y reconozco a la Universidad de las Ciencias Informáticas (UCI) los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los ___ días del mes de _____ del año 2015.

_____	_____
Robin Rodríguez Rodríguez	Darlon Antonio Santana Carvajal

AGRADECIMIENTOS

A mi tutor Darlon Santana por apoyarme durante todo el proceso de la investigación.

A mi compañera durante estos dos últimos años Katia Valdés Vallejo que ha sido para mí, después de mi madre, la persona más importante en mi vida.

A René Acosta por toda su ayuda incondicional y por ser mi amigo durante todos estos años.

A todos los trabajadores de DATEC que aportaron sus conocimientos para que lograra mi objetivo, Leonel Pérez, Yordanka Hechevarria e Idalmys Maza.

A todos mis compañeros que me han acompañado durante estos años.

A los profes del tribunal por corregirme cuando estuve equivocado.

DEDICATORIA

A mi mamá por apoyarme siempre en mis objetivos.

A mi papá que aunque no se encuentre conmigo siempre fue y será mi ejemplo a seguir.

A Katia Valdés Vallejo por soportarme a pesar de mis defectos y siempre estar ahí para mí cuando la necesité.

A mis abuelos Héctor Rodríguez y Oceanía Sánchez por darme todo su cariño y apoyo.

A mi abuela Xiomara que aunque se encuentre lejos siempre ha estado conmigo dándome todo el apoyo.

A mi hermano Alois que aunque tengamos nuestras cosas siempre lo voy a querer.

A mis amigos de toda la vida.

RESUMEN

Un mercado de datos permite realizar análisis estadísticos de procesos en una entidad; estas herramientas posibilitan almacenar datos históricos, obtener una visión de la evolución del negocio y tomar decisiones estratégicas. El proceso de análisis estadístico de los medios de prensa digitales en DOWAI presenta inconvenientes en la obtención de información de los datos recopilados debido principalmente al elevado volumen de estos, la falta de mecanismos automatizados completamente y la dependencia de herramientas basadas en aplicaciones de oficina como Excel para su almacenamiento. Mediante la presente investigación se desarrolló el mercado de datos que permitió centralizar y visualizar la información de manera eficiente y enlazar las variables requeridas para su estudio. Se utilizaron las herramientas PostgreSQL para el almacenamiento y las herramientas de Pentaho para los procesos ETL y de inteligencia de negocio. El sistema implementado se validó a través de listas de chequeo, pruebas de rendimiento utilizando la herramienta Apache-JMeter, pruebas de estrés y culminando con la carta de aceptación del cliente.

Palabras claves: análisis estadísticos, DOWAI, datos históricos, inteligencia de negocio, Mercado de datos.

CONTENIDO

INTRODUCCIÓN

En las últimas décadas el desarrollo de las Tecnologías de la Informática y las Comunicaciones (TIC) y sus grandes avances tecnológicos han causado un aumento considerable de la información. Debido al elevado volumen de datos almacenados principalmente por las empresas, se hace necesario una estructura capaz de soportar este crecimiento sin precedentes. Es así que surgen las Bases de Datos (BD). Estas son una herramienta indispensable en la actual sociedad de la información, su utilidad no sólo se debe a que es un conjunto de datos almacenados de alguna forma determinada, en una BD también existen una cantidad de elementos que ayudan a organizar sistemáticamente, relacionar, proteger, y administrar de manera eficiente los datos.

Las BD, pueden clasificarse de dos maneras atendiendo a la información que manejan: BD estáticas y dinámicas. En las dinámicas la Información almacenada se modifica con el tiempo, permitiendo operaciones como actualización, borrado y adición de datos, además de las operaciones fundamentales de consulta.

En cambio, las estáticas son de sólo lectura, utilizadas primordialmente para almacenar datos históricos que posteriormente se pueden utilizar para estudiar el comportamiento de un conjunto de datos a través del tiempo, realizar proyecciones y tomar decisiones. Por sus características le ha permitido ser la variante adoptada en proyectos que requieren de Inteligencia de Negocio.

Un tipo particular de BD son las de prensa, utilizadas entre otros por las herramientas de monitoreo de noticias, que permiten consultar noticias publicadas en muchos medios en internet. Este tipo de BD es un valioso instrumento para las búsquedas en un conjunto importante de información o en un largo período de tiempo. Permiten una localización rápida de los artículos y, en algunos casos, realizar búsquedas en el texto íntegro.

Sin embargo los datos por si solos no aportan ningún beneficio a la entidad u organización que los posea. Para que sean útiles tienen que brindar información que contribuya al desarrollo del negocio, que permita realizar análisis sobre ella aplicando diversas tecnologías y así reducir la incertidumbre sobre algún aspecto

y dar soporte a procesos de la organización, prever su evolución y tomar decisiones estratégicas para el futuro.

En Cuba, en la Facultad 1 de la Universidad de las Ciencias Informáticas (UCI) se encuentra el Departamento de Operaciones Web y Análisis de Información (DOWAI). El mismo es una estructura que procesa contenido de medios de información digitales, así como su comportamiento a partir del análisis estadístico. El objetivo principal del departamento es apoyar la toma de decisiones estratégicas. Una de las líneas con las que cuenta este departamento es la de Análisis de Información, que tiene la finalidad de presentar cifras estadísticas e indicadores que reflejen el desarrollo de la actividad o proceder de los medios de prensa en internet. Las principales deficiencias que posee DOWAI con el manejo de la información referente a dicha área son:

- El análisis estadístico se realiza a través de mecanismos no automatizados completamente, poco confiables y en algunas ocasiones resulta tedioso, a causa de que la información es almacenada en herramientas basadas en aplicaciones de oficina como Excel.
- La información solamente puede ser accedida por especialistas que tengan conocimientos del negocio.
- Se generan un gran número de datos anuales obstaculizando su análisis.
- Poseen variadas versiones de los datos lo que desencadena la mala calidad de los mismos y origina su no integración.
- La recuperación y creación de los informes se torna engorroso y a veces costoso en cuanto a tiempo y esfuerzo.

Todo esto deteriora la calidad de la información en cuanto a seguridad, disponibilidad e integridad, dificultando así el análisis estadístico de diferentes variables relacionadas con los datos que se procesan en el área de Análisis de Información del departamento DOWAI. Como resumen existen dificultades para almacenar, recuperar y presentar la información proveniente de los medios digitales de prensa tales como: principales reportes, cruces de variables, indicadores y porcentajes dificultando así la toma de decisiones.

A partir de lo anteriormente planteado surge como **problema a resolver:**

¿Cómo contribuir al análisis de los medios de prensa digitales para la toma de decisiones en el Departamento de Operaciones Web y Análisis de Información?

Para dar solución al problema científico se define como **objeto de estudio:** Almacenes de datos, enmarcado en el **campo de acción**: Mercados de datos para el proceso de análisis de información.

Objetivo general:

Desarrollar un mercado de datos para el análisis de los medios de prensa digitales en el Departamento de Operaciones Web y Análisis de Información.

Objetivos específicos:

1. Fundamentar la selección de la metodología y herramientas a utilizar para la investigación.
2. Realizar el análisis y diseño del mercado de datos para el análisis de los medios de prensa digitales para el Departamento de Operaciones Web y Análisis de Información.
3. Implementar el mercado de datos para el análisis de los medios de prensa digitales para el Departamento de Operaciones Web y Análisis de Información.
4. Validar el mercado de datos para el análisis de los medios de prensa digitales para el Departamento de Operaciones Web y Análisis de Información.

Se define como **idea a defender**: El desarrollo de un mercado de datos en el Departamento de Operaciones Web y Análisis de Información contribuirá al análisis de los medios de prensa digitales para la toma de decisiones.

Como **métodos de investigación teóricos** los siguientes:

- **Histórico-Lógico:** A través de este método se analiza el desenvolvimiento de los mercados de datos en el tiempo y la aplicación de las mejores prácticas en su desarrollo con herramientas de código abierto. También se establecieron las posibles fuentes de datos del departamento DOWAI para realizar el proceso de análisis.

- **Analítico-Sintético:** Este método permite el análisis y selección de las metodologías existentes para el desarrollo del mercado de datos según las características de la institución.
- **Modelado:** Se define el conjunto de modelos, arquitectura y esquemas para el desarrollo del mercado de datos.

Aportes de la investigación

- Informe detallado con toda la base teórico-práctico sobre la cual se sustente la solución propuesta.
- Mercado de datos para el Departamento de Operaciones Web y Análisis de Información.

La presente investigación está estructurada en 3 capítulos que son abordados de la siguiente manera:

Capítulo 1: Fundamentación teórica

En este capítulo se abordan los principales conceptos tratados en la investigación. La importancia que tiene para una organización un adecuado control estadístico, la descripción del proceso estadístico a modelar, la metodología a seguir en el diseño e implementación de un mercado de datos, y la selección de herramientas útiles para llevar a cabo un excelente almacenamiento de la información.

Capítulo 2: Análisis y diseño del mercado de datos

Este capítulo contiene la descripción de los pasos a seguir durante el análisis y el diseño de la solución. Se abordan aspectos concernientes a la descripción de las fuentes a integrar. Se definen los requisitos que debe cumplir el sistema, así como el modelo dimensional propuesto para el desarrollo del mercado de datos a partir de los indicadores que se seleccionaron.

Capítulo 3: Implementación y validación del mercado de datos

En este capítulo se efectúan todos los procesos de extracción, transformación y carga de los datos, los flujos de integración y los trabajos para enlazar todas las

transformaciones. Además, se realizan los procesos de inteligencia de negocio donde se implementa y prueba el modelo de datos, así como los cubos OLAP.

CAPÍTULO 1: Fundamentación teórica

1.1 Introducción

Desde su surgimiento, las bases de datos se convirtieron en la herramienta fundamental para almacenar gran cantidad de información. Debido a esto, en muy poco tiempo la información almacenada por las grandes empresas y negocios alcanzaron una dimensión considerablemente voluminosa. Con la acumulación de esta información se presentó la problemática de cómo darle un fin útil.

La solución sería unificar las diferentes fuentes de información de las cuales disponían, en un único lugar, al que sólo se le incorporaría información relevante, sobre la base de una estructura integrada. La respuesta a esto fueron los almacenes de datos como se conocen mundialmente.

En este capítulo se van a tratar aspectos sobre los almacenes de datos, conceptos, características, ventajas, componentes, metodologías y herramientas a utilizar en su implementación.

1.2 Estadística, medir para gestionar

La estadística es comúnmente considerada como una colección de hechos numéricos expresados en términos de una relación sumisa, y que han sido recopilados a partir de otros datos numéricos.

No podemos gestionar lo que no se mide. Las mediciones son la clave. Si no se puede medir, no se puede controlar. Si no se puede controlar, no se puede gestionar. Si no se puede gestionar, no se puede mejorar. La falta sistemática o ausencia estructural de estadísticas en las organizaciones impide una administración científica de las mismas. Dirigir sólo en base a datos financieros del pasado, realizar predicciones basadas más en la intuición o en simples extrapolaciones, y tomar decisiones desconociendo las probabilidades de éxito u ocurrencia, son sólo algunos de los problemas o inconvenientes más comunes hallados en las empresas. (1)

Carecer de datos estadísticos en cuanto a lo que acontece tanto interna como externamente, impide decidir sobre bases racionales, y adoptar las medidas preventivas y correctivas con el suficiente tiempo para evitar daños, en muchos casos irreparables, para la organización. (1)

1.2.1 Aplicaciones estadísticas existentes

Como resultado de la experiencia a lo largo de la existencia de una empresa, estas acumulan gran cantidad de información, considerada como un activo fundamental, en la toma de decisiones futuras. La conversión de esa información en conocimiento se ha convertido en la única fuente de competitividad sostenible, para incrementar su eficiencia, elevar la eficacia y su posición en el mercado. Viéndose obligadas las empresas a implementar herramientas para viabilizar esta operación.

1.2.2 Aplicaciones estadísticas en el mundo que utilizan almacenes de datos

Google Estadísticas de Búsqueda, también conocido como Google Insights, es una de las herramientas del conocido buscador (Google) que analiza y compara resultados de términos teniendo en cuenta distintas variables. (2)

Para usar Google Estadísticas de Búsqueda tienes tres posibilidades: hacerlo tomando como base los términos de búsqueda, la zona geográfica o el periodo de tiempo. Una vez elegida tu opción, introduce los términos y los gráficos aparecerán en tiempo real indicando la evolución en la tendencia de búsquedas. (2)

Para muchos de los usuarios más activos de Twitter revisar los trending topics se ha convertido en la forma más inmediata de informarse sobre los acontecimientos que estén sucediendo en el mundo. Los trending topics indican cuáles son las palabras más mencionadas más rápidamente a lo largo y ancho de la red social. (3)

1.2.3 Aplicaciones estadísticas en Cuba que utilizan almacenes de datos

El Observatorio del Periodismo Cubano en el Centro de Información para la Prensa (CIP) surgió por la necesidad de contar con un instrumento dedicado al

aprovechamiento sistemático de la información y el conocimiento y para elevar al máximo la capacidad de respuesta en lo que se refiere al mensaje cubano tanto a la opinión pública nacional como internacional. (4)

Con herramientas de software desarrolladas específicamente para estas tareas y la aplicación de novedosas estrategias de análisis, desde el observatorio cubano se genera el material necesario para medirle el ritmo a la información nacional y diagnosticar las tendencias sobre la presencia del tema Cuba a nivel mundial. (4)

Se conoce sobre la existencia de otras herramientas en instituciones como el MINREX, la Mesa Redonda y el Noticiero Nacional pero no se pudo obtener acceso a la información sobre los mismos debido a que no se encuentra pública.

1.3 Almacenes de Datos, conceptualizando

En el proceso de Inteligencia de Negocios[1] (IN) una herramienta característica la conforma los Almacenes de Datos (AD). Existen múltiples definiciones de los AD, a continuación se citan algunas de ellas:

Un Almacén de Datos o Datawarehouse, es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta. (5)

William H. Inmon, considerado el padre de los AD los define como: *una colección de datos orientada al sujeto, integrados, variantes en el tiempo y no volátiles que soportan el proceso administrativo de soporte a las decisiones. Su definición es útil porque utiliza atributos que pueden medirse.* (6)

[1] **Inteligencia de negocio**: es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios. Enmarca las distintas estrategias encaminadas a resolver problemas de administración de información y creación de conocimiento, y por lo tanto, constituye un componente esencial de los Almacenes de datos

Según Ralph Kimball en la segunda edición de su libro The Datawarehouse Toolkit, *un almacén de datos es una copia de los datos transaccionales específicamente estructurada para la consulta y el análisis.* (7)

Después de analizar las bibliografías y definiciones particulares de estos autores se puede definir que un AD es una estructura donde se almacenan datos orientados a un tema específico donde la información va a ser persistente y efectiva a lo largo del tiempo con el objetivo de dar soporte a la toma de decisiones de las empresas y organizaciones.

1.3.1 Características de los AD

William H. Inmon, define algunas de sus principales características:

1. **Temático**: Los datos contenidos en el AD están orientados a un tema o materia específicos. Esto posibilita a los usuarios finales un mejor entendimiento y acceso a la información. Además, las respuestas a los pedidos de información de un tema en particular se realizan de una forma más ágil ya que toda esta contenida en un mismo lugar (Anexo 1).
2. **Integrado**: Los datos que se encuentran en el AD están integrados en una estructura sólida. No contiene datos que no son de utilidad para los clientes, tampoco tiene datos redundantes (Anexo 2).
3. **No volátil**: En el AD solo se realizan tres operaciones fundamentales: carga periódica, carga inicial o histórica y consulta. No se actualizan ni eliminan los datos (Anexo 3).
4. **Histórico**: Los datos contenidos en un AD reflejan el estado de la actividad del negocio en el momento presente, a medida que el AD tiene más datos será más útil pues brindará la posibilidad de establecer comparaciones en diferentes períodos (Anexo 4).
5. **Contiene metadatos**: es decir, datos sobre los datos. Los metadatos permiten saber la procedencia de la información, su periodicidad de actualización y su fiabilidad.

Un AD generalmente no es estático pues debe crecer y cambiar a medida que lo vayan haciendo las necesidades del negocio. Esto significa que deben ser diseñados para adaptarse de forma constante ya que no es posible predecir los

requerimientos de la información que habrá en el futuro, pues mientras el negocio crezca estos requerimientos cambiarán.

1.3.2 Ventajas y desventajas de los AD

La existencia de los AD ha generado gran cantidad de posibilidades a las empresas para controlar sus actividades. Entre las ventajas que propicia el uso de AD están:

- ✓ Facilita la toma de decisiones en los negocios. Con la implementación de un AD se crea una herramienta que permite dar soporte a al proceso de toma de decisiones en las empresas basándose en información general del negocio. Se le puede sacar un gran provecho porque como contiene datos históricos, se puede aprender de estos y con este conocimiento se predicen situaciones que se presentarían en un futuro.
- ✓ Proporciona una comunicación fiable entre todos los departamentos de la empresa. Lo que permite a los usuarios medir los resultados de una mejor forma y a partir de esto, establecer prioridades en las acciones a realizar y decisiones a tomar con respecto a un determinado cliente.
- ✓ En un contexto de objetivos definidos en los negocios permite a empresas, explorar automáticamente, visualizar y comprender los datos e identificar patrones, relaciones y dependencias que impactan en los resultados finales de la cuenta de resultados (tales como el aumento de los ingresos, incremento de los beneficios, contención de costes y gestión de riesgos).
- ✓ El uso de los AD en los negocios posibilita que los usuarios comprendan mejor el mismo, así como también se garantiza la ejecución de las consultas de manera rápida y eficiente.

Entre las principales desventajas de un AD se encuentran el costo y la estandarización, su dinámica y exclusividad dificulta su mantención.

1.3.3 Representación arquitectónica de un AD

Con la arquitectura del AD se describe el flujo de los datos desde su obtención de las fuentes contenedoras, hasta estar listos para su utilización por la empresa. La arquitectura está compuesta por tres subsistemas (Figura 1):

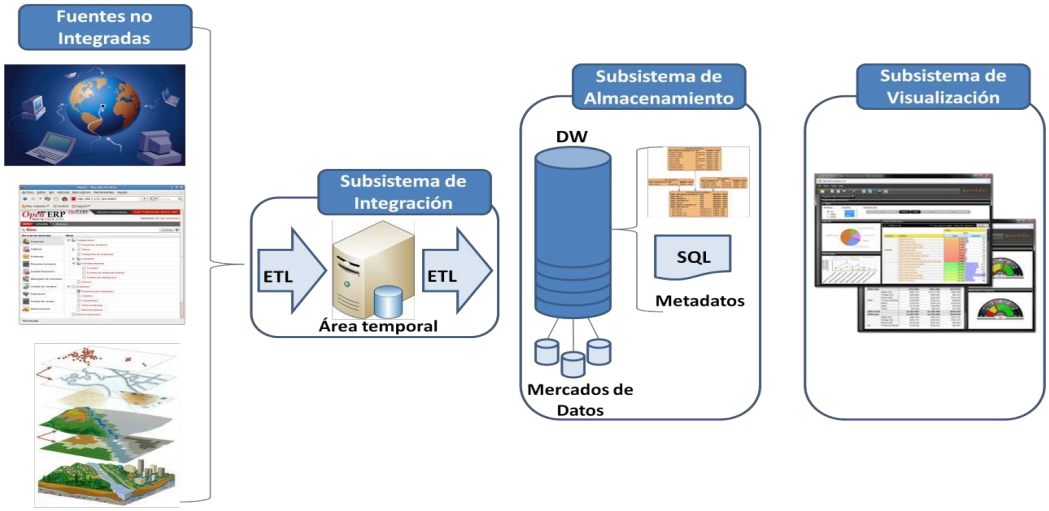


Figura 1: Arquitectura general de subsistemas de un AD (8)

Subsistema de integración de datos: Es donde se realizan los procesos de Extracción, Transformación y Carga (Extraction, Transformation and Load - ETL). Es la capa encargada de conectarse a las fuentes de datos y extraer la información que nutrirá el AD, realizar las transformaciones necesarias y finalmente carga la información en el Repositorio de Datos.

Subsistema de almacenamiento: Es el núcleo principal de la solución, son las estructuras de datos que soportan el almacenamiento de la información para que pueda ser consultada.

Subsistema de visualización de información: Es la capa que presenta la información al usuario final para su análisis. Los usuarios realizan este análisis por medio de herramientas de Procesamiento Analítico en Línea (On Line Analitycal Procesing - OLAP) y de técnicas de inteligencia de negocio que permiten explorar los datos almacenados y obtener conocimiento a partir de ellos.

1.4 Mercado de datos

Una estructura muy particular de los AD son los mercados de datos (MD). Los mismos son bases de datos departamentales, especializadas en el almacenamiento de los datos de un área del negocio específica. 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. Un mercado de datos puede ser alimentado desde los datos de un almacén de datos, o integrar por sí mismo un compendio de distintas fuentes de información. (5)

Según la definición anterior podemos decir que los MD son AD orientados a temas o aplicaciones específicas y contienen datos de sólo una línea del negocio. La mayor diferencia entre ellos es el ámbito de la información que contienen debido a que en los mercados es más pequeño y los datos se obtienen de un menor número de fuentes y comúnmente el tiempo de desarrollo es menor.

1.4.1 Características de un MD

Se caracteriza por disponer de una estructura de datos que analiza la información al detalle desde las perspectivas que afectan los procesos del departamento. Este puede obtener los datos desde un almacén o puede integrar un compendio de distintas fuentes de información.

1. Poseen características similares de integración, no volatilidad y orientación temática que un AD.
2. Representan una estrategia de divide y vencerás para ámbitos muy genéricos de un AD.
3. Los mercados de datos se enfocan a los requisitos de los usuarios que están asociados a un departamento específico de la empresa.
4. Su utilización y comprensión es sencilla debido que contienen menor número de información que los AD.
5. Poseen menor alcance histórico que los AD.

1.4.2 Ventajas del uso de los MD

- Mayor rapidez en las respuestas a las consultas y que su elaboración sea más fácil ya que poseen menos volumen de información.
- Permite realizar consultas SQL[2] o MDX[3] sencillas facilitando el acceso a los datos que son utilizados con frecuencia.
- Simplifican el desarrollo de todo el mecanismo de su base de datos y con ello baja substancialmente todo el coste del proyecto, así como su duración.
- Simples de implementar.
- Poco tiempo de construcción y puesta en marcha.
- Permiten manejar información confidencial.
- Reflejan rápidamente sus beneficios y cualidades.

1.5 Modelo multidimensional

Debido al enfoque analítico de la tecnología de los almacenes surge una forma diferente de pensamiento y procesamiento lo que se evidencia en un modelado de bases de datos propio llamado modelo multidimensional. El modelado dimensional es una técnica para modelar bases de datos simples y entendibles al usuario final. La idea fundamental es que el usuario visualice fácilmente la relación que existe entre los distintos componentes del modelo. (4)

En los AD es más conveniente utilizar un modelo multidimensional (MMD) atendiendo que este posee sus ventajas con respecto al modelo entidad-relación (MER) aunque ambas almacenan la misma información, el MMD se representa a través de las tablas de hechos con sus concernientes tablas de dimensiones (Anexo 5).

---

[2] SQL Lenguaje de consulta estructurado: es un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones en ellas.
[3] MDX Expresiones multidimensionales: es un lenguaje de consulta para bases de datos multidimensionales sobre cubos OLAP, se utiliza en IN para generar reportes para la toma de decisiones basados en datos históricos, con la posibilidad de cambiar la estructura o rotación del cubo.

1.5.1 Características del modelo multidimensional

La estructura básica de un AD está definida por dos elementos principales, esquemas y tablas. Hay dos tipos de tablas en el modelo multidimensional:

- **Tablas de hechos:** Representan la ocurrencia de un determinado proceso dentro de la organización y no tienen relación entre sí. Generalmente, almacenan medidas numéricas, las que representan valores de las dimensiones, aunque en ocasiones estas no están presentes y se les denominan “tablas de hechos sin hechos”, es decir, la relación entre las dimensiones que definen la llave de esta tabla de hecho implica por sí sola la ocurrencia de un evento. La llave de la tabla de hecho, es una llave compuesta, debido a que se forma de la composición de las llaves primarias de las tablas dimensionales a las que está unida (Anexo 6).

- **Tablas de dimensiones:** Contienen, generalmente, una llave simple y atributos que la describen. En dependencia del esquema de diseño que se asuma pueden contener llaves foráneas de otras tablas de dimensión. Existe una dimensión fundamental en todo AD, la dimensión tiempo. Esto ocurre porque todo registro que se incluya constituye la ocurrencia de un fenómeno en un instante de tiempo definido. Dicha dimensión es la que establece uno de los objetivos fundamentales de la construcción de un AD, la conservación de un “histórico”. Los atributos dimensionales son fundamentalmente textos descriptivos, estos juegan un papel determinante porque son la fuente de gran parte de todas las necesidades que deben cubrirse, además, sirven de restricciones en la mayoría de las consultas que realizan los usuarios. Esto significa, que la calidad del modelo multidimensional, dependerá en gran parte de cuan descriptivos y manejables, sean los atributos dimensionales escogidos. La dimensión más importante de un AD, es la dimensión tiempo, atendiendo que esta será la encargada de decir en qué momento ocurrió este hecho (Anexo 6).

1.5.2 Esquemas

- **Esquema de estrella:** La tabla de hechos se encuentra en el medio de la estrella y se relacionan a ella las tablas de dimensiones, no existen relaciones entre las mismas. Este esquema es ideal por su simplicidad y velocidad para ser usado en análisis multidimensionales. Es la opción con mejor rendimiento y velocidad pues permite indexar las dimensiones de forma individualizada sin que repercuta en el rendimiento de la base de datos en su conjunto. (Figura 2).

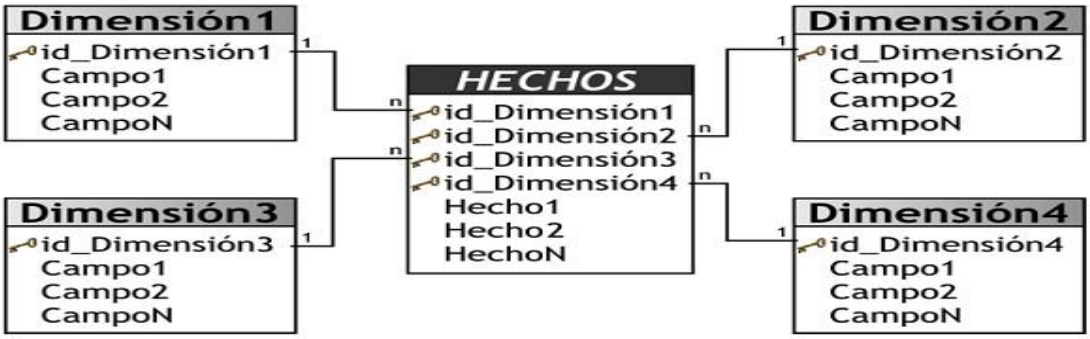


Figura 2: Esquema de estrella (9)

- **Esquema de copo de nieve:** Es similar al de estrella pero se evidencia jerarquía entre las dimensiones, las cuales están relacionadas, o sea existen caminos alternativos entre ellas. La finalidad es normalizar las tablas y así reducir el espacio de almacenamiento al eliminar la redundancia de datos; pero tiene la contrapartida de generar peores rendimientos al tener que crear más tablas de dimensiones y más relaciones entre las tablas (Figura 3).

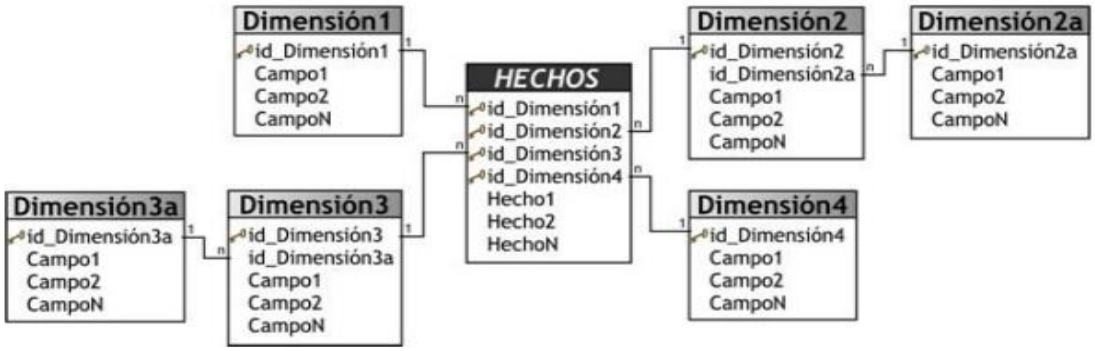


Figura 3: Esquema de copo de nieve (9)

- **Esquema constelación:** Está compuesto por una serie de esquemas en estrella. Lo conforman una tabla de hechos principal y una o más tablas de hechos auxiliares las cuales pueden ser resúmenes de la principal (Figura 4).

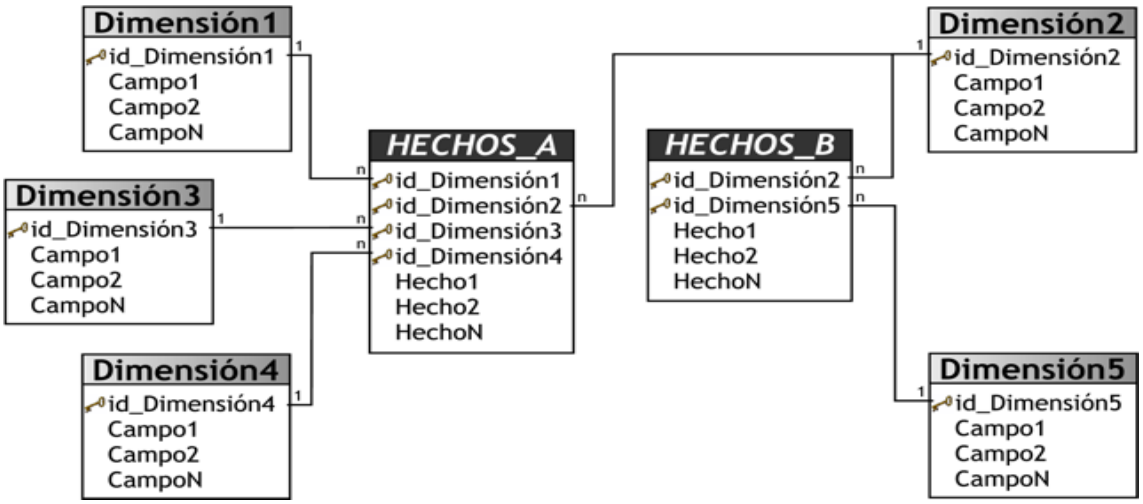


Figura 4: Esquema de constelación (9)

1.5.3 Modo de almacenamiento de datos OLAP

Procesamiento analítico en línea OLAP, estas herramientas se basan en el modelo multidimensional de datos, presentando a los usuarios una visión multidimensional de los datos, independientemente del servidor que soporte el AD. Permiten analizar las bases de datos de gran volumen de manera multidimensional lo que muestra cualquier correlación dentro de un volumen de datos importante del sistema de información con el fin de detectar alguna tendencia.

1.5.4 Razón de recurrir a OLAP para las consultas:

- **Velocidad de respuesta:** Una base de datos relacional almacena entidades en tablas discretas si han sido normalizadas. Esta estructura es buena en un sistema Procesamiento de Transacciones En Línea (OnLine Transaction Processing - **OLTP**) pero para las complejas consultas multitabla es relativamente lenta.

- **Un modelo mejor para búsquedas:** La principal característica que potencia a OLAP, es que es lo más rápido a la hora de ejecutar sentencias SQL de tipo **SELECT.**

1.5.5 Implementaciones OLAP

Existen tres tipos de implementaciones OLAP, cada una con sus peculiaridades según las exigencias del almacenamiento de los datos.

ROLAP (OLAP Relacional):

- Almacena los datos en un motor relacional.
- Los esquemas más comunes sobre los que se trabaja son estrella o copo de nieve, aunque es posible trabajar sobre cualquier base de datos relacional.
- La arquitectura está compuesta por un servidor de banco de datos relacional y el motor OLAP se encuentra en un servidor dedicado.
- La principal ventaja de esta arquitectura es que permite el análisis de una enorme cantidad de datos.

MOLAP (OLAP Multidimensional):

- Almacena los datos en una base de datos multidimensional.
- Para optimizar los tiempos de respuesta, el resumen de la información es usualmente calculado por adelantado.
- Estos valores pre calculados o agregaciones son la base de las ganancias de desempeño de este sistema.
- Algunos sistemas utilizan técnicas de compresión de datos para disminuir el espacio de almacenamiento en disco debido a los valores pre calculado.

HOLAP (OLAP Híbrido):

- Almacena algunos datos en un motor relacional y otros en una base de datos multidimensional.
- Es la integración de los anteriores.

1.6 Metodologías para el desarrollo de los AD

Metodología es un vocablo generado a partir de tres palabras de origen griego: *metà* (**"**más allá**"**), *odòs* (**"**camino**"**) y *logos* (**"**estudio**"**). El concepto hace referencia al plan de investigación que permite cumplir ciertos objetivos en el marco de una **ciencia**. Puede entenderse a la metodología como el conjunto de **procedimientos** que determinan una investigación de tipo científico o marcan el rumbo de una exposición doctrinal. (10)

Existen gran variedad de metodologías para la creación de AD, cada experto ha adaptado su forma de trabajo creando una metodología que satisfaga sus necesidades. Existen diferentes modelos que las metodologías deben seguir para crear un AD, claro está dependiendo del objetivo de la empresa será la selección del modelo que más se adecue.

- **Modelo Top Down** traducido al español (de arriba hacia abajo): tiene como base un sistema de AD para toda la empresa y a partir de este se desarrollan los MD para cada departamento. Está basado en la estructura del almacén central, la cual se construye de los datos que se obtienen de los sistemas operacionales o externos (datos aislados) a través de un proceso de extracción, transformación y carga (ETL). El enfoque top-down se adapta a la visión de Bill Inmon.
- **Modelo Bottom Up** traducido al español (de abajo hacia arriba): tiene como base los distintos MD de los departamentos y a partir de estos se construye el almacén principal para toda la empresa. Este modelo se construye a través de dos procesos diferentes de ETL, un primer proceso para la construcción de cada mercado con los datos aislados que son necesarios para las funciones de cada departamento y un segundo proceso en el sentido de los MD hacia el AD. Este modelo es defendido por Ralph Kimball.

1.6.1 Metodología R. Kimball

La metodología se basa en lo que Kimball denomina Ciclo de Vida Dimensional del Negocio (Business Dimensional Lifecycle). Este ciclo de vida (Anexo 8) del proyecto de AD, está basado en cuatro principios básicos (8):

- Centrarse en el negocio: Hay que concentrarse en la identificación de los requerimientos 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 requerimientos de negocio identificados en la empresa.
- Realizar entregas en incrementos significativos: Crear el AD en incrementos entregables en plazos de 6 a 12 meses. Hay que usar el valor de negocio de cada elemento identificado para determinar el orden de aplicación de los incrementos. En esto la metodología tiene similitud con las metodologías ágiles de construcción de software.
- Ofrecer la solución completa: Proporcionar todos los elementos necesarios para entregar valor a los usuarios de negocios. Esto significa tener un almacén de datos sólido, bien diseñado, con calidad probada, y accesible.

La metodología se divide en cuatro fases donde se realizan un grupo de actividades:

Fase I - Requerimientos y gestión del proyecto

- Definición del proyecto.
- Gestión y planeación del proyecto.
- Gestión y planeación del programa.
- Definición de requisitos del negocio.

Fase II - Arquitectura técnica AD/IN

- Diseño de la arquitectura.
- Selección de productos.
- Gestión de metadatos.
- Implementación de medidas tácticas de seguridad.
- Desarrollo del plan estratégico de seguridad.
- Desarrollo del plan de infraestructura.

- Instalación de productos.

Fase III – Diseño e implementación

- Diseño del modelo de datos dimensional.
- Diseño físico de la BD.
- Implementación física de la BD.
- Diseño del subsistema ETL.
- Desarrollo del subsistema de ETL.
- Diseño del subsistema de IN.
- Desarrollo del subsistema de IN.

Fase IV - Implantación y operaciones

- Pruebas de pre-implantación.
- Pruebas de datos y procesos.
- Optimización del rendimiento.
- Implantación del sistema.
- Capacitación y transferencia tecnológica.
- Operaciones de mantenimiento.
- Operación de soporte.

1.6.2 Metodología Hefesto

La metodología Hefesto creada por el Ing. Bernabeu Ricardo Dario en su libro ¨HEFESTO: Metodología para la Construcción de un Data Warehouse¨, permite la construcción de un AD de forma sencilla, ordenada e intuitiva. Hefesto es una metodología bien fundamentada y explícita que permite desarrollar un almacén de datos de manera metódica y sencilla, guiándose por pasos lógicos relacionados sólidamente durante todas las etapas del proceso de confección (9). Hefesto tiene como principales características las siguientes:

- Los objetivos y resultados esperados en cada fase se distinguen fácilmente y son sencillos de comprender.
- Se basa en los requerimientos 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 AD.
- Utiliza modelos conceptuales y lógicos, los cuales son sencillos de interpretar y analizar.
- Es independiente del tipo de ciclo de vida que se emplee para contener la metodología.
- Es independiente de las herramientas que se utilicen para su implementación.
- Es independiente de las estructuras físicas que contengan el AD 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 AD como para MD.

Esta metodología es ágil y madura, propone un conjunto de fases (Figura 6) que con pocos recursos, tiempo y documentación, permite realizar un AD. Lo que se busca es entregar una primera implementación que satisfaga una parte de las necesidades, para demostrar las ventajas del AD y motivar a los usuarios.

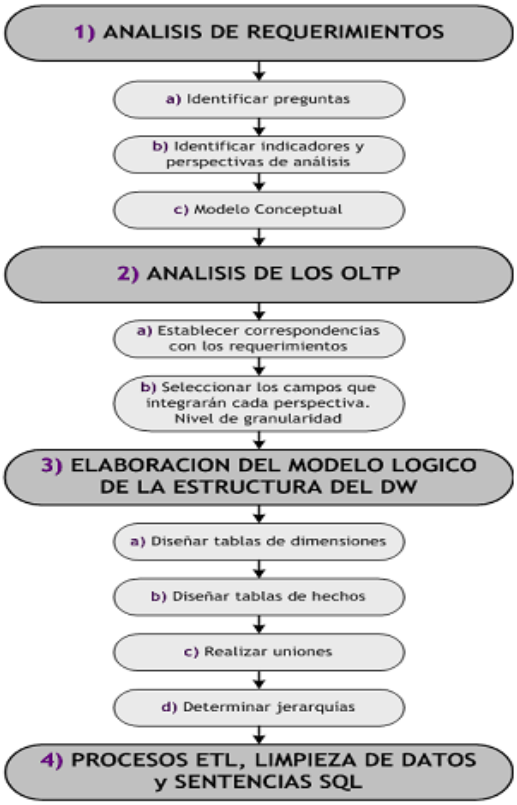


Figura 6: Fases de la metodología Hefesto. (9)

1.6.3 Metodología de W. Inmon.

W. Inmon define su metodología en el año 1992 en el libro “Building the Data Warehouse”, donde propone los mecanismos necesarios para llevar a cabo la correcta realización de un AD. A Inmon se le asocia con los AD a nivel empresarial, que involucran desde un inicio todo el ámbito corporativo, sin centrarse en un área específica hasta después de haber terminado completamente el diseño del AD. En su filosofía, un MD es sólo una de las capas del AD y los MD son dependientes del depósito central de datos o AD Corporativo y por lo tanto se construyen después de él. Inmon es defensor de utilizar el modelo relacional para el ambiente para el diseño del AD Corporativo. Afirma, que la creación de una base de datos relacional con una ligera normalización, es la base de los MD, o lo que es lo mismo, a partir de los esquemas relacionales, a los que se les irán añadiendo complejidad, se obtendrán finalmente los MD. (8)

1.6.4 Metodología seleccionada

Después de realizar un análisis de las metodologías existentes para la realización de AD se selecciona Hefesto para dar solución al problema de la investigación, sin obviar la posible inclusión de características de las otras metodologías existentes. Esta metodología resulta ser la escogida ya que se puede adaptar fácilmente a las características existentes del negocio lo que no va a modificar el funcionamiento del mismo. Se basa en los requisitos del cliente, permitiendo que se puedan hacer cambios en el futuro de ser necesarios para ampliar el alcance de esta solución llevándola a otros procesos del sistema. También resulta importante la agilidad de esta metodología, así como la utilización de pocos recursos en la realización de sus fases.

1.7 Herramientas a utilizar en la construcción del mercado de datos

Existen gran número herramientas que facilitan la realización de trabajos de diseño e implementación de aplicaciones informáticas. A continuación se hace una descripción de las utilizadas para llevar a cabo la solución de la investigación.

La Open BI Suite de Pentaho provee un completo espectro de funcionalidades de IN, incluyendo reportes, análisis, tableros de control, minería de datos, integración de datos y una plataforma de IN que la han convertido en la suite de código abierto más popular del mundo. (11)

1.7.1 Herramienta de modelado

Visual Paradigm versión 8.0

Se selecciona esta herramienta CASE (computer aided software engineering, ingeniería de software asistida por computadora) porque se ajusta a las políticas de utilización de software libre en la UCI y es la herramienta de modelado definida a utilizar por el centro de desarrollo de DOWAI. Las características que la hacen confiable y robusta para dar solución al modelado son las siguientes: (12)

- Disponibilidad en múltiples plataformas (Windows, Linux) y en varios idiomas.
- Permite realizar diseños centrados en casos de uso y enfocados al negocio que generan un software de mayor calidad.
- Uso de un lenguaje estándar común a todo el equipo de desarrollo que facilita la comunicación.
- Capacidades de ingeniería directa e inversa.
- Licencia: gratuita y comercial.
- Permite el diseño de diagramas de interacción, clases, modelado de base de datos etc.
- Generación de código, modelo a código, diagrama a código.
- Permite diseñar diagramas de flujo de datos.
- Tiene incluido editor de figuras.

1.7.2 Sistema Gestor de Bases de Datos (SGBD)

PostgreSQL v9.4

PostgreSQL es un sistema de gestión de bases de datos objeto-relacional, distribuido bajo licencia BSD[4] (Berkeley Software Distribution) y con su código fuente disponible libremente. Es el sistema de gestión de bases de datos de código abierto más potente del mercado y en sus últimas versiones no tiene nada que envidiarle a otras bases de datos comerciales. (12)

PostgreSQL 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. Entre las ventajas que lo convierten en el gestor seleccionado se encuentran:

1. **Estabilidad y confiabilidad legendarias:** En contraste a muchos sistemas de bases de datos comerciales, es extremadamente común que

[4] La **licencia BSD** es la licencia de software otorgada principalmente para los sistemas BSD (*Berkeley Software Distribution*). Es una licencia de software libre que permite el uso del código fuente en software no libre.

compañías reporten que PostgreSQL nunca ha presentado caídas en varios años de operación de alta actividad.

2. **Extensible:** El código fuente está disponible para todos sin costo. Si su equipo necesita extender o personalizar PostgreSQL de alguna manera, pueden hacerlo con un mínimo esfuerzo, sin costos adicionales. Esto es complementado por la comunidad de profesionales y entusiastas de PostgreSQL alrededor del mundo que también extienden PostgreSQL todos los días.
3. **Multiplataforma:** PostgreSQL está disponible en casi cualquier sistema Unix (34 plataformas en la última versión estable), y en Windows.
4. **Diseñado para ambientes de alto volumen:** PostgreSQL usa una estrategia de almacenamiento de filas llamada MVCC[5] para conseguir una mejor respuesta en ambientes de grandes volúmenes. Los principales proveedores de sistemas de bases de datos comerciales usan también esta tecnología, por las mismas razones.

Además de las características anteriores esta herramienta es seleccionada, ya que cumple con las políticas de utilización de software libre en la UCI y es muy utilizada en los proyectos productivos de la universidad debido a las ventajas anteriormente planteadas.

PgAdmin v1.16.1

Es una herramienta de código abierto que utilizará para la administración de bases de datos PostgreSQL y derivados, incluye:

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

PgAdmin se diseña para responder a las necesidades de la mayoría de los usuarios: desde escribir simples consultas SQL hasta desarrollar bases de datos

[5] **Control de concurrencia mediante versiones múltiples** (*Multiversion concurrency control* o MVCC): es un método para control de acceso generalmente usado por SGBDs para proporcionar acceso concurrente a los datos, y en lenguajes de programación para implementar concurrencia.

complejas. La interfaz gráfica soporta todas las características de PostgreSQL y hace simple la administración. Está disponible en más de una docena de lenguajes y para varios sistemas operativos, incluyendo Microsoft Windows, Linux, FreeBSD, Mac OSX y Solaris. (14)

1.7.3 Herramienta para el proceso ETL

Pentaho Data Integration (Kettle) v4.1.0

Esta herramienta reúne un conjunto de componentes que permiten modelar y ejecutar transformaciones sobre flujos de datos. Es una de las herramientas ETL de código abierto más antigua, cuenta con una gran comunidad de usuarios y su interfaz gráfica permite un aumento de la productividad. Puede funcionar sobre varias plataformas a través de un sistema que soporte un entorno de ejecución de Java. Incluye procesamiento optimizado de los ficheros planos.

Ofrece soporte para metadatos e incluye operaciones de transformación, así como funciones que posibilitan operar con los campos en el flujo de datos, renombrando, calculando campos en función de otros, correlacionando valores y realizando búsquedas auxiliares en bases de datos.

Brinda la posibilidad de copiar y leer del mismo fichero en paralelo, permitiendo maximizar la capacidad de entrada/salida en el entorno ETL. Su rendimiento se puede ver afectado cuando se realizan operaciones de join (unión) con numerosos volúmenes de datos, pues maneja pequeñas cantidades de información en el flujo. Permite ejecutar código JavaScript dentro de las transformaciones e incorpora un evaluador de expresiones regulares. (15)

1.7.4 Herramientas para la IN.

Pentaho BI-Server v3.5

Con esta herramienta se suministra soporte e infraestructura para crear soluciones de inteligencia de negocio. Proporciona servicios básicos además de incluir autenticación, registro, auditoría y servicios web. Incorpora un motor de solución que integra reportes, análisis, tableros de comandos y componentes de minería de datos. Funciona como un sistema basado en

administración web de informes, el servidor de integración de aplicaciones y un motor de flujo de trabajo ligero (secuencias de acción). Además, está diseñada para integrarse fácilmente en cualquier proceso de negocio. Permite que puedan ejecutarse los informes y aplicaciones, se puede usar como base para construir un sistema propio de IN. (15)

Pentaho Schema Workbench v3.2

Es un entorno visual para el desarrollo y prueba de cubos OLAP Mondrian. Provee un mecanismo para buscar datos con rapidez y tiempo de respuesta uniforme independientemente de la cantidad de datos en el cubo o la complejidad del procedimiento de búsqueda. Permite la ejecución de consultas MDX contra el esquema y la base de datos y la navegación por la base de datos subyacente. (15)

Mondrian OLAP Server

Es un servidor OLAP de código abierto que gestiona la comunicación entre la aplicación OLAP y la base de datos. Permite crear cubos de información para análisis multidimensional. Este proporciona la conexión a la base de datos y ejecuta las sentencias SQL. (15)

Conclusiones parciales

En este capítulo se realizó un estudio teórico de los conceptos y herramientas para realizar un MD lo que facilitó la comprensión de los AD. Después de efectuar el análisis del modelo multidimensional fue seleccionado de los modos de almacenamiento OLAP el ROLAP por su compatibilidad con la herramienta PostgreSQL utilizada por el departamento, con la que se va a administrar el MD. Se escogió el enfoque de abajo hacia arriba (bottom-up) propuesto por Kimball, con una arquitectura de datos de tres capas. A pesar de que el MD será poblado desde un único origen ella permite recuperarse de un error sin tener que reiniciar un proceso desde el inicio. Además con su aplicación se podrán obtener los datos desde el origen de una forma mucho más rápida y se considera una arquitectura muy completa. El proceso de desarrollo estará guiado por la metodología Hefesto, con ella se realiza un análisis completo que permite

abarcar los procesos principales de la organización, examinando e interpretando de forma óptima las necesidades de información del negocio. Los resultados y objetivos esperados se distinguen fácilmente siendo los mismos sencillos de comprobar.

CAPÍTULO 2: Análisis y diseño del MD

2.1 Introducción

En este capítulo se llevará a cabo el análisis y el diseño del MD guiándose por las tres primeras fases de la metodología seleccionada. En cada una de ellas se explicará el procedimiento y se evidenciarán los resultados mediante los artefactos resultantes.

2.2 Análisis de requerimientos

Según el Ing. Bernabeu Ricardo Dario, creador de Hefesto: *el análisis de los requerimientos de los diferentes usuarios, es el punto de partida de esta metodología, ya que ellos son los que deben, en cierto modo, guiar la investigación hacia un desarrollo que refleje claramente lo que se espera del depósito de datos, en relación a sus funciones y cualidades.*

El objetivo principal de esta fase, es la de obtener e identificar las necesidades de información clave de alto nivel, que es esencial para llevar a cabo las metas y estrategias de la empresa, y que facilitará una eficaz y eficiente toma de decisiones. (9)

2.2.1 Identificar preguntas

El primer paso comienza con el acopio de las necesidades de información, el cual puede llevarse a cabo a través de muy variadas y diferentes técnicas, cada una de las cuales poseen características inherentes y específicas, como por ejemplo entrevistas, cuestionarios y observaciones.

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

Se realizó una entrevista al cliente donde plantea que necesita que el MD permita realizar el análisis estadístico combinando campos y que genere determinados reportes sin necesidad de recurrir a las herramientas de oficina como Excel, acción que actualmente no es posible en el sistema.

De la entrevista se obtienen los siguientes resultados:

1. Se desea conocer el total de noticias publicadas en un período de tiempo determinado.
2. Se desea conocer el total de noticias tendenciosas publicadas en un período de tiempo determinado.
3. Se desea conocer el total de noticias negativas publicadas en un período de tiempo determinado.
4. Se desea conocer el total de noticias positivas publicadas en un período de tiempo determinado.
5. Se desea conocer el total de noticias equilibradas publicadas en un período de tiempo determinado.
6. Se desea conocer el total de noticias sin postura publicadas en un período de tiempo determinado.
7. Se desea conocer el total de noticias publicadas por un autor en un período de tiempo determinado.
8. Se desea conocer el total de noticias publicadas en una fuente en un período de tiempo determinado.
9. Se desea conocer el total de noticias publicadas en un idioma en un período de tiempo determinado.
10. Se desea conocer el total de noticias subidas por un perfil en un período de tiempo determinado.
11. Se desea conocer el total de noticias publicadas sobre una temática en un período de tiempo determinado.
12. Se desea conocer el total de noticias publicadas de un género en un período de tiempo determinado.
13. Se desea conocer el total de noticias publicadas de una postura ideológica dada en un período de tiempo determinado.
14. Se desea conocer el total de noticias publicadas de un descriptor en un período de tiempo determinado.
15. Se desea conocer la cantidad de noticias tendenciosas publicadas por un

autor en una fuente en un período de tiempo determinado.

16. Se desea conocer la cantidad de noticias positivas publicadas por un autor en una fuente en un período de tiempo determinado.
17. Se desea conocer la cantidad de noticias negativas publicadas por un autor en una fuente en un período de tiempo determinado.
18. Se desea conocer la cantidad de noticias publicadas por un autor en una fuente en un período de tiempo determinado.
19. Se desea conocer la cantidad de noticias tendenciosas publicadas en una fuente dado un país en un período de tiempo determinado.
20. Se desea conocer la cantidad de noticias positivas publicadas en una fuente dado un país en un período de tiempo determinado.
21. Se desea conocer la cantidad de noticias negativas publicadas en una fuente dado un tipo de fuente en un período de tiempo determinado.
22. Se desea conocer la cantidad de noticias tendenciosas publicadas dada una temática y un autor en una fuente en un período de tiempo determinado.
23. Se desea conocer la cantidad de noticias positivas publicadas dada una temática y un autor en una fuente en un período de tiempo determinado.
24. Se desea conocer la cantidad de noticias negativas publicadas dada una temática y un autor en una fuente en un período de tiempo determinado.

2.2.2 Indicadores y perspectivas de análisis

Una vez que se han establecido las preguntas de negocio, se debe proceder a su descomposición para descubrir los indicadores que se utilizarán y las perspectivas de análisis que intervendrán. Para ello, se debe tener en cuenta que los indicadores, para que sean realmente efectivos son, en general, valores numéricos y representan lo que se desea analizar concretamente.

En cambio, las perspectivas se refieren a los objetos mediante los cuales se quiere examinar los indicadores, con el fin de responder a las preguntas planteadas.

A partir de las preguntas realizadas al cliente en la entrevista se identificaron los indicadores y perspectivas a analizar:

1. Total de noticias publicadas en un período de tiempo determinado.
2. Cantidad de noticias tendenciosas publicadas en un período de tiempo determinado.
3. Cantidad de noticias negativas publicadas en un período de tiempo determinado.
4. Cantidad de noticias positivas publicadas en un período de tiempo determinado.
5. Cantidad de noticias equilibradas publicadas en un período de tiempo determinado.
6. Cantidad de noticias objetivas publicadas en un período de tiempo determinado.
7. Cantidad de noticias sin postura publicadas en un período de tiempo determinado.
8. Total de noticias publicadas por un autor en un período de tiempo determinado.
9. Total de noticias publicadas en una fuente en un período de tiempo determinado.
10. Total de noticias publicadas en un idioma en un período de tiempo determinado.
11. Total de noticias subidas por un perfil en un período de tiempo determinado.
12. Total de noticias publicadas por una temática en un período de tiempo determinado.
13. Total de noticias publicadas de un género en un período de tiempo determinado.
14. Total de noticias publicadas de una postura ideológica en un período de tiempo determinado.

15. Total de noticias publicadas de un descriptor en un período de tiempo determinado.
16. La cantidad de noticias tendenciosas publicadas por un autor en una fuente en un período de tiempo determinado.
17. La cantidad de noticias positivas publicadas por un autor en una fuente en un período de tiempo determinado.
18. La cantidad de noticias negativas publicadas por un autor en una fuente en un período de tiempo determinado.
19. La cantidad de noticias tendenciosas publicadas en una fuente dado un país en un período de tiempo determinado.
20. La cantidad de noticias positivas publicadas en una fuente dado una postura ideológica en un período de tiempo determinado.
21. La cantidad de noticias negativas publicadas en una fuente dado un tipo de fuente en un período de tiempo determinado.
22. La cantidad de noticias tendenciosas publicadas dada una temática y un autor en una fuente en un período de tiempo determinado.
23. La cantidad de noticias positivas publicadas dada una temática y un autor en una fuente en un período de tiempo determinado.
24. La cantidad de noticias negativas publicadas dada una temática y un autor en una fuente en un período de tiempo determinado.

2.2.3 Modelo conceptual

Después de identificar los indicadores y las perspectivas se pasa a elaborar el modelo conceptual de MD de DOWAI (Figura 7) donde se evidencian los indicadores a la derecha y las perspectivas a la izquierda, relacionados por las noticias asociadas, proporcionando una idea precisa y clara del alcance del almacén.

Modelo de datos conceptual: El modelo conceptual captura la información fundamental acerca de las entidades del dominio del problema y sus

relaciones. Este modelo es más cercano al espacio del problema que al espacio de la solución.

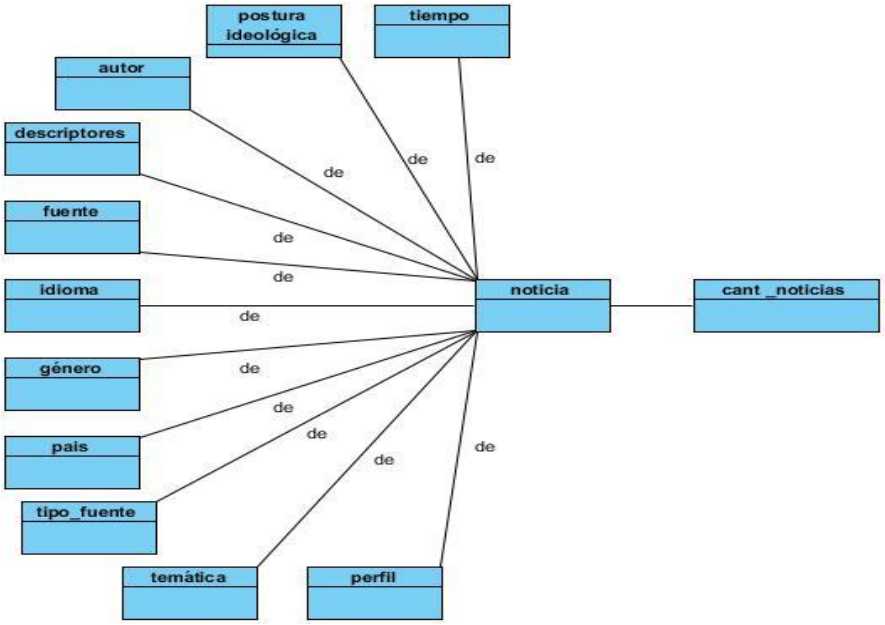


Figura 7: Modelo conceptual DOWAI

2.3 Análisis de los OLTP

Seguidamente, 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.

2.3.1 Conformar los indicadores

En este paso se explica cómo se calcularán los indicadores, definiendo como estándar de codificación las palabras separadas por guion bajo:

Indicadores

A continuación se describe cómo serán calculados cada uno de los indicadores con el objetivo de que todos los elementos del modelo conceptual estén correspondidos en los OLTP.

Indicadores	Explicación
Total de noticias **cant_noticias**	Se calcula realizando un contador de las noticias almacenadas en las perspectivas seleccionadas para el análisis.

2.3.2 Establecer correspondencias

El objetivo de este paso, es el de examinar los OLTP disponibles que contengan la información requerida, así como también sus características, para poder identificar las correspondencias entre el modelo conceptual y las fuentes de datos. Todos los elementos del modelo conceptual deben estar correspondidos en los OLTP.

Las relaciones identificadas fueron las siguientes:

- La tabla ´autor´ se relaciona con la perspectiva dimensión autor.
- La tabla ´temática´ se relaciona con la perspectiva dimensión temática.
- La tabla ´tipo_fuente´ se relaciona con la perspectiva dimensión tipo fuente.
- La tabla ´país´ se relaciona con la perspectiva dimensión país.
- La tabla ´postura_ideológica´ se relaciona con la perspectiva dimensión postura ideológica.
- La tabla ´fuente´ se relaciona con la perspectiva dimensión fuente.
- La tabla ´género´ se relaciona con la perspectiva dimensión género.
- La tabla ´descriptor´ se relaciona con la perspectiva dimensión descriptor.
- La tabla ´perfil´ se relaciona con la perspectiva dimensión perfil.

2.3.3 Nivel de granularidad

Una vez que se han establecido las relaciones con los OLTP, se deben seleccionar los campos que contendrá cada perspectiva, ya que será a través de estos por los que se examinarán y filtrarán los indicadores. Para ello, basándose

en las correspondencias establecidas en el paso anterior, se debe presentar a los usuarios los datos de análisis disponibles para cada perspectiva.

Luego de exponer frente a los usuarios los datos existentes, explicando su significado, valores posibles y características, estos deben decidir cuáles son los que consideran relevantes para consultar los indicadores y cuáles no.

Con respecto a la perspectiva “Tiempo”, es muy importante definir el ámbito mediante el cual se agruparán o sumarán los datos. Sus campos posibles pueden ser: día de la semana, quincena, mes, trimestres, semestre, año, etc.

Al momento de seleccionar los campos que integrarán cada perspectiva, debe prestarse mucha atención, ya que esta acción determinará la granularidad de la información encontrada en el MD.

A continuación se especifican los campos seleccionados por tabla para poblar las perspectivas.

Tabla	Campos	Perspectiva
autor	Id, nombre, apellidos	Dimensión autor
temática	Id, nombre	Dimensión temática
tipo_fuente	Id, nombre	Dimensión tipo fuente
país	Id, nombre	Dimensión país
postura_ideológica	Id, nombre	Dimensión postura ideológica
fuente	Id, nombre	Dimensión fuente
género	Id, nombre	Dimensión género
descriptor	Id, nombre	Dimensión descriptor
perfil	Id	Dimensión perfil

2.3.4 Modelo conceptual ampliado

En este paso, y con el fin de graficar los resultados obtenidos en los pasos anteriores, se ampliará el modelo conceptual, colocando bajo cada perspectiva los campos seleccionados y bajo cada indicador su respectiva fórmula de cálculo.

Luego de haberse recolectado toda la información, el cliente presenta los campos principales y de interés de cada perspectiva seleccionada de la base de datos de DOWAI que ayudaran a consultar los indicadores, los resultados se exponen en la (Figura 8) que representa el modelo conceptual con cada uno de los campos o atributos elegidos para cada perspectiva:

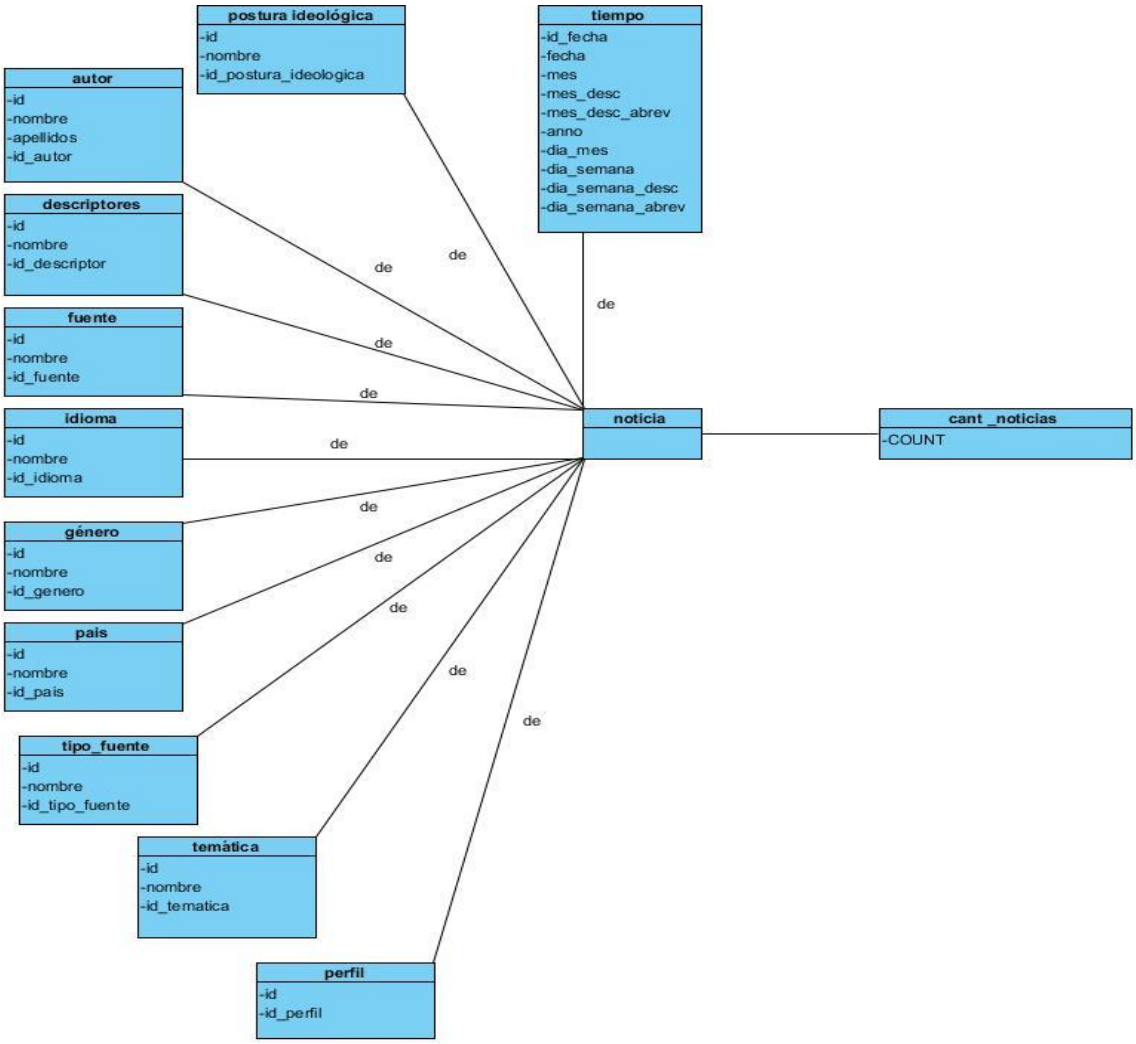


Figura 8: Modelo conceptual ampliado DOWAI.

2.4 Modelo lógico del MD

A continuación, se confeccionará el modelo lógico de la estructura del DW, 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.

2.4.1 Tipo de modelo lógico del MD

Para este paso la metodología plantea que se debe seleccionar cuál será el tipo de esquema que se utilizará para contener la estructura del depósito de datos, que se adapte mejor a los requerimientos y necesidades de los usuarios. Es muy importante definir objetivamente si se empleará un esquema en estrella, constelación o copo de nieve, ya que esta decisión afectará considerablemente la elaboración del modelo lógico.

El esquema seleccionado para llevar a cabo la solución de la investigación es el de estrella (Figura 9) ya que a partir de los requerimientos actuales del cliente se puede dar solución a estos con una única tabla de hechos. A medida que el cliente decida analizar más procesos pueden ir surgiendo más tablas de hecho y el tipo de esquema se transformaría ajustándose a las nuevas necesidades basándose en la capacidad de adaptación que poseen los AD.

2.4.2 Tablas de dimensiones

Se define como estándar de codificación para las dimensiones la estructura ¨dim_nombre¨, para la tabla hecho ¨hecho_nombre¨ y para los campos se separa las palabras que se vayan a incluir con guión bajo. Se utiliza el patrón llaves subrogadas para el tratamiento de los identificadores en la dimensiones. Una llave subrogada es un identificador único que se asigna a cada registro de una tabla de dimensión. Esta clave, generalmente, no tiene ningún sentido específico de negocio. Son siempre de tipo numérico. Preferiblemente, un entero autoincremental.

Los campos ¨id_nombre_perspectiva¨ almacenan las llaves primarias de las tablas origen asociadas a cada perspectiva y el campo ¨id¨ es la llave primaria de la dimensión destino, generada de manera secuencial.

Dimensiones	Campos
dim_autor	Id, nombre, apellidos, id_autor
dim_tematica	Id, nombre, id_tematica
dim_tipo_fuente	Id, nombre, id_tipo_fuente
dim_pais	Id, nombre, id_pais
dim_postura_ideologica	Id, nombre, id_postura_ideologica
dim_fuente	Id, nombre, id_fuente
dim_genero	Id, nombre, id_genero
dim_descriptor	Id, nombre, id_descriptor
dim_perfil	Id, id_perfil
dim_tiempo	id_fecha, fecha, dia_semana, dia_semana_desc, dia_semana_abrev, dia_mes, mes, mes_desc, mes_desc_abrev, anno

En la siguiente tabla se define la estructura de los campos y la explicación correspondiente a cada uno pertenecientes a la dimensión tiempo:

Campos	Explicación
id_fecha	Llave primaria generada automáticamente que se añaden valores a la tabla
Fecha	Fechas ejemplo (2012-11-21)
Mes	Mes numérico ejemplo (1, 2,…,12)

mes_desc	Nombre del mes ejemplo (Enero, Febrero,…, Diciembre)
mes_desc_abrev	Nombre del mes abreviado ejemplo (ENE, FEB)
dia_mes	Día numérico del mes que aparece en la fecha ejemplo (1, 2,…,31)
dia_semana	Día de la semana numérico ejemplo (1, 2, 3,…7)
dia_semana_desc	Nombre del día de la semana ejemplo (Lunes,…, Domingo)
dia_semana_abrev	Nombre del día de la semana abreviado ejemplo (LUN, MAR)
Anno	Año que aparece en la fecha ejemplo (2012, 2013)

2.4.3 Tablas de hechos

El campo ¨id_dimensiones_asociadas¨ representa a los id independientes de cada dimensión relacionada a la tabla hecho. La llave primaria de la tabla hecho la constituyen la unión de todas estas llaves y ¨cant_noticias¨ es la constante donde se va a almacenar el valor resultante de los requisitos.

Hechos	Indicadores y llaves
Hecho_cantidad_noticias	cant_noticias, id_dimensiones_asociadas

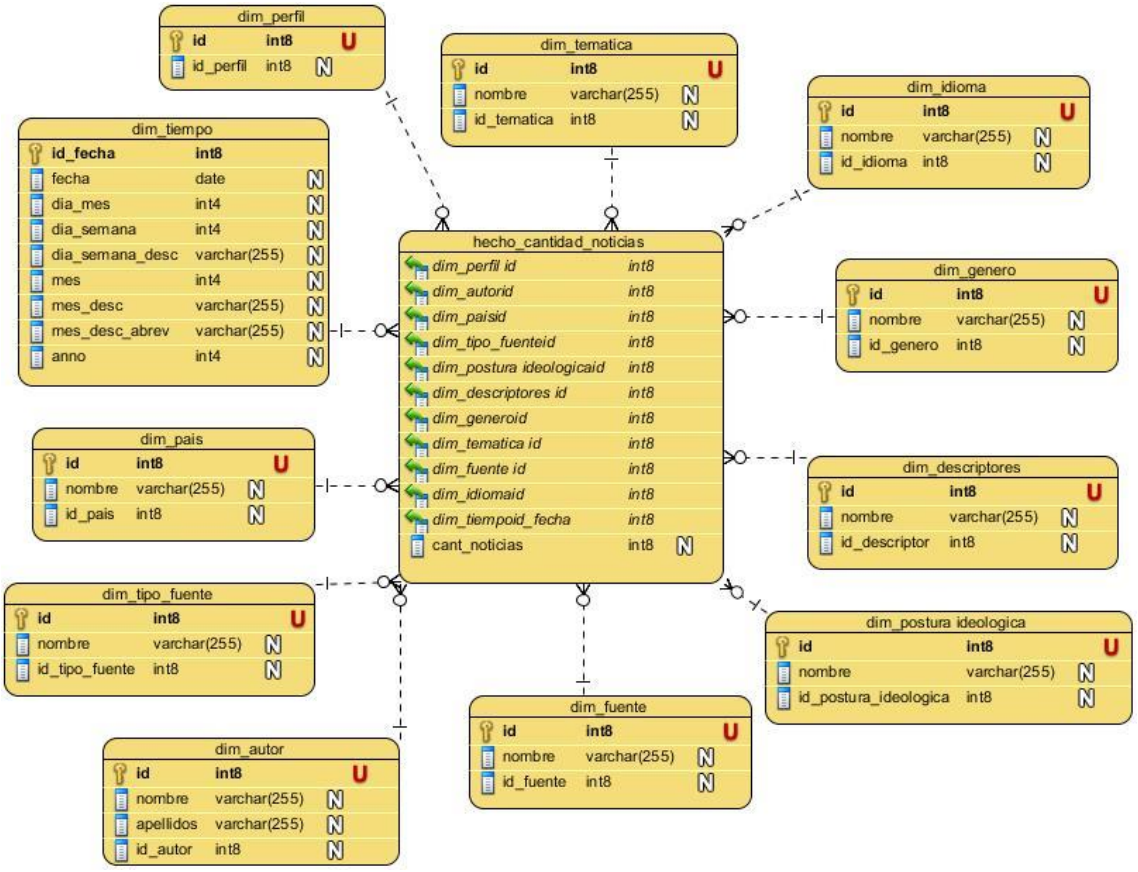


Figura 9: Esquema de estrella DOWAI.

Conclusiones parciales

En el presente capítulo se llevó a cabo el proceso de análisis y diseño del MD que va a dar solución a las interrogantes actuales del cliente. Para la realización del MD se utilizó como guía el diseño de la metodología de Hefesto describiendo en cada fase el cómo y porqué de cada uno de los pasos realizados. Dando cumplimiento a las tres primeras fases (de un total de cuatro) de dicha metodología.

En la primera fase se realizó un análisis de los requerimientos del cliente y se identificaron las preguntas, de las cuales se elaboran los indicadores y perspectivas para la creación del modelo conceptual del MD donde se evidencia la relación que existe entre estos. Esta etapa permite adentrarse en las necesidades que posee el cliente y poder aterrizarlas para su mejor comprensión.

En la segunda fase se llevó a cabo el análisis de los OLTP con que cuenta el cliente y así dar respuesta a los problemas de la primera fase. En este espacio

son detectados inconvenientes en cuanto al análisis estadístico ya que el cliente no cuenta con datos claves en la base de datos que faciliten la realización del mismo. No obstante se trabaja con los que poseen conformando así los indicadores y las correspondencias que son factibles hasta el momento, seleccionando el nivel de granularidad, lo que da lugar a la confección del modelo conceptual ampliado del MD, que contiene los campos de las perspectivas y las formas de calcular los indicadores.

En la tercera fase se diseñó el modelo lógico del MD. En este se definió el tipo de modelo seleccionado (estrella), las dimensiones y hechos que lo componen incluyendo los campos que poseen así como las relaciones que van a servir en el análisis.

CAPÍTULO 3: Implementación del MD

3.1 Introducción

Para dar cumplimiento al actual capítulo se realizará inicialmente la cuarta fase de la metodología Hefesto con los procesos ETL. Se modelará el cubo OLAP para permitir un mejor acceso a la información y dar paso a la visualización con el Pentaho BI server. Se finalizará con la realización de pruebas al MD para validar que se da solución a las exigencias del cliente.

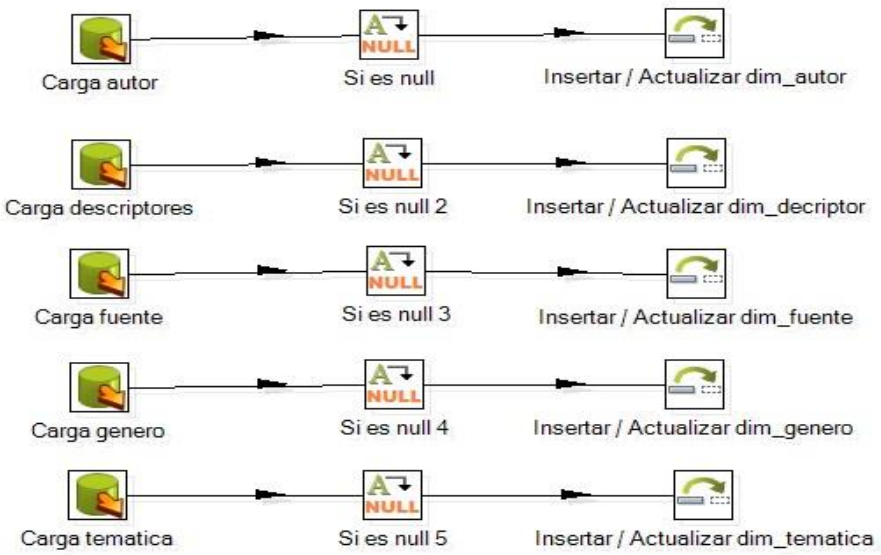
3.2 Integración de los datos (ETL)

Después de creado el modelo lógico del AD se pasa a poblarlo con datos, utilizando técnicas de limpieza, calidad de datos y procesos ETL. Luego se definirán las reglas y políticas para su respectiva actualización, así como también los procesos que la llevarán a cabo.

3.2.1 Carga inicial

En este paso se realiza la carga de los datos almacenados en la fuente de datos para ser guardados en las dimensiones correspondientes. Para esto se llevó a cabo una serie de transformaciones y trabajos con la herramienta Pentaho Data Integration.

3.2.2 Carga de las dimensiones.



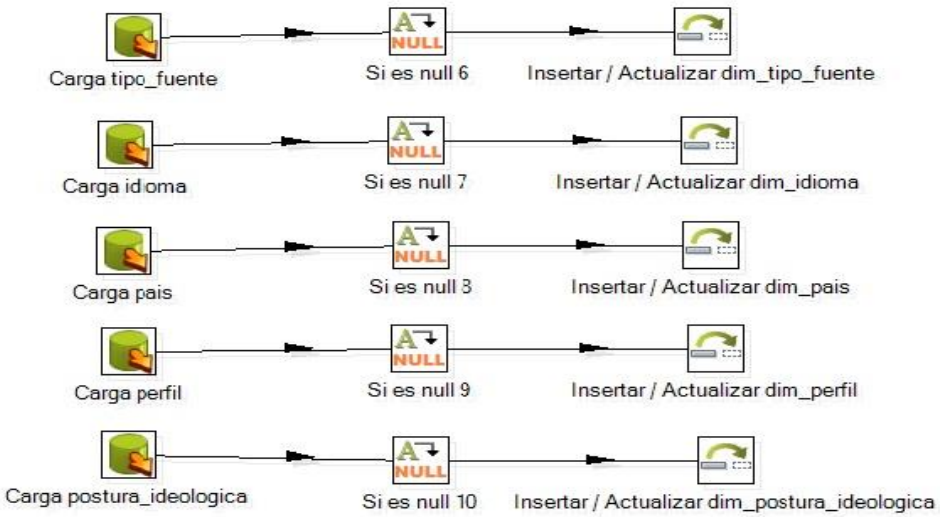


Figura 10: Proceso ETL para la carga de las dimensiones del MD

En la imagen anterior se muestran tres pasos para el llenado de las dimensiones:

- **Obtención de los datos de las tablas:** Para la carga de los datos, las consultas hacen referencias a las tablas donde se encuentra almacenada la última actualización de la información a solicitar.
- **Validación de los campos nulos existentes en las tablas:** Aquí se llenan los campos nulos o vacíos de las tablas con los valores ¨Desconocido¨ para los campos nominales y ¨0¨ para los numéricos.
- **Inserción y actualización de los campos de las dimensiones:** Se procede a insertar los datos obtenidos, en las dimensiones correspondientes.

3.2.2 Carga de la dimensión tiempo

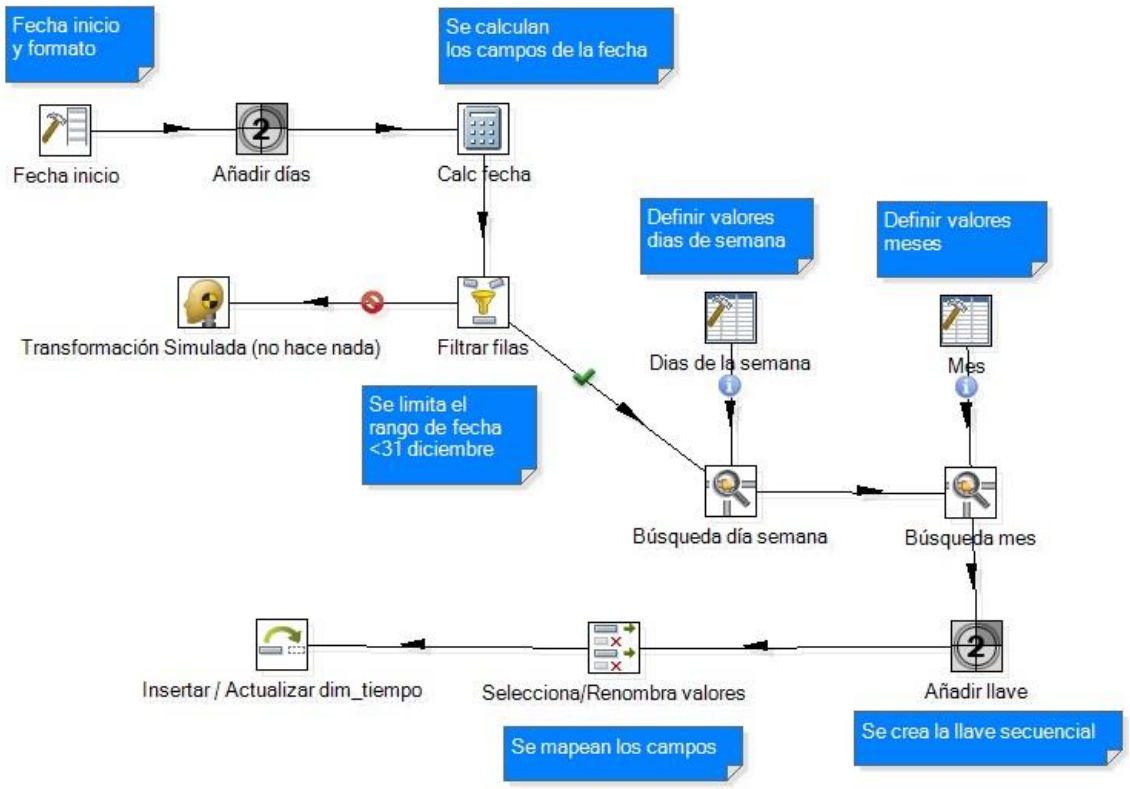


Figura 11: Proceso ETL para el llenado de la dimensión tiempo.

En esta transformación se determina la fecha inicial para a partir de ella generar todas las fechas que se quieren almacenar. Las fechas almacenadas comienzan desde el ´2012-11-20´ hasta el ´2020-12-30´, la fecha inicial se define por la existente en la fuente OLTP y la final fue definida por el cliente lo cual no quiere decir que ahí culmine ya que posteriormente se pueden añadir más fechas.

3.2.3 Carga de la tabla de hechos.

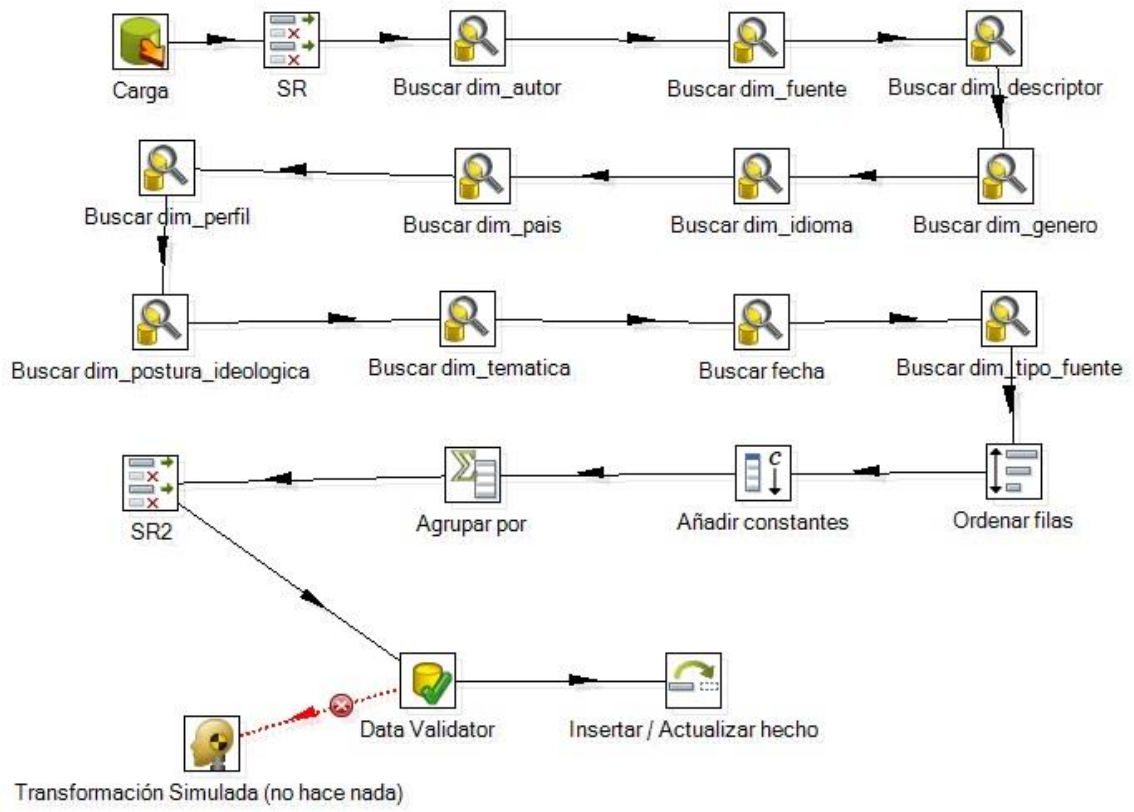


Figura 12: Proceso ETL para la carga de la tabla hecho del MD

Esta transformación consiste en extraer los identificadores de las tablas en la fuente relacionándolas a la tabla noticias. Se comparan cada uno de los id con los almacenados en las dimensiones correspondientes y se obtiene la llave primaria asociada a la fila donde se encuentra este. Se ordenan los datos de manera ascendente, se crea la constante cant_noticias donde se va a almacenar los resultados de los requisitos del cliente y se agrupan. Antes de insertar los datos en la tabla destino se verifica que no existan valores nulos, si existen sigue al paso ¨Transformación simulada¨ donde no ocurre nada, sino el flujo culmina satisfactoriamente.

3.2.4 Automatización del proceso de carga

Después de culminar la carga inicial de las tablas en el MD se procede a la automatización del proceso de carga diaria. Para esto se crea el trabajo ¨Carga diaria¨ el cual se ejecuta a las 12:00 am definido por el cliente.

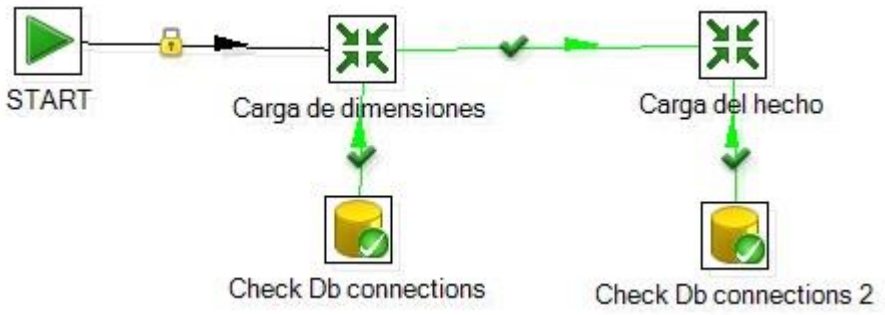


Figura 13: Trabajo para automatizar la carga diaria del MD

3.3 Creación del cubo multidimensional

Para realizar el análisis sobre los datos almacenados en el MD se hace necesario la creación del cubo OLAP donde se va a definir la información que va a ser utilizada en las vistas de análisis en el Pentaho BI server.

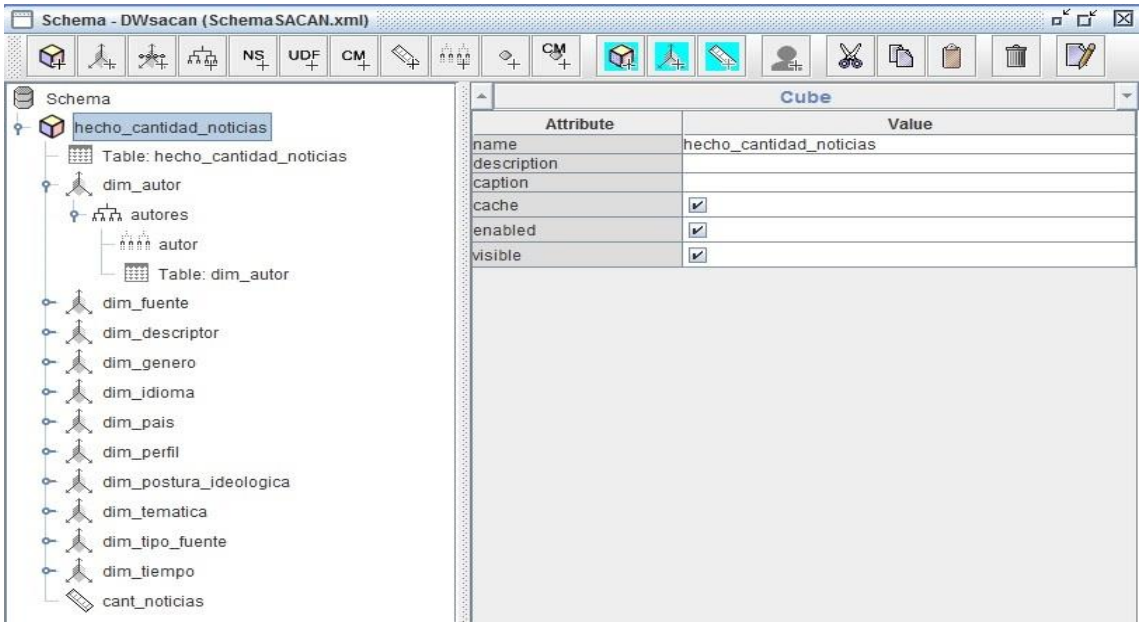


Figura 14: Cubo OLAP del MD

A partir del cubo creado se puede visualizar toda la información del MD desde todos los niveles de detalle que se definió en el diseño sin necesidad de realizar consultas SQL manuales. El cubo permite realizar el análisis de las cantidades de noticias referentes a cada una de las dimensiones por separado y en relación con las que se desee.

3.4 Visualización de los datos

Para la visualización de los datos se utilizó el Pentaho BI server v3.5 con la vista de análisis. Para esto se publica el cubo generado por el Pentaho Schema Workbench en un fichero .xml.

			Medidas
autores	fuente	tiempo	cant_noticias
All dim_autor.autoress	All dim_fuente.fuentes	All dim_tiempo.tiempos	18
ANDRÉS	All dim_fuente.fuentes	All dim_tiempo.tiempos	2
		2013.0	2
		3.0	2
		12.0	2
	El Nacional	All dim_tiempo.tiempos	2
		2013.0	2
		3.0	2
		12.0	2
César	All dim_fuente.fuentes	All dim_tiempo.tiempos	2
		2013.0	2
		3.0	2
		12.0	2
Gabriela	All dim_fuente.fuentes	All dim_tiempo.tiempos	1
Grabiela	All dim_fuente.fuentes	All dim_tiempo.tiempos	3
Juan	All dim_fuente.fuentes	All dim_tiempo.tiempos	1

Figura 15: Tabla de análisis con las perspectivas autor, fuente y tiempo

Esta tabla permite visualizar la cantidad de noticias por las perspectivas autor, fuente y tiempo, al igual con el resto de las demás perspectivas solo hay que seleccionar cuales se desean para el análisis. A continuación una gráfica que representa la misma información generada también por la herramienta.

Figura 16: Gráfica de análisis con las perspectivas autor, fuente y tiempo

3.4.1 Roles y usuarios

Para el acceso a la visualización de la información se crearon los roles ¨Administrador¨ y ¨Autenticado¨ con el objetivo de limitar las funcionalidades que pueden realizar los usuarios del sistema. También se definieron los usuarios ¨administrador¨ y ¨analista¨ y se le asignaron los roles anteriormente planteados respectivamente.

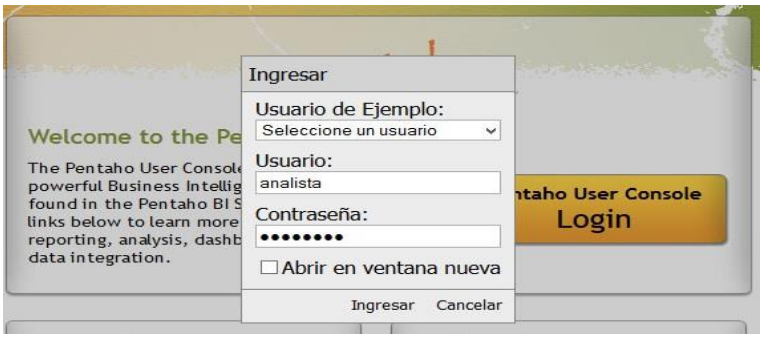


Figura 17: Interfaz de autenticación definida por Pentaho BI server

3.5 Validación del mercado de datos

Para la validación del MD se elaboraron un conjunto de pruebas, con las que se garantiza el correcto funcionamiento del mismo a partir de comprobar que se han

cumplido con todas las metas y requisitos planteados por el cliente al inicio del ciclo de desarrollo.

3.5.1 Listas de chequeo

Las listas de chequeo son formatos creados para realizar actividades repetitivas, controlar el cumplimiento de una lista de requisitos o recolectar datos ordenadamente y de forma sistemática**.** Se usan para hacer comprobaciones regulares de actividades o productos asegurándose de que el trabajador o inspector no se olvida de nada importante.

La lista de chequeo elaborada para la solución fue la siguiente:

Indicadores	Evaluación	Comentarios
Documento		
¿Posee una correcta ortografía?		
¿Se entiende claramente lo que se ha especificado en el documento?		
MD		
¿Se seleccionó la metodología adecuada?		
¿Se concluyó cada una de las fases correctamente?		
¿Se generaron los artefactos definidos?		
Cliente		
¿Se dio solución a todos los requerimientos?		
¿El cliente quedó satisfecho?		

3.5.2 Pruebas de rendimiento

Para verificar el rendimiento del MD se le realizó una prueba con la herramienta Apache JMeter v2.12 basándose en las siguientes condiciones:

- Sistema a probar: Mercado de datos para el análisis de los medios de prensa digitales en el Departamento de Operaciones Web y Análisis de Información.
- Operación a realizar: Consulta SQL.
- Servidor de PostgreSQL en ejecución.
- Hardware: 2Gb de RAM, 1Tb de capacidad en el Disco duro y Procesador Intel Core i3.
- Cantidad de conexiones: 25 usuarios.
- Contador del bucle: 1.
- Rendimiento esperado: menos de 30 segundos.
- Consulta SQL: Ver (Anexo 9) similar a las consultas que deberá realizar el MD constantemente.

Al finalizar, la prueba arrojó los siguientes resultados donde se evidencia la ausencia de errores y un rendimiento de 25,7 segundos lo que significa que se tardó aproximadamente 1 segundo por cada consulta. También se puede observar que el sistema se tarda una media de 6 segundos para la realización de las peticiones de este tipo.

Etiqueta	# Muestras	Media	Mediana	Linea de 90%	Mín	Máx	% Error	Rendimiento	Kb/sec
Petición JDBC	25	6	6	7	5	20	0,00%	25,7/sec	4,5
Total	25	6	6	7	5	20	0,00%	25,7/sec	4,5

Figura 18: Resultado de la prueba de rendimiento con JMeter.

Para verificar el MD en una situación de estrés se aumentó el bucle a 1000 o sea se va a repetir la sentencia 1000 veces para observar la variación del rendimiento.

Etiqueta	# Muestras	Media	Mediana	Linea de 90%	Mín	Máx	% Error	Rendimiento	Kb/sec
Petición JDBC	25000	21	5	51	3	1091	0,00%	373,8/sec	64,6
Total	25000	21	5	51	3	1091	0,00%	373,8/sec	64,6

Figura 19: Resultado de la prueba de rendimiento con bucle en 1000.

Como se puede observar la media de respuesta ante las solicitudes ascendió a los 21 segundos lo cual no es una variación significativa y no se generaron errores.

3.5.3 Pruebas de estrés

Un aspecto importante para probar un MD es el cómo responde anta cargas de grandes volúmenes de datos. La herramienta Generador de datos para PostgreSQL (Data Generator for PostgreSQL) permite realizar esta acción y verificar la ocurrencia de errores. Para la prueba se realizó la carga de todas las dimensiones y del hecho estableciendo la cantidad de tuplas a insertar en 10000. El archivo de logs o registros de la prueba arrojó como resultado que se habían insertado satisfactoriamente los datos en las dimensiones.

Generating data for table public.dim_autor...

Generation started

Generation complete

10000 records generated, 0 errors occurred

Logs de la carga de dim_autor con el Generador de datos para PostgreSQL

Conclusiones parciales

En el capítulo finalizado se llevaron a cabo los procesos ETL para la carga inicial de las dimensiones y la tabla hecho del MD así como su actualización diaria. Se elaboró el cubo OLAP para una mejor comprensión y visualización de la información generando el fichero xml. Se creó el servidor para el análisis del cubo a partir del fichero anteriormente obtenido permitiendo mostrar los datos sin necesidad de realizar consultas SQL. También se elaboraron pruebas enfocadas a verificar las actividades realizadas, mediante listas de chequeo. Se probó el rendimiento del mercado con la herramienta Apache-JMeter v2.12 dando resultados alentadores en comparación a los esperados. Para validar la respuesta del MD ante las situaciones de estrés se realizó la carga de todas las tablas con la herramienta Generador de datos para PostgreSQL devolviendo resultados satisfactorios.

CONCLUSIONES

Con el objetivo de implementar un MD para el análisis de los medios de prensa digital en DOWAI se elaboraron un conjunto de metas a cumplir, las mismas consistieron en:

- El estudio del estado del arte, los conceptos y características de los AD, permitieron la selección de las herramientas y metodologías empleadas en la construcción del MD.
- Basándose en la metodología seleccionada se pudo analizar el negocio y obtener un total de veinticuatro preguntas enfocadas a las necesidades del cliente.
- Gracias a los artefactos generados en cada fase, se facilitó la verificación entre las necesidades del cliente y el diseño elaborado.
- Las herramientas seleccionadas para el proceso ETL y de IN garantizaron una población del MD satisfactoria así como el modelado y visualización de una manera sencilla, asequible y eficiente para los usuarios.
- Mediante las pruebas realizadas se pudo chequear la validez de la implementación realizada y su correspondencia según los requisitos del cliente.

Referencias bibliográficas

1. **Lefcovich, Mauricio León.** [En línea] http://manuelgross.bligoo.com/content/view/218974/La-estadistica-es-fundamental-para-la-gestion-eficiente.html.

2. **http://www.softonic.com/. [En línea] http://google-estadisticas-de-busqueda.softonic.com/aplicaciones-web.**

3. http://hipertextual.com. [En línea] http://hipertextual.com/archivo/2010/09/twirus-mas-trending-topics-de-los-que-nos-ofrece-twitter/.

4. Leidys García Chico. http://www.cubahora.cu. [En línea] http://www.cubahora.cu/sociedad/medirle-el-pulso-a-la-informacion.

5. Sinexxus. [En línea] 19 de enero de 2015. http://www.sinnexus.com/empresa/index.aspx.

6. Inmon, W. H. *Building the Data Warehouse.* New York : John Wiley & Sons, 2002.

7. Kimball, Ralph. *The Data Warehouse Toolkit: the complete guide to dimensional modeling.* New York : John Wiley & Sons, 2002.

8. Hernández, Yanisbel González. *METODOLOGÍA DE DESARROLLO PARA PROYECTOS DE ALMACENES DE DATOS.* La Habana : s.n., 2013.

9. Bernabeu, Ricardo Dario. *HEFESTO: Metodología propia para la Construcción de un Data Warehouse.* 2015.

10. http://definicion.de. [En línea] http://definicion.de/metodologia/.

11. http://www.cognus.cl. [En línea] http://www.cognus.cl/content/view/271452/Pentaho-Open-BI.html.

12. http://www.visual-paradigm.com/. [En línea] http://www.visual-paradigm.com/.

13. http://www.postgresql.org.es/. [En línea] http://www.postgresql.org.es/.

14. http://www.pgadmin.org/. [En línea] http://www.pgadmin.org/.

15. http://www.pentaho.com/. [En línea] http://www.pentaho.com/.

16. https://modelosbd2012t1.wordpress.com. [En línea] https://modelosbd2012t1.wordpress.com/2012/03/02/almacenes-de-datos/.

Bibliografía

1. **Lefcovich, Mauricio León.** [En línea] http://manuelgross.bligoo.com/content/view/218974/La-estadistica-es-fundamental-para-la-gestion-eficiente.html.

2. **http://www.softonic.com/. [En línea] http://google-estadisticas-de-busqueda.softonic.com/aplicaciones-web.**

3. http://hipertextual.com. [En línea] http://hipertextual.com/archivo/2010/09/twirus-mas-trending-topics-de-los-que-nos-ofrece-twitter/.

4. Leidys García Chico. http://www.cubahora.cu. [En línea] http://www.cubahora.cu/sociedad/medirle-el-pulso-a-la-informacion.

5. Sinexxus. [En línea] 19 de enero de 2015. http://www.sinnexus.com/empresa/index.aspx.

6. Inmon, W. H. ***Building the Data Warehouse.*** **New York : John Wiley & Sons, 2002.**

7. Kimball, Ralph. ***The Data Warehouse Toolkit: the complete guide to dimensional modeling.*** **New York : John Wiley & Sons, 2002.**

8. Hernández, Yanisbel González. ***METODOLOGÍA DE DESARROLLO PARA PROYECTOS DE ALMACENES DE DATOS.*** **La Habana : s.n., 2013.**

9. Bernabeu, Ricardo Dario. ***HEFESTO: Metodología propia para la Construcción de un Data Warehouse.*** **2015.**

10. http://definicion.de. [En línea] http://definicion.de/metodologia/.

11. http://www.cognus.cl. [En línea] http://www.cognus.cl/content/view/271452/Pentaho-Open-BI.html.

12. http://www.visual-paradigm.com/. [En línea] http://www.visual-paradigm.com/.

13. http://www.postgresql.org.es/. [En línea] http://www.postgresql.org.es/.

14. http://www.pgadmin.org/. [En línea] http://www.pgadmin.org/.

15. http://www.pentaho.com/. [En línea] http://www.pentaho.com/.

16. https://modelosbd2012t1.wordpress.com. [En línea] https://modelosbd2012t1.wordpress.com/2012/03/02/almacenes-de-datos/.

17. Torres Llanes, Jany y Alfonso Collado, Andy Carlos. ***Mercado de Datos para la Dirección de Colaboración Económica.*** **La Habana : s.n., 2012.**

18. Nleblas, Leonel Pérez. *Sistema de Información de Gobierno. Mercado de datos para el área de Construcción.* La habana : s.n., 2011.

19. Mora, Sergio Luján. *Diseño de almacenes de datos con UML.*

20. Mascareño Hodge, Leidys Susel y Peña Consuegra, Patricia. *Sistema de Información de Gobierno. Mercado de datos Comercio exterior.*

21. López, Ing. Yadini Pérez. *Mercado de Datos para la gestión de la información .* La Habana : s.n., 2013.

22. Gallego, J. C. *Tecnologías de la Información y de la Comunicación. Técnicas básicas.* Madrid : Editex, 2010.

23. Ferrer, Alejandro Tiana. http://www.campus-oei.org. [En línea] http://www.campus-oei.org/calidad/tiana.htm.

24. Abreu, Eddy David Amaya. *Mercado de datos Series históricas de industria .* La Habana : s.n., 2012.

25. Ms. C. Martha Denia Hernández Ramírez. www.acimed.sld.cu. [En línea] http://www.acimed.sld.cu/index.php/acimed/rt/printerFriendly/208/168.

26. http://www.ecured.cu. [En línea] http://www.ecured.cu/index.php/Bases_de_datos.

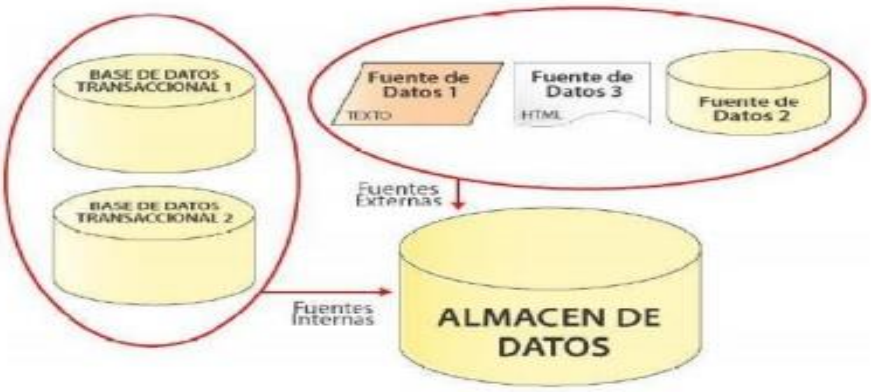
27. http://www.bnf.fr. [En línea] http://www.bnf.fr/es/colecciones_y_servicios/prensa/s.bases_de_datos_prensa.html?first_Art=non.

28. http://gravitar.biz. [En línea] http://gravitar.biz/bi/metodologia-business-intelligence/.

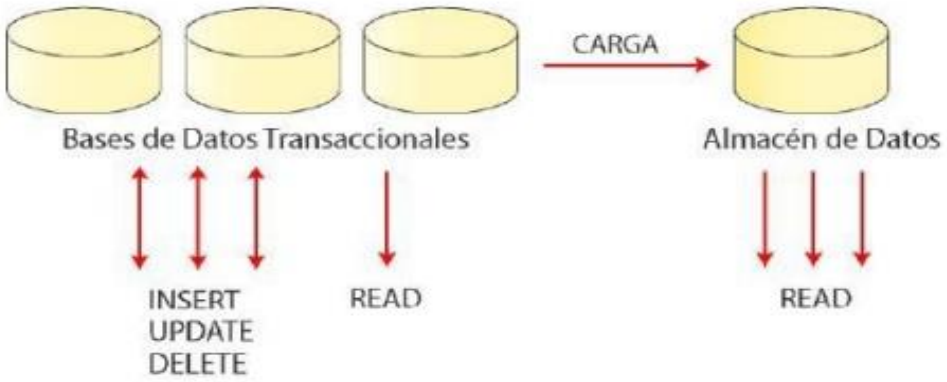
29. http://gravitar.biz. [En línea] http://gravitar.biz/bi/pentaho-ejemplo-cubo-mondrian/.

Anexos

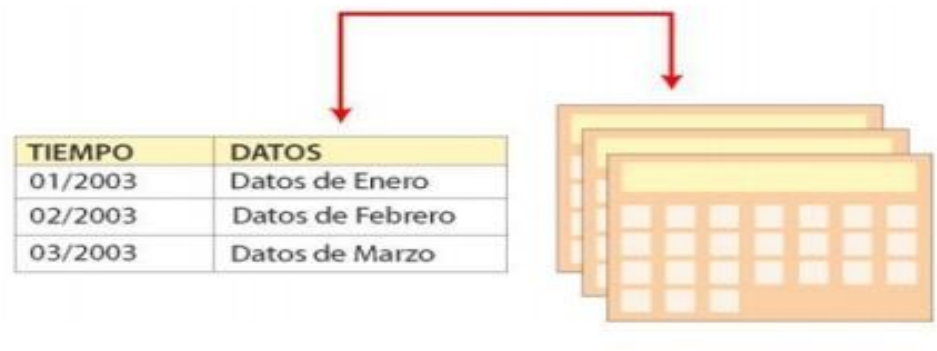
Anexo 1: Orientado a temas



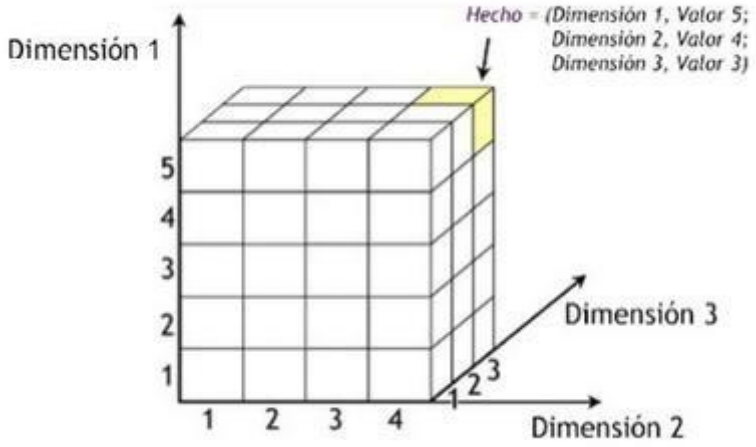
Anexo 2: Integrado (12)



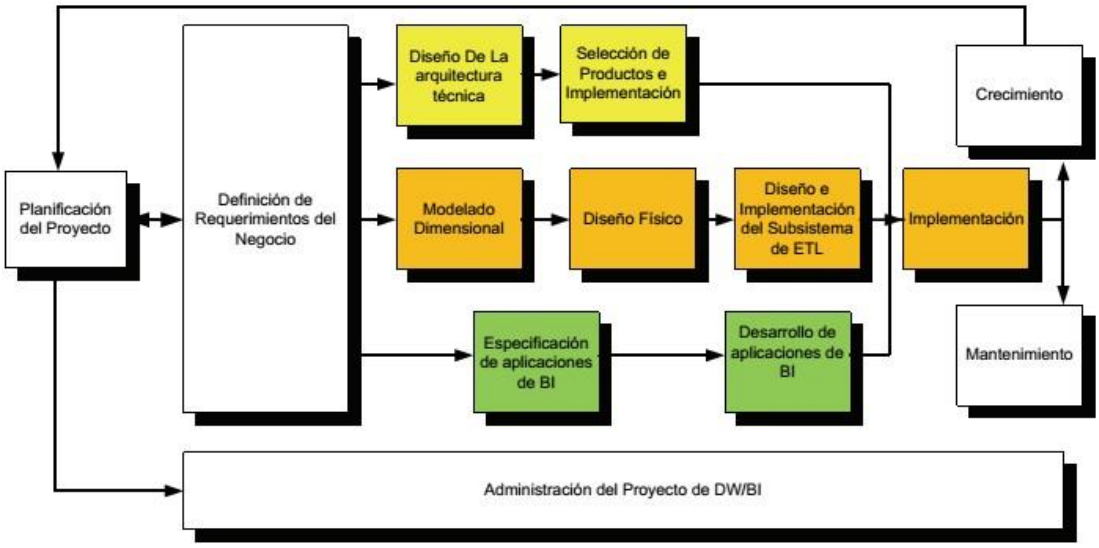
Anexo 3: No volátil (12)



Anexo 4: Datos Históricos (12)



Anexo 5: Modelo de datos multidimensional.



Anexo 6: Tareas de la metodología Kimball. (8)

SELECT "public".hecho_cantidad_noticias.dim_autorid,

```
"public".hecho_cantidad_noticias.dim_descriptorid,
"public".hecho_cantidad_noticias.dim_fuenteid,
"public".hecho_cantidad_noticias.dim_generoid,
"public".hecho_cantidad_noticias.dim_idiomaid,
"public".hecho_cantidad_noticias.dim_paisid,
"public".hecho_cantidad_noticias.dim_perfilid,
"public".hecho_cantidad_noticias.dim_postura_ideologicaid,
"public".hecho_cantidad_noticias.dim_tematicaid,
"public".hecho_cantidad_noticias.dim_tiempoid,
"public".hecho_cantidad_noticias.dim_tipo_fuenteid,
"public".hecho_cantidad_noticias.cant_noticias
FROM
"public".hecho_cantidad_noticias
INNER JOIN "public".dim_autor ON
"public".hecho_cantidad_noticias.dim_autorid = "public".dim_autor."id"
INNER JOIN "public".dim_descriptor ON
"public".hecho_cantidad_noticias.dim_descriptorid = "public".dim_descriptor."id"
INNER JOIN "public".dim_tiempo ON
"public".hecho_cantidad_noticias.dim_tiempoid = "public".dim_tiempo.id_fecha
WHERE
"public".dim_autor.id_autor = 839 AND
"public".dim_descriptor.id_decriptor = 1028 AND
"public".dim_tiempo.fecha = '2012-12-12'
GROUP BY
"public".hecho_cantidad_noticias.dim_autorid,
```

"public".hecho_cantidad_noticias.dim_descriptorid,

"public".hecho_cantidad_noticias.dim_tiempoid,

"public".hecho_cantidad_noticias.dim_fuenteid,

"public".hecho_cantidad_noticias.dim_generoid,

"public".hecho_cantidad_noticias.dim_idiomaid,

"public".hecho_cantidad_noticias.dim_paisid,

"public".hecho_cantidad_noticias.dim_perfilid,

"public".hecho_cantidad_noticias.dim_postura_ideologicaid,

"public".hecho_cantidad_noticias.dim_tematicaid,

"public".hecho_cantidad_noticias.dim_tipo_fuenteid

Anexo 7: Consulta SQL para probar el rendimiento