Facultad 2

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

Título: Vista Minable a partir del desarrollo de un Mercado de Datos para el Área de Informatización de la Biblioteca UCI.

Autores: Yoel Hernández Merlín
Antonio Sevila Hidalgo

Tutores: Ing. Evelyn Guindo Betancourt
Ing. Yadier Mesa Pérez
Co-Tutor: Dr.C. Jorge Sergio Menéndez Pérez

La Habana, junio 2015
"Año 57 de la Revolución"

Declaramos que somos los únicos autores de este trabajo y autorizamos al Centro de Informatización de Gestión Documental (CIGED) de la Universidad de las Ciencias Informáticas (UCI) a hacer uso del mismo en su beneficio.

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

Yoel Hernández Merlin

_____
Firma del autor

Antonio Sevila Hidalgo

_____
Firma del autor

Ing. Evelyn Guindo Betancourt

_____
Firma del tutor

Ing. Yadier Mesa Pérez

_____
Firma del tutor

Dr. Jorge Sergio Menéndez Pérez

_____
Firma del co-tutor

"El profesionalismo del ingeniero radica en el detalle"

Jorge Sergio Menéndez Pérez

AGRADECIMIENTOS

Antonio Sevila Hidalgo.

Agradezco a la revolución por proporcionarme todos los medios que he utilizado durante el decursar de mi vida como estudiante. Agradezco a mi familia, mi novia y su madre, por el apoyo incondicional con el que siempre he contado y el abrigo, calor y cariño que me han proporcionado.

A mi compañero de tesis, por ser más que un amigo, un hermano. Al profesor Isbel Herrera del Sol y su esposa Elianis Cepero, así como toda la familia del ajedrez por formar parte de mi vida para siempre. A los amigos que nunca olvidaré: el Piri, Alejandrito, Alejandro que es mi tío, Luisa, Martín, Adrián, Yaliana, La cosi, Karel, Douglas, Carlos el ñaña, Jessica, a mis compañeros de aula en el pre universitario, a Osvaldo, Jorge Carlos el niche, Yosvani, Abelito, Reinier Arias y muchos que por cuestiones de espacio albergaré en mi corazón con mucho cariño. A los profesores que me han guiado a lo largo de este camino que ha sido mi vida como universitario, al profesor Menéndez, al profesor Yadier –tremendo tipo- a la profesora Evelyn por aportarnos nuevo conocimiento para nuestras vidas. A todas las personas que han contribuido de una forma u otra con el desarrollo de esta tesis, muchas gracias.

Yoel Hernández Merlin.

Primeramente a los 2 pilares de mi vida, mi madre y mi difunto abuelo, por la fuerza, esfuerzo y sacrificio que siempre me inculcaron, este importante paso es dedicado especialmente a ellos.

A toda mi familia y personas que de una forma u otra fueron un apoyo para ponerme de pie y seguir echando para adelante.

A mi novia Wendy, a veces azul a veces gris, pero el apoyo siempre fue incondicional.

A mi hermano, que se pela bajito para pasar por rubio, mil gracias por esta larga travesía juntos.

Al profesor Menéndez, por significar un padre desde el día 1 para ambos en la UCI, con sus lecciones inolvidables pues muchas veces los hechos dicen más que mil palabras

A Yadier, siempre le dije esto no era papa de él, y su paso siempre fue firme.

A todas las nuevas amistades conocidas, algunos aquí: Yusbiel, Jefrey, Riki, El gordo (Osvaldo), el Yosva, Mellizo y el equipo UCI de beisbol; otros que por una razón u otra no están presentes: Yaimel, Luis Miguel, Rubio, Noel, todos de aquel inolvidable 7106.

En fin, a toda persona que influyó positivamente en mi vida universitaria, tengan en mí un amigo.

A la revolución.

RESUMEN

El descubrimiento de conocimientos en bases de datos se ha convertido en una herramienta muy potente en el mundo actual, por la necesidad de aplicar métodos que descubran la información oculta en grandes volúmenes de datos. Por esta razón, es considerada una disciplina para la búsqueda de conocimiento no trivial, a partir de los datos previamente desconocidos.

La Universidad de las Ciencias Informáticas (UCI), cuenta con un servicio de repositorio, en el cual se almacenan los resultados científicos de profesores, especialistas y estudiantes de este Centro de Enseñanza Superior. La dirección del Centro de Informatización Científico Técnica (CICT) y la Dirección de Investigación UCI carecen de argumentos basados en el conocimiento almacenado para el proceso de toma de decisiones, debido a que no se aprovecha la información almacenada en el repositorio.

En la elaboración del presente trabajo de Diploma, se lleva a cabo la creación de una vista minable a partir de un mercado de datos, que usa como fuentes: el repositorio institucional y los servicios del Protocolo Ligero de Acceso a Directorios. Ambos propósitos, tanto el proceso parcial de descubrimiento de conocimiento de bases de datos, como el mercado de datos, fueron guiados por las metodologías: “CRISP-DM” y “Metodología de Desarrollo para Proyectos de Almacenes de Datos (DATEC)” respectivamente

Palabras Clave: agrupamiento, almacén de datos, mercado de datos, minería de datos.

INDICE

INTRODUCCION

La necesidad de información nace con el surgimiento de los primeros hombres que habitaron la tierra. Dicha información era asociada a características de objetos y fenómenos, posteriormente su integración dio paso a herramientas, invenciones y otros beneficios. El conocimiento fue en ascenso, y los bienes que proveía a su portador como capacidad de adaptación y solución de problemas, también. Debido a esto comienzan a surgir razones con la intensión de almacenar el mismo.

En la actualidad, con la madurez y desarrollo del hombre como ser social, los sistemas de gestión han adquirido una importancia vital, y necesitan cada vez más de información nueva o actualizada, multiplicándose la cantidad de datos almacenados en forma exponencialmente creciente.

Con la potencia de cómputo existente, no representa un problema almacenar estos grandes cúmulos de información, pero se hace énfasis, en la necesidad de administrar la misma. Gestión que permita comprender, clasificar, analizar los datos y reportes que se pudieran generar, mediando una fuente confiable de patrones, que sirva de soporte a especialistas para la toma de decisiones.

Ante la situación global y necesaria de analizar los datos generados por los incontables sistemas de gestión de información y otros factores, surge una tecnología con tales fines, denominada “Minería de Datos”. La *minería de datos* no es más que el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (1). Es decir, su tarea fundamental es encontrar modelos más entendibles a partir de los datos y los patrones a obtener deberían ayudar a tomar decisiones más seguras en beneficio de la organización.

Cuba no está ajena a la necesidad antes descrita, tomando mayor importancia en los últimos años por los lineamientos de la política económica y social del Partido y la Revolución unido al extenso proceso de informatización de la sociedad. Donde a través de ambos se pretende impulsar, facilitar y promover el uso masivo de servicios y productos de las tecnologías de la información, comunicación, electrónica y automatización para satisfacer las expectativas de todas las esferas de la sociedad. (2)

Las universidades también forman parte de este paulatino proceso. Particularmente, la UCI en una de sus áreas, la biblioteca, cuenta con el servicio del Repositorio Institucional. El cuál es la fuente fundamental de los resultados científicos de profesores, especialistas y estudiantes del centro, donde los últimos mencionados elaboran tesis de culminación de estudios y representan el mayor número de publicaciones almacenadas, acumulándose además tesis de maestría, tesis de doctorado y otros artículos científicos.

Esta valiosa información almacenada, que aumenta cada año, actualmente no está organizada ni estructurada de una forma que permita un análisis objetivo para apoyar el proceso de toma de

decisiones por los directivos del Centro de Informatización Científico Técnica (CICT) y de la Dirección de Investigación de la UCI. Ante esta situación y para revertir la misma, se hace necesario el estudio de los datos almacenados en el repositorio, para contribuir al perfeccionamiento del proceso de toma de decisiones por los directivos de estas áreas.

Atendiendo a las ideas expuestas se identificó como **problema a resolver** el siguiente**:** *se dificulta el proceso de toma de decisiones de los directivos del Centro de Informatización Científico Técnica y la Dirección de Investigación de la UCI, por la insuficiente organización y estructuración de la información almacenada en el repositorio institucional de la universidad.*

Teniendo en cuenta el problema anteriormente referenciado, se considera como **idea a defender** que: *con la elaboración de una Vista Minable a partir de los datos del Repositorio Institucional UCI, se contribuirá al perfeccionamiento del proceso de toma de decisiones de los directivos del Centro de Informatización Científico Técnica y de la Dirección de Investigación de la universidad.*

Se ha identificado como **objeto de estudio**: *proceso de descubrir conocimiento en bases de datos utilizando almacenes de datos*; concibiendo como **campo de acción**: *vistas minables utilizando un mercado de datos en el Área de Informatización de la Biblioteca UCI*.

El **objetivo general** del presente trabajo es: *crear una vista minable a partir de un mercado de datos del Área de Informatización de la Biblioteca UCI, para contribuir al perfeccionamiento del proceso de toma de decisiones por los directivos del Centro de Informatización Científico Técnica y la Dirección de Investigación de la UCI*.

Para alcanzar el objetivo propuesto se plantean los **objetivos específicos**:

1. *Estudiar y realizar el marco teórico de la investigación.*
2. *Diseñar e implementar un mercado de datos para el área de informatización de la Biblioteca UCI.*
3. *Pre procesar los datos como paso inmediato a la creación de la vista minable para aplicar en un futuro la tarea de agrupamiento.*

Tareas de Investigación

- *Estudio del estado del arte y elaboración del marco teórico para enmarcar la situación actual existente con respecto a la investigación llevada a cabo.*
- *Estudio y selección de las herramientas y métodos a utilizar para la implementación del mercado de datos y la creación de la vista minable.*
- *Estudio y exploración de los datos almacenados en el repositorio y los relacionados con el acceso al mismo, para la ejecución del proceso ETL (extracción, transformación y carga) con respecto a la selección de los metadatos necesarios para la confección el mercado de datos.*
- *Estudio y selección de las técnicas de pre-procesamiento a emplear para garantizar la*

heterogeneidad de los datos del mercado de datos.

- *Aplicar las técnicas de pre-procesamiento sobre los datos del mercado de datos para la obtención de la vista minable.*

Tipo de Estudio

Descriptivo-Correlacional: ambos tipos de estudio se complementan, debido a que no existe experiencia previa en este tipo de soluciones. Es necesario enfocar como objeto de la ciencia la minería de datos aplicada en contextos bibliográficos, para posteriormente observar si se está aplicando o como se está llevando a cabo en la UCI, a través de variables medibles y sus relaciones. La investigación se rige por una estrategia descriptiva pues el estudio realizado corresponde al patrón de esta misma categoría (descriptivo), añadiendo, que sin tener previo conocimiento del fenómeno, provee el planteamiento de un claro objetivo a nivel de detalles, que tributa a una específica representación del problema y de los objetivos específicos a alcanzar de forma estructural y funcional. (3)

Métodos Científicos de Investigación

El método científico de investigación es la forma de abordar la realidad, de estudiar la naturaleza, la sociedad y el pensamiento, con el propósito de descubrir su esencia y sus relaciones. Se clasifican en teóricos y empíricos, los cuales están dialécticamente relacionados. (4)

Como sustento al empleo de los métodos teóricos y empíricos que a continuación se describen los autores utilizaron la dialéctica materialista.

Teóricos:

Analítico-sintético: se utilizó para analizar elementos bibliográficos y definiciones sobre sistemas existentes, que realicen un trabajo similar al propuesto, con el propósito de arribar a conclusiones que sustenten la necesidad de la investigación. Así como para la apropiación de conocimientos necesarios para el desarrollo de la presente investigación.

Histórico-lógico: se utilizó en la realización del estudio del estado del arte sobre las soluciones informáticas que se han utilizado a través de los almacenes de datos para el análisis de datos bibliográficos, con el objetivo de identificar tendencias que coadyuven a la elaboración de la propuesta que se realiza en este informe de investigación y a conocer cómo estas aplicaciones manejan la información para contribuir al perfeccionamiento de la toma de decisiones.

Modelación: se utilizó para lograr la relación entre el modelo y el objetivo, para ello se definen las dimensiones del mercado de datos, los metadatos, el hecho asociado a las dimensiones definidas además se estructura el modelo dimensional y se transforma al diseño físico.

Empíricos:

Encuesta: método utilizado para la obtención de información relevante ante el cliente, procesos del

negocio y datos importantes a tener en cuenta.

Observación científica: La observación científica es la percepción planificada dirigida a un fin y relativamente prolongada de un hecho o fenómeno. Es el instrumento universal del científico, se realiza de forma consciente y orientada a un objetivo determinado. (4)

Estructura de la Tesis:

Capítulo 1: Se identifican las principales definiciones relacionados con minería de datos, sus fases, y el contexto que da origen a la vista minable. También se exponen temas sobre los almacenes y mercados de datos. Por último, se describen las herramientas utilizadas durante el desarrollo de la investigación.

Capítulo 2: Este capítulo abarca todo el proceso de creación del mercado de datos, guiado por la "*Metodología de Desarrollo para Proyectos de Almacenes de Datos*", incluyendo desde los requisitos necesarios, hasta la integración de los datos provenientes de las distintas fuentes de datos.

Capítulo 3: Este es el último capítulo, en el mismo se procede a realizar parte del proceso de extraer conocimiento a partir de datos, precisamente todo el contexto que conlleva generar la vista minable con los datos adecuados para la futura aplicación de técnicas de minería de datos a un conjunto de datos específicos. Se explica el proceso guiado por la metodología para proyectos de minería de datos "*CRISP-DM*".

El presente informe cuenta con las conclusiones finales orientadas a exponer con claridad el cumplimiento de los objetivos planteados, recomendaciones de la investigación, referencias bibliográficas, bibliografías y anexos complementarios con algunos de los artefactos desarrollados durante el transcurso del proyecto investigativo.

Capítulo 1: Fundamentos teóricos para un proceso de descubrimiento de conocimiento en bases de datos mediante Almacenes de Datos.

En este capítulo se pretende identificar los principales conceptos relacionados con minería de datos, sus principales fases. Se hace mayor énfasis en la fase de pre procesamiento, pues la vista minable es el producto de esta fase. Se abordan temas sobre almacén y mercado de datos. Además, se describen las técnicas y herramientas necesarias para dar solución al problema identificado en el diseño de investigación expuesto.

1.1 Análisis de soluciones similares

Aunque el uso de técnicas de procesamiento analítico en línea para el análisis bibliográfico es reciente, en el mundo se han desarrollado soluciones en diferentes campos de las ciencias. Los aportes de este análisis son catalogados como importantes, teniendo en cuenta la cantidad de datos que se manejan y que la variedad de estudios que se realizan rebasan la capacidad de cálculo manual.

Entre las soluciones existentes se encuentra bgMath/OLAP, que se utiliza para el almacenamiento y procesamiento analítico en línea de datos bibliográficos sobre literatura científica relacionada con Matemática. Las funcionalidades básicas del sistema se dividen en cuatro grupos:

- Carga de los datos en el almacén de datos periódicamente por un horario determinado.
- Calcular y mantener los datos resumidos en el cubo de datos.
- Navegar por los datos resumidos con el propósito de su análisis por diferentes dimensiones en una tabla y una vista gráfica mediante la aplicación Microsoft Excel.
- Exportación de los datos que se resumen en PDF, HTML, XML, otros formatos. (5)

El almacén de datos fue desarrollado con la herramienta SQL Server Integration Services (Servidor de Servicios Integrados SQL). La cual cuenta con una licencia privativa y no es una aplicación de código abierto (5). Las características ya mencionadas de bgMath/OLAP indican que no es viable seleccionarla como solución al problema planteado porque está desarrollada para un área específica que no está relacionada con el problema en cuestión, con tecnologías privativas y su código no está disponible.

Otra solución existente, DBPus, es un sistema que integra la búsqueda de palabras clave en el cuerpo de los documentos con técnicas OLAP (Procesamiento Analítico en línea) en bases de datos para analizar mejor y visualizar grandes cantidades de datos. En esta solución se utilizan documentos de grandes conferencias y periódicos como SIGMOD (*Special Interest Group on Management Of Data), ICDE (International Council for Open and Distance Education), IEEE (Instituto de Ingeniería Eléctrica y Electrónica) y VLDB Journal (Very Large DataBases)* (6).

No es viable seleccionar la solución DBPus porque: aunque trabaja con datos bibliográficos y técnicas de almacenes de datos o procesamiento analítico en línea, no se posee el código de la solución y no tiene como objetivo final ofrecer una solución al problema planteado en la presente investigación.

En la Universidad de las Ciencias Informáticas se desarrolló un mercado de datos para el análisis de datos bibliográficos para el Instituto de Cibernética, Matemática y Física de Cuba (ICIMAF). El mercado de datos desarrollado tiene como objetivo principal facilitar el proceso de análisis de datos bibliográficos para ayudar en la toma de decisiones que realizan los especialistas del ICIMAF, a través de indicadores bibliométricos (7). El diseño del mismo está enfocado a dos procesos: productividad y predicción, para poder aplicar métodos matemáticos y estadísticos a los datos existentes en el almacén.

Aunque la solución en cuestión fue desarrollada con herramientas libres y de código abierto no se conoce su fuente de datos principal, el objetivo principal de la misma es apoyar a la toma de decisiones a través de indicadores bibliométricos y no a través de la aplicación de descubrimiento de conocimientos en bases de datos, por lo que, no tiene como objetivo final ofrecer una solución al problema planteado en la presente investigación. Por lo antes descrito en las soluciones analizadas se evidencia la necesidad de realizar un mercado de datos a través del cual se genere una vista minable que contribuya a mejorar el proceso de toma de decisiones por el CICT.

1.2 Minería de datos

El uso de información aprendida desde los datos, actualmente es un recurso necesario en muchos ámbitos sociales, como mantener una competitividad adecuada en entornos empresariales u optimizar decisiones de las instituciones públicas. No cabe duda que la minería de datos se ha convertido en los últimos años en un término muy popular, por lo que ha ido evolucionando desde su aparición, desarrollando nuevos métodos para adaptarse a las necesidades de una amplia variedad de aplicaciones.

Cada vez son más los usuarios, las aplicaciones, las investigaciones y los desarrollos relacionados con la misma, y a su vez crece también el número de sistemas que suelen ser productos de esta reciente disciplina.

La minería de datos no es más que: "*el proceso de extraer conocimiento útil y comprensible, previamente desconocido desde grandes cantidades de datos almacenados en distintos formatos*." (8)

1.2.1 Etapas del proceso de extracción de conocimiento a partir de datos

Existen términos que se utilizan frecuentemente como sinónimos de la minería de datos. Uno de

ellos es la extracción o descubrimiento de conocimiento en bases de datos (KDD, por sus siglas en inglés). Aunque ambos términos se han utilizado indistintamente existen claras diferencias entre los dos. Se define KDD como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos. Es el proceso global de descubrir conocimiento útil desde las bases de datos, mientras que la minería de datos se refiere a la aplicación de los métodos estadísticos y de aprendizaje para la obtención de patrones y modelos. (1)

Entonces se puede entender que la minería de datos forma parte de las etapas del KDD, lo cual se muestra en la Figura 1:

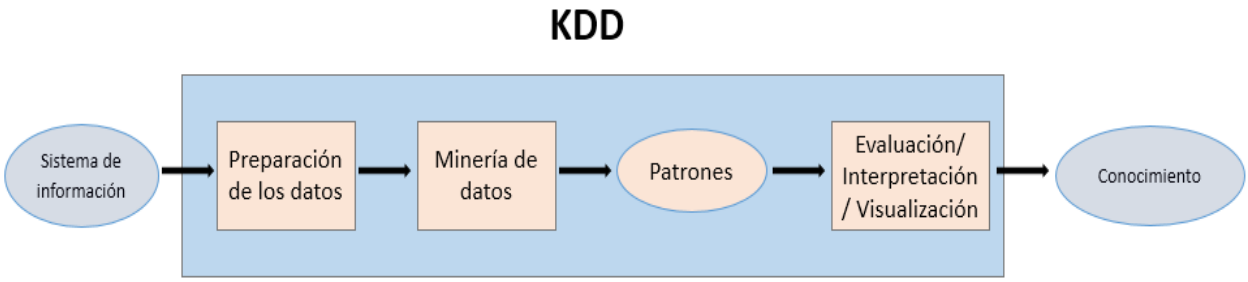

Figura 1: Proceso de KDD. (5)

Preparación de los datos

Dentro de la fase de preparación de los datos se realiza la integración y recopilación de los datos y la selección, limpieza y transformación de los mismos, después de esto se realiza el preprocesado de los datos.

- Integración y recopilación.

El problema de reunir un conjunto de datos que posibilite la extracción de conocimiento requiere decidir, entre otros aspectos, de qué fuentes (internas y externas) se van a obtener los datos iniciales, cómo se van a organizar, cómo serán conservados con el tiempo y de qué forma se va a extraer el mismo.

El mantenimiento de esta información plantea cuestiones técnicas, pues se requerirá añadir al sistema nueva información, ya sean actualizaciones como nuevas incorporaciones provenientes de la propia organización (fuentes internas) o de fuentes externas. Estos detalles junto a las características propias del proceso analítico de datos, han dado lugar al desarrollo de una tecnología nueva y específica, denominada "***almacenes de datos***". Estas y otras cuestiones relacionadas con los Almacenes de Datos, se explicarán en el epígrafe con dicha denominación.

- Selección, Limpieza y Transformación.

La recopilación de datos debe ir acompañada de una limpieza e integración de los mismos, para que éstos estén en condiciones para su análisis. Los beneficios del análisis y de la extracción de

conocimiento a partir de datos dependen, en gran medida, de la calidad de los datos recopilados. Además, generalmente, debido a las características propias de las técnicas de minería de datos, es necesario realizar una transformación de los datos para obtener una "*materia prima*" que sea adecuada para el propósito concreto y las técnicas que se quieren emplear. (1)

El concepto de "*calidad de datos*" se asocia cada vez más frecuentemente a los sistemas de información. En la mayoría de las bases de datos existe información incorrecta y datos inconsistentes respecto al dominio real que se desea cubrir. Dichos problemas suelen incrementarse cuando se integran datos de distintas fuentes. Para resolver esta disparidad existen técnicas de integración, de limpieza y de algunas transformaciones, para así convertir los datos en otros más apropiados para la minería. A continuación, se procede a describir cada una de estas fases individualmente por la importancia que tiene para la investigación:

Selección de datos: En esta etapa, se selecciona un subconjunto de los datos adquiridos en la fase anterior, apoyándose en criterios previamente establecidos en las fases anteriores: calidad de los datos en cuanto a completitud y corrección de los datos y limitaciones en el volumen, o en los tipos de datos que están relacionadas con las técnicas de minería de datos seleccionadas. (9)

Limpieza de los datos: Esta tarea complementa a la anterior, y es una de las que más tiempo y esfuerzo consume, debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos a objeto de prepararlos para la obtención de patrones. Algunas de las técnicas a utilizar en la presente investigación son: el tratamiento de valores ausentes en búsqueda de aumentar la calidad de los datos y reducción del volumen de datos, con el objetivo de sintetizar los datos a una cantidad representativa pero acorde a las exigencias de la vista minable a realizar. (9)

Estructuración o Transformación de los datos: Esta tarea incluye las operaciones de preparación de los datos tales como: la generación de nuevos atributos a partir de atributos ya existentes, integración de nuevos registros o transformación de valores para atributos existentes. (9)

- Técnicas de Pre procesamiento

El propósito del pre procesamiento de datos es principalmente corregir las inconsistencias de los mismos, que serán la base de análisis en procesos de minería de datos. Existen muchísimas técnicas creadas con estos fines y cada una de ellas con objetivos específicos, pues la presencia de errores en los datos puede reflejarse de disímiles maneras.

Dentro de las técnicas empleadas para el tratamiento de los datos se pueden mencionar:

Reemplazar valores perdidos: la aplicación de los estadígrafos **moda** o **media** en caso de ausencia de datos y dependiendo del tipo que sean. Si el atributo es nominal se utiliza la moda para los remplazar los campos nulos; de ser numérico entonces se aplica la media. En el caso de WEKA, esta técnica se aplica mediante el filtro “*Replace Missing Values*”, el cual determina el tipo de atributo

y sustituye por el valor correspondiente.

Lematización y stemming: esta técnica se utiliza para la eliminación de palabras vacías, tal es el caso de los artículos, preposiciones, prefijos y sufijos. De forma muy sencilla, se puede definir la lematización como el proceso de representar mediante un único término (el lema), todas las posibilidades flexivas de una palabra. Desde el punto de vista lingüístico, un lema es un término que representa y unifica todos los elementos de un conjunto de palabras morfológicamente similares. De forma similar, el stemming reduce un conjunto de palabras a su stem o raíz común. Así como *camión* sería la raíz de *camiones, camioneros, etc*. Uno de los primeros algoritmos de stemming, desarrollado para el idioma inglés en 1980, se debe a Martin Porter, por lo que se conoce como "*algoritmo de Porter*". (10)

La aplicación de esta técnica en WEKA se manipula mediante el filtro "*String To Word Vector*", el cual a través de los parámetros "*stemmer*" y "*useStopList*", permite elegir algoritmos que tratan de reducir las palabras a sus raíces y determinar si suprime las palabras vacías, respectivamente.

En la presente investigación se realizó una modificación del algoritmo que posee WEKA, para que se efectúe de forma correcta la lemantización de las palabras. El algoritmo está implementado por defecto para trabajar sobre idioma Inglés y posee una opción para utilizar el español pero no está bien diseñada en cuanto a la eliminación de los sufijos de las palabras, por esto se modificó para que se utilizara correctamente con el idioma español.

Minería de datos

El objetivo de esta fase es producir nuevos conocimientos que pueda utilizar el usuario. Esto se realiza construyendo un modelo basado en los datos recopilados para este efecto. El modelo es una descripción de los patrones y relaciones entre los datos que pueden usarse para hacer predicciones, para entender mejor los datos o para explicar situaciones pasadas. Para ello es necesario tomar una serie de decisiones antes de empezar el proceso (1):

- determinar qué tipo de tarea de minería es el más apropiado. por ejemplo, podríamos usar la clasificación para predecir una entidad bancaria los clientes que dejarán de serlo.
- elegir el tipo de modelo. por ejemplo para una tarea de clasificación usar un árbol de decisión, porque queremos obtener un modelo en forma de reglas.
- elegir el algoritmo de minería que resuelva la tarea y obtenga el tipo de modelo que estamos buscando. esta selección es pertinente porque existen muchos métodos para construir los modelos (1).

Como salida de esta fase se tienen los patrones obtenidos en la misma que son evaluados, interpretados y analizados en la siguiente etapa.

Evaluación, interpretación y visualización

Medir la calidad de los patrones descubiertos por un algoritmo de minería de datos no es un problema trivial, ya que esta medida puede llevar a varios criterios; algunos de ellos, bastante subjetivos. Idealmente, los patrones descubiertos deben tener tres cualidades: ser precisos, comprensibles (es decir, entendibles) e interesantes (útiles y novedosos). Según las aplicaciones puede interesar mejorar algún criterio y sacrificar ligeramente otro, como en el caso del diagnóstico médico que prefiere patrones comprensibles aunque su precisión no sea muy buena. (1)

En esta etapa se evalúan los patrones obtenidos para analizar si son posibles o no y se interpretan para inducir conocimiento.

Fase de difusión, uso y monitorización

Una vez construido y validado el modelo puede usarse principalmente con dos finalidades: para que un analista recomiende acciones basándose en el modelo y en sus resultados, o bien para aplicar el modelo a diferentes conjuntos de datos. También puede incorporarse a otras aplicaciones, como por ejemplo a un sistema de análisis de créditos bancarios, que asista al empleado bancario a la hora de evaluar a los solicitantes de los créditos, o incluso automáticamente, como los filtros de *spam* o la detección de compras con tarjetas de créditos fraudulentas. (1)

1.2.2 Contexto de Vista Minable

Una herramienta para realizar análisis de datos, no puede digerir un conjunto de datos y producir algo razonable, si no se le orienta. La razón fundamental del por qué, radica no sólo en la incapacidad actual de las herramientas de realizar algunas tareas de una manera completamente automática, sino, fundamentalmente, en que la extracción de conocimiento viene a cubrir unas necesidades y expectativas, que deben indicarse, en cierto modo, de forma interactiva. (1)

No será posible extraer conocimiento, sin expresar y proporcionar respuestas a las siguientes preguntas:

¿Qué parte de los datos es pertinente analizar?

Una **vista minable**: consiste en el sentido más clásico de base de datos: una tabla. La mayoría de métodos de minería de datos son solo capaces de tratar una tabla en cada tarea. Por tanto, la vista minable ha de recoger toda (y solo) la información necesaria para realizar la tarea de minería de datos.

¿Qué tipo de conocimiento se desea extraer y cómo se debe presentar?

Se trata de decidir qué **tarea** (clasificación, regresión, agrupamiento, reglas de asociación, etc.), con qué **método**, entre los existentes para cada tarea (árboles de decisión, redes neuronales, regresión logística, etc.) y de qué manera, se van a **presentar** o se van a navegar los **resultados**

(gráficamente, como un árbol, como un conjunto de reglas, etc.).
Obtener la vista minable, la tarea, el método, el conocimiento previo necesario, es un proceso iterativo, que se tornará más sencillo, según se vayan descubriendo los datos y su contexto, los usuarios, las técnicas de exploración y de minería de datos.

1.2.3 Tareas de la Minería de Datos

Dentro de la minería de datos se distinguen varios tipos de tareas, cada una de las cuales puede considerarse como un tipo de problema a ser resuelto por un algoritmo de minería de datos. Esto significa que cada tarea tiene sus propios requisitos, y que el tipo de información obtenida con una tarea puede diferir mucho de la obtenida con otra.
Las tareas pueden ser **predictivas** o **descriptivas**. Entre las predictivas se encuentran la clasificación, la regresión y las series temporales, mientras que el agrupamiento (clustering), las reglas de asociación, las reglas de asociación secuenciales y las correlaciones son tareas descriptivas.

Tarea de minería de datos a la que se perfilan los datos en la propuesta de solución

El **agrupamiento** (clustering) es la tarea descriptiva por excelencia y consiste en obtener grupos "naturales" a partir de los datos. Los datos son agrupados basándose en el principio de maximizar la similitud entre los elementos de un grupo minimizando la similitud entre los distintos grupos. Es decir, se forman grupos tales que los objetos de un mismo grupo son muy similares entre sí y, al mismo tiempo, son muy diferentes a los objetos de otro grupo. Al agrupamiento también se le suele llamar segmentación, ya que parte o segmenta los datos en grupos que pueden ser o no disjuntos. (1)

1.2.4 Técnicas de minería de datos

Todas las tareas anteriormente mencionadas, requieren métodos o técnicas que representen el enfoque conceptual para extraer conocimiento y obtener modelos a partir de los datos; además, existen varios algoritmos que se pueden emplear para su implementación, cada uno de ellos con sus características específicas.
Existen paradigmas estipulados detrás de las disímiles técnicas de minería de datos que se pudieran utilizar, entre las más comunes están: técnicas de inferencia estadística, árboles de decisión, redes neuronales, inducción de reglas, aprendizaje basado en instancias, algoritmos genéticos, aprendizaje bayesiano, programación lógica inductiva y varios tipos de métodos basados en núcleos, entre otros. También coexisten otros tipos de restricciones que hacen que la efectividad del algoritmo dependa del dominio de aplicación, no existiendo lo que podríamos llamar el "*método universal*"

aplicable a todo tipo de aplicación.

En el caso particular de agrupamiento, existen varios algoritmos y variaciones de ellos, que pudieran ser factibles para implementar dicha tarea de minería de datos. Se puede citar como los más utilizados al k-means o k-medias, agregando también una variante del mismo, conocida como k-medoides. (11)

1.2.5 Objetivos de la Minería de Datos en la solución.

Con el desarrollo de esta investigación, se pretende generar una vista minable, a la cual posteriormente, un analista de datos le aplicará técnicas de minería de datos, para la obtención de patrones.

En la Figura 2, se muestra hasta donde abarcará la presente investigación, siguiendo la guía del proceso KDD.

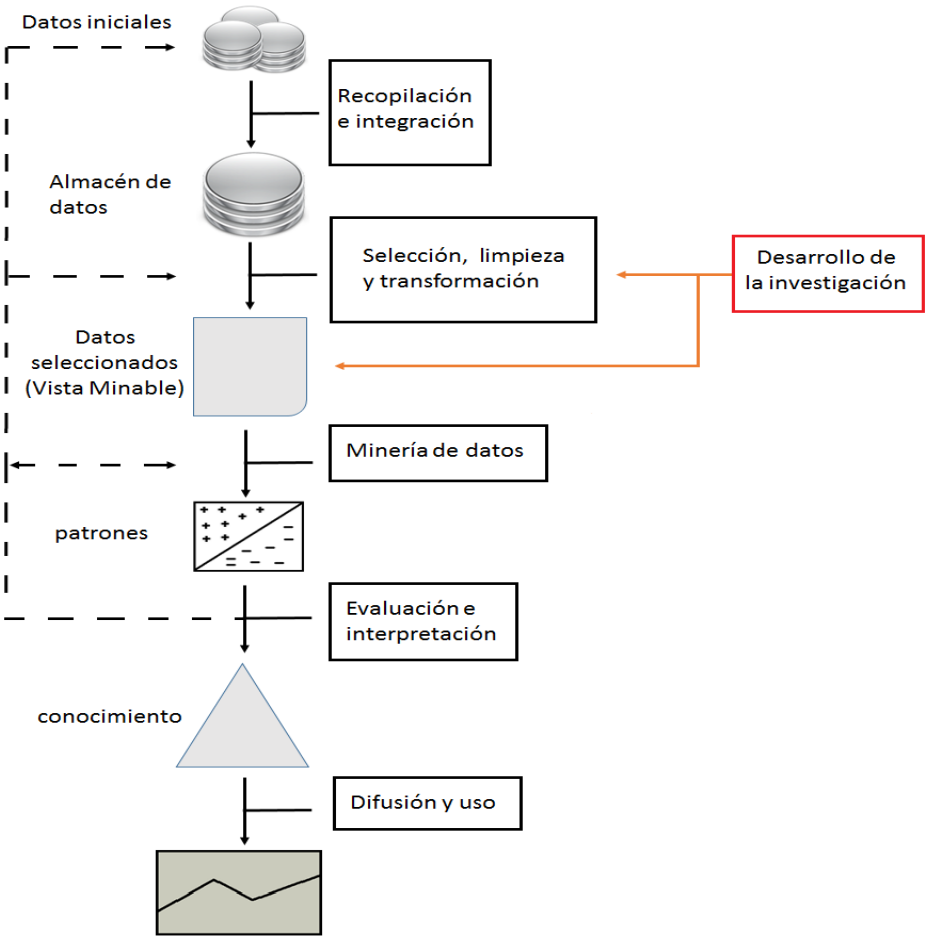

Figura 2 Fase hasta donde abarca la presente Investigación en el proceso de KDD.

1.2.6 Metodología para el desarrollo de un proceso de extracción de conocimiento a partir

de datos

Para un mejor desarrollo de un proceso de KDD resulta de utilidad el empleo de alguna metodología como guía para poner en marcha un proyecto de este tipo. De esta forma, se facilita la planificación y dirección del proyecto, así como permite un mejor seguimiento de este. (12) (sic)

Varias son las metodologías que pudieran emplearse ante la ejecución de un proceso de KDD. Entre las conocidas y empleadas se pueden incluir: **SEMMA**, creada por SAS Institute y su nombre se debe al acrónimo correspondiente a las cinco fases básicas del proceso: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado), Assess (Valoración). Se encuentra enfocada especialmente en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema que se está abordando. Su principal desventaja, es que fue propuesta especialmente para trabajar con el software de minería de datos de la compañía SAS.

Otra metodología existente es **Catalyst**, también nombrada como **P3TQ** (Product, Place, Price, Time, Quantity), plantea la formulación de dos modelos: el Modelo de Negocio proporciona una guía de pasos para identificar un problema de negocio (o la oportunidad del mismo) y los requerimientos reales de la organización. En cambio el Modelo de Explotación de Información sintetiza una guía de pasos para la construcción y ejecución de modelos de minería de datos a partir del Modelo de Negocio.

La última metodología estudiada y actualmente la guía más utilizada en el desarrollo de proyectos de Minería de Datos es **CRISP–DM**, creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000. Estructura el proceso en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación, y la sucesión de estas no es necesariamente rígida. Esta metodología establece un conjunto de tareas y actividades para cada fase del proyecto. (13)

Justificación de la metodología utilizada en la propuesta de solución.

DMAMC (Definir, Medir, Analizar, Mejorar, Controlar) o CRISP-DM (Cross Industry Standard Process for Data Mining, por sus siglas en inglés), es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de minería de datos y la más documentada. Además de contar con ejemplos palpables de su eficiencia una vez en práctica. (9)

Fases de la Metodología CRISP-DM

Está dividida en cuatro niveles de abstracción organizados de forma jerárquica (fases, tareas generales, tareas especializadas e instancias de proceso), tareas que van desde el nivel más general hasta los casos más específicos y organiza el desarrollo del proyecto en una serie de seis fases. (9) Dichas fases se exponen a continuación:

1. **Comprensión del negocio o problema**: La primera fase es la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. (9)
2. **Comprensión de los datos**: La fase de comprensión de los datos, comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. (9)
3. **Preparación de los datos**: En esta fase y una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se utilicen posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos. La preparación de datos incluye tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. (9)
4. **Modelado**: La fase presente, procede a la selección las técnicas de modelado más apropiadas para el proyecto de minería de datos específico. (9)
5. **Evaluación**: En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que posiblemente se haya cometido algún error. Considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados. (9)
6. **Implementación**: Una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso. Generalmente un proyecto de minería de datos no concluye en la implantación del modelo, pues se deben documentar y presentar los resultados de manera comprensible para el usuario, con el objetivo de lograr un incremento del conocimiento. (9)

Es válido acotar, que el objetivo de la presente investigación es la generación de una vista minable, que precisamente es el artefacto de salida de la fase "*Preparación de los datos*" de un proceso KDD. El alcance de la investigación abarca hasta esta importante fase, aunque los datos contemplados por la tabla de datos o vista minable se perfilarán para la realización de la tarea de agrupamiento

mediante el algoritmo K-Means.

En realidad, se puede hacer minería de datos sobre un simple archivo de datos. Sin embargo, las ventajas de organizar un almacén de datos se amortizan sobradamente a mediano y largo plazo. Esto es especialmente patente cuando se enfrentan grandes volúmenes de datos o aumentan con el tiempo, o provienen de fuentes heterogéneas o se van a querer combinar de maneras arbitrarias y no predefinidas. Incluso si todos los datos originalmente no provienen de bases de datos puede ser conveniente la realización de un almacén de datos. En gran medida, facilita la limpieza y la transformación de datos (en especial para generar "vistas minables" en tiempo real). (1)

Por los beneficios que brinda y aunque no son imprescindibles para hacer extracción de conocimiento a partir de datos, se opta por implementar un mercado de datos auxiliar en la presente investigación.

1.3 Almacén de datos

Cada día es más necesario distinguir dos usos diferentes de los sistemas dedicados a la gestión de información: el procesamiento transaccional y el procesamiento analítico.

El **procesamiento transaccional en línea** (**OLTP**, por sus siglas en inglés), constituye el trabajo primario de un sistema de información. Este trabajo consiste en realizar transacciones, es decir, actualizaciones y consultas a la base de datos con un objetivo operacional: hacer funcionar las aplicaciones de la organización, proporcionar información sobre el estado del sistema de información y permitir actualizarlo conforme va variando la realidad del contexto de la organización.

Por su parte, el **procesamiento analítico en línea** (**OLAP**, por sus siglas en inglés) engloba un conjunto de operaciones, exclusivamente de consulta, en las que se requiere agregar y cruzar gran cantidad de información. El objetivo de estas consultas es realizar informes y resúmenes, generalmente para apoyar la toma de decisiones. (1)

Hasta hace pocos años en muchas organizaciones y empresas, la práctica de ambos tipos de procesamiento (OLTP y OLAP) se realizaban sobre la misma base de datos transaccional, sin tener en cuenta dos problemas fundamentales: primero, las consultas OLAP perturban el trabajo transaccional diario de los sistemas de información originales. Al ser consultas complejas y que involucran muchas tablas y agrupaciones, suelen consumir gran parte de los recursos del sistema de gestión de base de datos.

El resultado es que durante la ejecución de estas consultas, las operaciones transaccionales normales se resienten: las aplicaciones van más lentas, las actualizaciones toman más tiempo y el sistema puede incluso llegar a colapsar. De este hecho viene el nombre familiar que se les da a las consultas OLAP: "killer queries" (consultas asesinas). Por este motivo, estas consultas se deben ejecutar en horarios de poca o ninguna demanda transaccional.

Segundo, la base de datos está diseñada para el trabajo transaccional, no para el análisis de los datos. Esto significa que, aunque tuviéramos el sistema dedicado exclusivamente para realizar una consulta OLAP, dicha consulta puede requerir mucho tiempo, no sólo por ser compleja intrínsecamente, sino porque el esquema de la base de datos no es el más adecuado para este tipo de consultas. (1)

Ambos problemas implican que es prácticamente imposible, a un costo razonable de *hardware*, realizar un análisis complejo de la información en tiempo real, si ambos procesamientos se realizan sobre la misma base de datos operativa. Por lo que parece razonable recoger (copiar) los datos en un sistema unificado y diferenciado del sistema tradicional transaccional u operacional, aunque esto vaya contra la filosofía general de bases de datos. Desde esta perspectiva, se separa definitivamente la base de datos con fines transaccionales de la base de datos con fines analíticos, y nacen los almacenes de datos. (1)

Se asume que la definición original de un almacén de datos (AD) data de hace más de una década según Inmon: *"una colección de datos, orientada a un dominio, integrada, no volátil y variante en el tiempo para ayudar en las decisiones de dirección"*. (14)

El almacén de datos está pensado para cubrir las necesidades de información de toda la organización, debido a su gran alcance resulta difícil aprovechar todas las ventajas que brinda a nivel departamental. La necesidad de obtener vistas más finas de los datos, respondería al interés de poder orientar las potencialidades de esta tecnología a cada área de la organización, dando paso al denominado mercado de datos.

1.3.1 Mercado de Datos

Immon plantea, "*un mercado de datos es una versión especial de un almacén de datos con el objetivo de responder a un determinado análisis, función o necesidad y con una población de usuarios específica*." (14) La diferencia entre un almacén de datos y un mercado de datos es su alcance.

El almacén de datos de una empresa es la unión de cada uno de los mercados de datos por departamentos de la misma. (15)"

Al presente trabajo se ajusta más el concepto de Mercado de datos (MD) como: "*un conjunto de hechos y datos organizados para soporte decisional basados en la necesidad de un área o departamento específico*." (16)

Los datos son orientados a satisfacer las necesidades particulares de un departamento dado, teniendo solo sentido para el personal de ese departamento y sus datos no tienen por qué tener las mismas fuentes que los de otro mercado de datos.

Justificación del tipo de Almacén empleado en la propuesta de solución

Los almacenes de datos no son estrictamente necesarios para realizar minería de datos, aunque suelen ser extremadamente útiles si se va a trabajar con grandes volúmenes de información provenientes de diversas fuentes, que varían al pasar del tiempo y a los que se le pudieran realizar tareas de minerías de datos variadas, abiertas y cambiantes. Esto se debe, a que han hecho posible el almacenamiento de grandes volúmenes de datos en un mismo repositorio, y junto al incremento en potencia de la computación, son las causas de que las empresas de hoy en día busquen herramientas y tecnologías capaces de extraer información útil de los datos.

Enmarcado el contexto del negocio, la biblioteca UCI solo constituye un sub-ámbito o área de la comunidad UCI. Teniendo en cuenta las ventajas antes expuestas de los mercados de datos en un proyecto de minería de datos, se deduce que es bastante factible la implementación del mismo.

1.3.2 Modelo multidimensional

El análisis de los datos en un rango de tiempo determinado, trajo consigo estudios en búsqueda de optimizar la forma de almacenar y representar los mismos, en aras de que puedan ser consultados de forma eficiente. El modelo conceptual de datos más utilizado para almacenes de datos es el multidimensional y su uso es una de las aproximaciones más acertadas y seguidas por los especialistas.

El modelo dimensional se basa en la dualidad hecho-dimensión, un hecho representa una actividad objeto de análisis, actividad que está caracterizada por un conjunto de dimensiones y que posee además, un grupo de variables cualitativas llamadas medidas. (17) Entonces, un hecho no es más que el "*que*" se quiere analizar, así como las cualidades "*como*", "*cuando*" y "*donde*" son representadas mediante las dimensiones y las medidas representan el "*cuanto*".

Las bases de datos multidimensionales implican tres variantes posibles de modelación, que permiten realizar consultas de soporte de decisión: esquema en estrella, esquema copo de nieve y esquema constelación o copo de estrellas.

Esquema estrella

El esquema estrella consta de una tabla de hecho central y varias tablas de dimensiones relacionadas a esta, a través de sus respectivas llaves primarias. Es el esquema más simple que hay de interpretar, (18) tal como se muestra en la Figura 3:

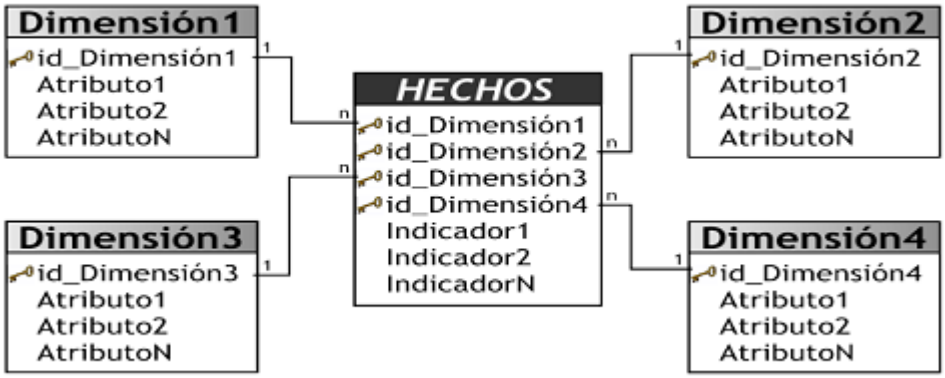

Figura 3: Ejemplo de Esquema estrella. (18)

Esquema copo de nieve

Este esquema representa una extensión del modelo en estrella cuando las dimensiones se organizan en jerarquías de dimensiones. Al igual que en el modelo en estrella, existe una tabla de hecho central y varias tablas de dimensiones, las cuales, a su vez, pueden estar relacionadas con otras tablas de dimensiones. Uno de los motivos principales para utilizar este tipo de modelo es la posibilidad de segregar los datos de las dimensiones y proveer un esquema que sustente los requerimientos de diseño. (18) En la Figura 4 se visualiza el mismo:

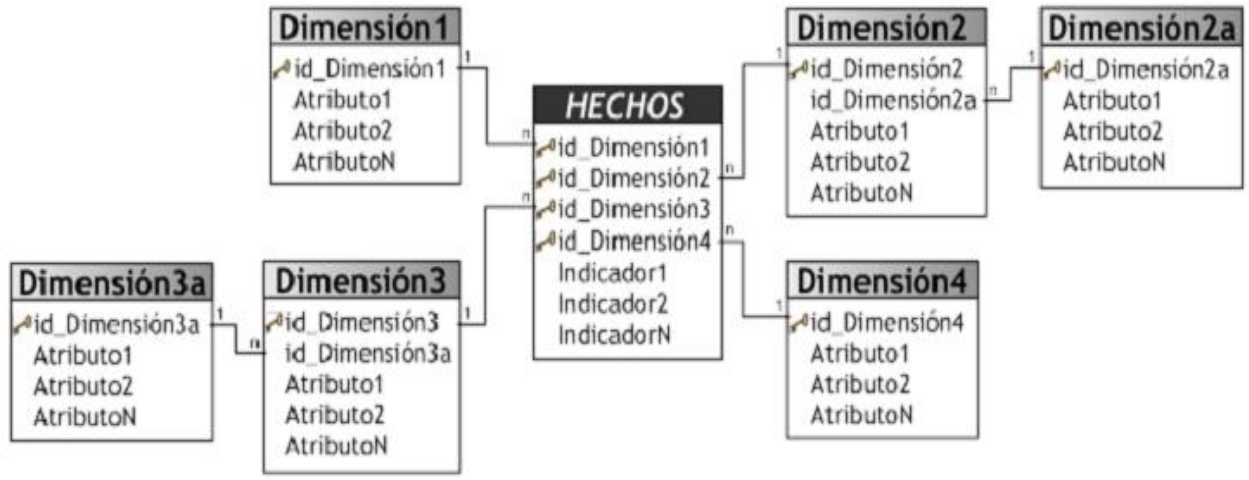

Figura 4: Ejemplo de Esquema copo de nieve. (18)

Esquema constelación o copo de estrellas

Un esquema constelación o copo de estrellas está formado por una tabla de hecho principal y por una o más tablas de hechos auxiliares. Cada una de estas tablas tiene relación con sus respectivas

tablas de dimensiones. No es necesario que las diferentes tablas de hechos compartan las mismas tablas de dimensiones, aunque se puede dar el caso. (18) Seguidamente se muestran los detalles del esquema copo de estrellas en la Figura 5:

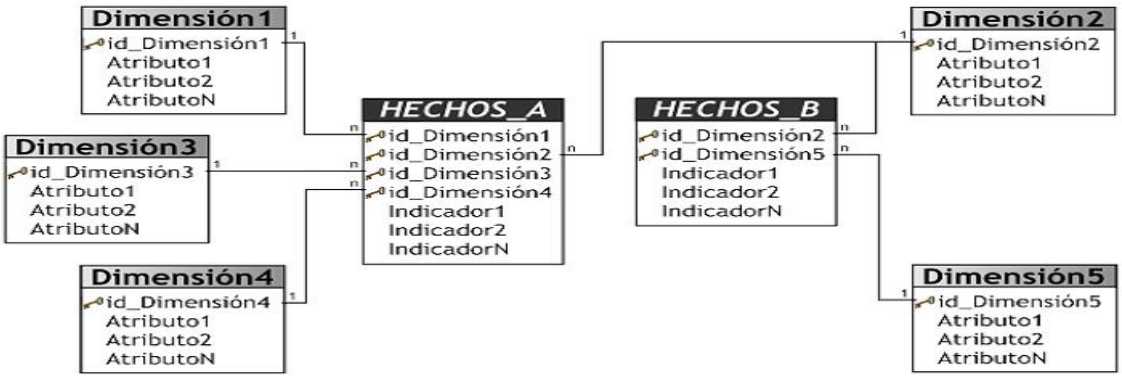

Figura 5: Ejemplo de Esquema constelación o copo de estrellas. (18)

En el presente trabajo investigativo se precisa que el esquema "***constelación o copo de estrella***" es el idóneo para el modelado del mercado de datos; debido a que permite varias tablas de hechos, adicionando que en algunos casos, las mismas comparten algunas dimensiones, como en el caso de tabla *dim_tiempo*.

1.3.3 Sistema de Extracción, Transformación y Carga (ETL)

La carga y mantenimiento de un almacén de datos es uno de los aspectos más delicados y de mayor esfuerzo. Para ello suele existir un sistema especializado, denominado "Sistema de Extracción Transformación y Carga" (ETL, por sus siglas en inglés).

El sistema 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, transformarlos a través de diferentes reglas (reglas ETL) para aumentar la calidad y consistencia de los mismos, consolidar la información proveniente de distintos sistemas, y finalmente cargar (grabar) la información en el AD en un formato acorde para la utilización por parte de las herramientas de análisis. (19)

1.3.4 Metodología para el desarrollo de almacenes de datos

Se puede citar Metodología como: un conjunto de procedimientos racionales que se utilizan para lograr una serie de objetivos a lo largo de una investigación científica. (3)

Actualmente existen muchas metodologías de diseño y construcción de almacén de datos. Cada fabricante de software de inteligencia de negocios busca imponer una metodología con sus productos. Sin embargo, se imponen entre la mayoría dos metodologías: la de **Kimball** y la de **Immon**, denominadas así debido al nombre de sus autores, quienes son especialistas considerados

líderes atendiendo al prestigio y los planteamientos realizados en el dominio de AD. La principal diferencia en los planteamientos realizados por ambos autores se centra en el contexto arquitectónico, o sentido de la construcción del AD. (20)

Enfatizando en el concepto de MD descrito anteriormente, Kimball (19) propone un método de almacenamiento de datos bottom up, en el que los MD individuales ofrecen vistas finas de los datos de cada área de la organización, que podrían ser combinados en un almacén de datos general.

Por otra parte, Inmon (14) refleja el punto de vista opuesto. Es decir, que el AD debe ser diseñado desde arriba hacia abajo (top down) para incluir todos los datos corporativos. Por este método, los MD son creados solo después de que los datos estén completos en el almacén.

Existen diversas metodologías que han surgido a raíz de la combinación entre algunos elementos de las metodologías anteriores y necesidades particulares de los propios fabricantes de productos. Entre las notorias se pueden mencionar las siguientes:

- Metodología **Hefestos**: entre sus principales directrices, plantea que la construcción e implementación de un AD puede adaptarse a cualquier ciclo de vida de desarrollo de software, con la salvedad de que para algunas fases en particular, las acciones a realizar serán diferentes. Se basa en no poseer fases extensas o un desarrollo monolítico que conlleve demasiado tiempo. Busca entregar una primera implementación que satisfaga parte de las necesidades y de esta manera motivar a los usuarios. (21)
- Metodología **DM2**: se basa en las necesidades de información a nivel gerencial, donde la información debe estar accesible a quien la necesite. Por la propia naturaleza del ambiente, el modelo cumple con su objetivo (atender las necesidades de información del nivel gerencial y ejecutivo de una empresa), esta metodología se asemeja a la forma descendente que propone Inmon y acorta en función razonable el tiempo entre el inicio del análisis y la implantación. Esta rapidez no solo es buena para el cliente sino que también es exigida y necesaria por el propio ambiente que lo rodea. (22)
- **Metodología de Desarrollo para Proyectos de AD**: toma como base la metodología de Kimball para definir los aspectos específicos del desarrollo de almacenes de datos. Los temas asociados al modelo de calidad Capability Maturity Model Integration (CMMI, por sus siglas en inglés) en su nivel dos, se incorporan a partir del Programa de Mejora, cuyo objetivo es definir e implementar los procesos necesarios para reducir los problemas en la producción de software en la universidad. (23)

Justificación y descripción de la metodología utilizada en la propuesta de solución.

La selección de la **"Metodología de Desarrollo de Proyectos de Almacenes de Datos**", se debe al estrecho vínculo con el entorno de la organización, pues fue diseñada y elaborada para proyectos de la universidad. Pretendiendo como objetivos, adaptar y formar a los especialistas y estudiantes del centro que se enfrentan por primera vez al desarrollo o liderazgo de un proyecto de AD.

Fases de la Metodología de Desarrollo para Proyectos de Almacenes de Datos

El ciclo de vida de la metodología se divide en siete fases y un flujo de trabajo. Algunas fases podrán ser implementadas de forma paralela como es el caso de la fase de Requisitos y Arquitectura, además durante la fase de diseño e implementación podrán desarrollarse varios componentes al mismo tiempo, esto permite un desarrollo más ágil. El flujo de trabajo de Gestión del Proyecto se ejecuta durante todo el ciclo de vida del proyecto. (23) Ver figura 6:

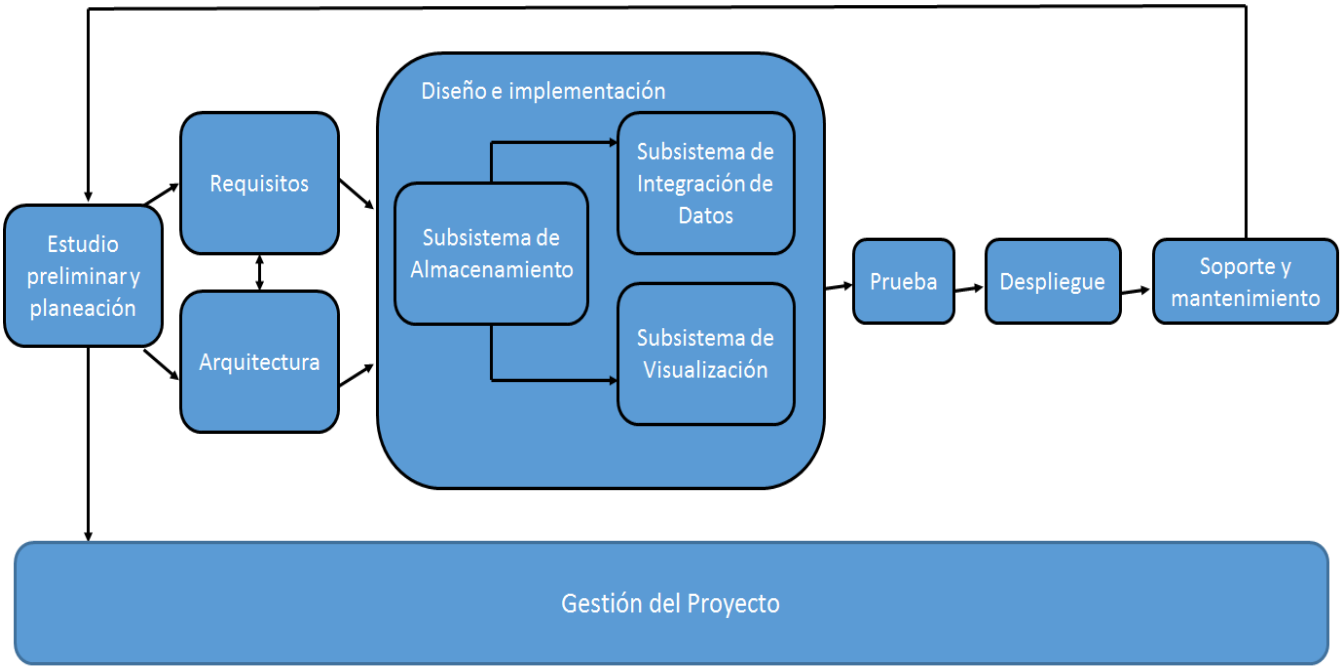

Figura 6 Ciclo de vida de la metodología. (20)

A continuación se detalla una breve descripción de cada fase del ciclo de vida de la metodología, incluyendo el flujo de trabajo de la gestión del proyecto.

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, definidos a partir del Programa de Mejora en sus libros de procesos (CALISOFT) y los procesos de las áreas de conocimiento de la dirección de proyectos definidos en la Guía del PMBOK. Los mismos fueron propuestos en función de gestionar y controlar aspectos claves del desarrollo del proyecto como son: los gastos, las utilidades, los recursos, las

adquisiciones, los planes y cronogramas. (23)

1. **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: Se analizan los principales procesos de negocio, objetivos y metas a los que responden, información que manejan, recursos humanos implicados, entidades externas involucradas entre otros aspectos.
    - Diagnóstico de los datos: Se analiza el estado de los datos que se manejan en la organización teniendo en cuenta aspectos como: fuentes de datos de los procesos del negocio, formato en el cuál se encuentran almacenados, sistemas informáticos que los manejan, utilización e importancia de los mismos y volumen almacenado.
    - Diagnóstico de la infraestructura tecnológica: Se analiza el estado tecnológico de la organización, características de los locales tecnológicos, calidad de los centros de datos, características de la redes de comunicación, entre otros aspectos.

    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.
2. **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: 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.
    - Requisitos funcionales: Describen las funcionalidades a realizar por el equipo de desarrollo. Estos requisitos incluyen las funcionalidades que deben implementarse en los subsistemas que se desarrollan en soluciones de AD.
    - Requisitos no funcionales: Los requisitos no funcionales describen las propiedades y cualidades que debe tener la solución. Representan las características del producto. (24)

3. **Arquitectura**: En esta fase se definen los aspectos arquitectónicos de la solución.
4. **Diseño e Implementación**: En esta etapa se obtiene el producto de software, se diseñan e implementan los tres subsistemas que conforman el AD (almacenamiento, integración y visualización). Cada subsistema puede verse como un componente de software que se desarrolla de forma independiente, para luego ser integrados conformando el producto final.
5. **Prueba**: En esta fase se realizan las pruebas necesarias para validar la calidad del software una vez implementado el mismo.
6. **Despliegue**: En esta fase se realiza la instalación del sistema en la organización o entidad cliente para que pueda ser utilizado por los usuarios del negocio. Por lo general consta de un despliegue piloto, donde se configuran los servidores necesarios y se instalan las herramientas según la arquitectura definida, se carga una muestra de los datos en un ambiente controlado, con el fin de mostrarle al cliente final el sistema en funcionamiento.
7. **Soporte y Mantenimiento**: Su objetivo es evitar que el sistema quede obsoleto o fuera de servicio por fallos en su funcionamiento. Comienza cuando la solución está implantada y en explotación, y se ejecuta según las condiciones de soporte establecidas. (23)

Debido a que el objetivo del mercado de datos a implementar es brindar soporte decisional al área de informatización de la biblioteca UCI mediante minería de datos y sus respectivas técnicas, en la presente investigación no se tendrán en cuenta las tres fases finales planteadas por la metodología seleccionada, que son: Prueba, Despliegue y Soporte y mantenimiento.

No se utilizaron pruebas para validar el MD pues la convencional Prueba de estrés no brinda datos relevantes sobre la validez del MD por las condiciones en las cuales se utilizará el mismo. No se realizará el despliegue del MD para la utilización del mismo por los usuarios porque se utilizará para la realización de la Vista Minable, y por último, no es objetivo de la investigación proporcionar soporte y mantenimiento al MD.

1.4 Almacenes de datos y minería de datos

Los almacenes de datos no son solamente útiles en ámbitos empresariales e institucionales, sino que también, pueden brindar grandes ventajas para disímiles procesos y análisis que lo requieran. Ejemplo de ello se muestra en la Figura 7, donde se pueden observar las distintas aplicaciones asociadas a la aplicación de AD. La señal en rojo indica el propósito del AD en el ciclo general de la investigación:

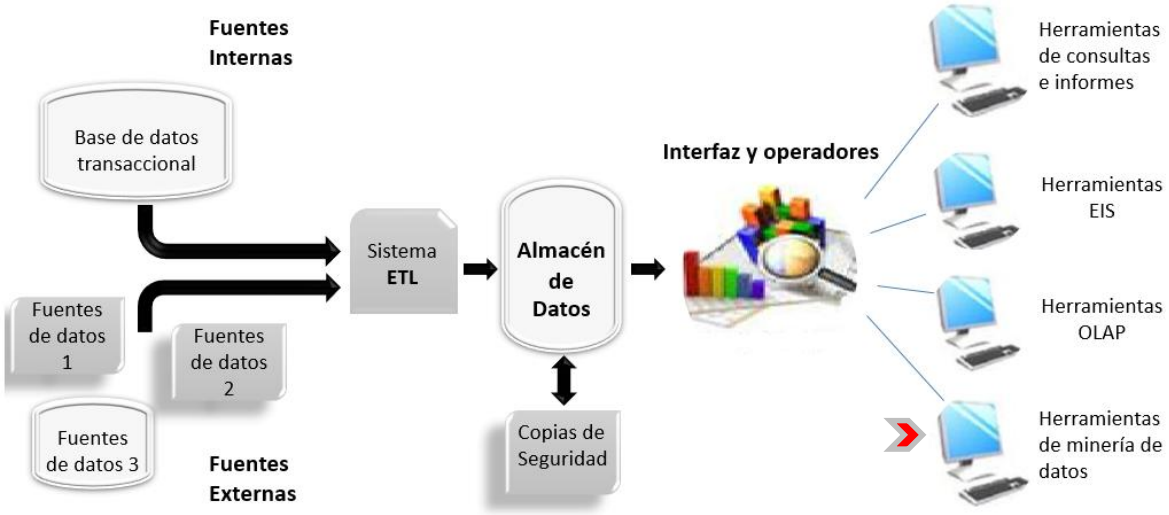

Figura 7: Aplicaciones y usos de los AD.

1.5 Herramientas a utilizar

A continuación se exponen y definen las herramientas seleccionadas por los autores del trabajo de diploma para alcanzar el objetivo propuesto en el diseño investigativo.

Sistemas gestores de bases de datos:

La empresa **Oracle** es una de las mejores en el campo de los sistemas gestores de bases de datos, pero sus productos para ser utilizados requieren la compra de una costosa licencia (25). **MySQL** por otra parte es uno de los gestores más utilizado por los usuarios en Internet, por la velocidad ofrecida a la hora de realizar las operaciones, convirtiéndolo en uno de los gestores que brindan mayor rendimiento, también es válido destacar características como bajo consumo, grandes utilidades de administración y facilidad de configuración e instalación.

Por último, el gestor **PostgreSQL** además de ser una derivación libre, lo cual es ventajoso pues permite modificar el código de la aplicación para utilizarlo como necesite el usuario, también es libre de costo. Teniendo en cuenta también que es el sistema gestor de bases de datos que emplea el Repositorio Institucional, se ha seleccionado a este último en su versión 9.3 para cumplir con el objetivo general de investigación.

A continuación se muestran las algunas características además de las antes mencionadas, que se tienen en cuenta del sistema gestor de bases de datos seleccionado:

- Implementación del estándar SQL92/SQL99.
- Soporta distintos tipos de datos: además del soporte para los tipos base, también soporta datos de tipo fecha, monetarios, elementos gráficos, datos sobre redes (MAC, IP), cadenas de bits, etc. También permite la creación de tipos propios.

- Permite la declaración de funciones propias, así como la definición de disparadores.
- Soporta el uso de índices, reglas y vistas.
- Permite la gestión de diferentes usuarios, como también los permisos asignados a cada uno de ellos. (26)

Herramientas de modelado

Las herramientas de modelado o CASE (***c****omputer* ***a****ided* ***s****oftware* ***e****ngineering*, ingeniería asistida por computadora) son diversas aplicaciones informáticas destinadas a aumentar la productividad en el desarrollo de software reduciendo el costo de las mismas en términos de tiempo y de dinero. (27)

Se pueden clasificar teniendo en cuenta los siguientes parámetros:

1. Las plataformas que soportan.
2. Las fases del ciclo de vida del desarrollo de sistemas que cubren.
3. La arquitectura de las aplicaciones que producen.
4. Su funcionalidad.

Entre las herramientas de modelado de más preferencia están:

Microsoft Project, un software de administración de proyectos diseñado, desarrollado y comercializado por Microsoft para asistir a administradores de proyectos en el desarrollo de planes, asignación de recursos a tareas, dar seguimiento al progreso, administrar presupuesto y analizar cargas de trabajo.

MagicDraw, herramienta de modelado con completas características UML.

Visual Paradigm: 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. Presenta licencia gratuita y comercial. Es fácil de instalar, actualizar y compatible entre ediciones. (28)

Dentro de la alta gama de herramientas CASE se selecciona Visual Paradigm, pues además de contar con previa experiencia con la misma, cuenta con características fundamentales, las cuales se mencionan a continuación:

- Disponibilidad en múltiples plataformas (Windows, Linux).
- Capacidades de ingeniería directa e inversa.
- Soporte de UML versión 2.1.
- Ingeniería inversa Java, C++, Esquemas XML, XML, NET exe/dll, CORBA IDL.
- Diagramas de flujo de datos.
- Ingeniería inversa de bases de datos - Desde Sistemas Gestores de Bases de Datos (DBMS)

existentes a diagramas de Entidad-Relación. (28)

Se utilizarán para el modelado en el Visual Paradigm el "*Lenguaje Unificado de Modelado*" (UML, por sus siglas en inglés), para artefactos como: diagrama de casos de uso y el modelo de datos del MD, y el estándar internacional de modelado de procesos "*Notación de Modelado de Procesos de Negocio*" (BPMN, por sus siglas en inglés), para el proceso de extracción, transformación y carga de los datos del MD.

Lenguaje Unificado de Modelado (UML): Es un lenguaje gráfico para visualizar, especificar, construir y documentar un sistema. UML ofrece un estándar para describir un plano del sistema (modelo), incluyendo aspectos conceptuales tales como procesos de negocio, funciones del sistema, y aspectos concretos como expresiones de lenguajes de programación y esquemas de bases de datos. (29)

Business Process Modeling Notation (BPMN) es una notación gráfica que describe la lógica de los pasos de un proceso de Negocio. Esta notación ha sido especialmente diseñada para coordinar la secuencia de los procesos y los mensajes que fluyen entre los participantes de las diferentes actividades. (30)

Plataforma para el desarrollo del mercado de datos

La plataforma Open Source Pentaho Business Intelligence cubre las amplias necesidades de Análisis de los Datos y de los Informes empresariales. Las soluciones de Pentaho están escritas en Java y tienen un ambiente de implementación también basado en Java. Eso hace que Pentaho sea una solución muy flexible para cubrir una amplia gama de necesidades empresariales tanto las típicas como las sofisticadas y específicas al negocio. (31)

Para cumplir con esa flexibilidad que lo caracteriza, Pentaho cuenta con tres módulos fundamentales:

- **Kettle ETL** (Pentaho Data Integration) se encarga de implementar los procesos ETL para la integración de datos.
- **Mondrian** suministra a los usuarios un sistema avanzado de análisis de información.
- **Pentaho Reporting** permite generar informes de gran capacidad, la distribución de los resultados del análisis en múltiples formatos y la programación de tareas de ejecución automática de informes con una determinada periodicidad.

En la investigación presente se utiliza el módulo Kettle ETL para la integración de los datos en el MD.

Herramienta para la creación de la vista minable

Actualmente existen muchas aplicaciones que se pueden emplear para realizar un proceso de KDD, destacando SPSS Clementine y SAS Enterprise Miner, que constituyen aplicaciones líderes en el mercado, basadas en las metodologías CRISP-DM y SEMMA respectivamente. No obstante, son herramientas comerciales, y su adquisición puede ser costosa.

Por otro lado, existen aplicaciones de código abierto y de libre distribución como los casos de YALE (Rapid Miner) y WEKA. El caso de YALE, constituye una herramienta creada en la Universidad de Dortmund para el descubrimiento del conocimiento y minería de datos, WEKA, por su parte, es una de las aplicaciones de minería de datos más populares en la actualidad, desarrollada por un equipo de investigadores de la Universidad de Waikato (Nueva Zelanda). Ambos casos no comprometen su uso con una metodología en particular, por lo que son mucho más atractivas, en este sentido. (12)

La Weka (gallirallus australis) es un ave endémica de Nueva Zelanda. Esta gallinácea en peligro de extinción es famosa por su curiosidad y agresividad, da nombre a una extensa colección de algoritmos de máquinas de conocimiento implementados en Java, útiles para ser aplicados sobre datos mediante las interfaces que ofrece, o para embeberlos dentro de cualquier aplicación pues es independiente de la arquitectura, y funciona en cualquier plataforma sobre la que haya una máquina virtual Java disponible. Además WEKA contiene las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, agrupamiento, asociación y visualización. Está diseñado como una herramienta orientada a la extensibilidad por lo que añadir nuevas funcionalidades es una tarea sencilla. (32)

Interiorizando las características descritas de las aplicaciones más comunes ante un proceso de KDD, se opta por WEKA 3.6.3 en versión de desarrollo que se utilizará para el pre procesamiento de los datos, y al ser compatible con Pentaho y de ser necesario se puede modificar su código fuente, lo cual significa un gran apoyo a la investigación presente.

Herramienta auxiliar

NetBeans es un Entorno Integrado de Desarrollo o IDE (Integrated Development Environment, por sus siglas en inglés) libre. Desarrollado principalmente para el lenguaje de programación “Java” y cuenta además con un número importante de módulos para extenderlo. (33)

Razones de su uso

Aunque puede ser costoso aprender a interactuar con el IDE NetBeans, los beneficios superan las dificultades pues cuenta con características como:

- Simplifica alguna de las tareas que, principalmente en proyectos grandes, suelen ser tediosas.
- Asiste (parcialmente) en la escritura de código.
- Ayuda en la navegación de las clases predefinidas en la plataforma. (33)

Exactamente se utiliza la herramienta en su versión 8.0, como plataforma para integrar *Postgres* con W*eka*, y permite modificar algunas funcionalidades de esta última.

Después del análisis anterior, se concluye que:

1) La minería de datos sintetiza una buena práctica para deducir conocimiento basado en datos, y actualmente con la cantidad de sistemas dedicados a la gestión de información, sin duda es una tecnología a tener en cuenta.
2) Se comprende que los AD no son necesarios para la minería de datos, pero indudablemente, sus aplicaciones y rendimiento son potenciados con la utilización de un almacén de datos auxiliar que de soporte a la misma.
3) Para realizar parcialmente el proceso de descubrimiento de conocimiento en bases de datos a través de AD, es necesario la selección de dos metodologías que guíen el proceso de desarrollo del MD y el proceso de KDD. La "*Metodología de Desarrollo de Proyectos de Almacenes de Datos*" y "CRISP-DM", fueron las seleccionadas para la construcción del almacén de datos y administrar el proceso parcial de KDD respectivamente. Donde se utilizan otras herramientas factibles para la investigación, tales son los casos de "*WEKA*" para la preparación de datos, "*Pentaho BI*" para soluciones de negocio, principalmente las de análisis e integración de datos.

Capítulo 2: Diseño e implementación del mercado de datos.

En el presente capítulo se abarca todo el proceso de creación del mercado de datos, guiado por la "*Metodología de Desarrollo para Proyectos de AD*". El ciclo de vida de la misma se divide en siete fases. En el presente capítulo los subepígrafes del mismo están acorde a las fases de la metodología seleccionada, teniendo en cuenta hasta cuál de las fases se llegará para la implementación del MD según lo descrito en el capítulo anterior; garantizando que se tengan en cuenta todos los detalles requeridos para el correcto diseño del mercado de datos. Los subsistemas de almacenamiento e integración se desarrollarán en los subepígrafe 2.4 y 2.5 respectivamente, pues se desarrollan en la fase de Diseño e implementación y se sustituirá el subepígrafe correspondiente a Diseño e implementación por los anteriormente descritos.

2.1 Estudio preliminar del negocio

2.1.1 Diagnóstico del negocio, datos e infraestructura tecnológica

El Repositorio Institucional de la UCI es un servicio habilitado para todo profesional del Centro, cuyo objetivo principal, es almacenar todo el contenido científico generado por especialistas, estudiantes y profesores. Parte importante de este servicio, es garantizar un acceso expedito para consultar los mismos.

La dirección del Centro de Informatización Científico Técnica (CICT) apoya la gestión del repositorio desde la biblioteca de la UCI, aprovechando así, la infraestructura tecnológica que posee la misma, teniendo en cuenta la disponibilidad de este importante servicio. También domina la idea, de cuan beneficioso pueden llegar a ser los resultados estadísticos e investigativos derivados de algún estudio realizado al repositorio de datos, incluso para la propia Dirección de Investigación UCI.

Actualmente estos datos almacenados pueden significar una matriz importante de argumentos contundentes basados en tendencias y comportamientos, no se exploran, ni son analizados. Motivos por lo cual no se tiene en cuenta en el proceso de toma de decisiones por la dirección del Centro de Informatización Científico Técnica (CICT).

Tanto el CICT como la Dirección de Investigación UCI, se interesan por descubrir patrones de conocimiento que ayuden a la toma de decisiones y diseño de estrategias. Para alcanzar dichos beneficios, se hace necesario aplicar al repositorio, técnicas de minería de datos.

2.2 Requisitos

2.2.1 Requisitos de Información

La información necesaria para satisfacer las necesidades del cliente, son los llamados requisitos de información. Aunque no sea objetivo de la investigación presente realizar determinado tipo de

reporte, atendiendo a las necesidades para la construcción del MD y al tipo de conocimiento que se desea extraer; se asume que los requisitos informacionales resultarán solo los datos relevantes que aporten información válida, novedosa y relevante a la minería de datos, independientemente de la fuente que provenga, en este caso el repositorio interno o servicios de Protocolo Ligero de Acceso a Directorios (LDAP, por sus siglas en inglés).

Parte de los requisitos de información fueron recopilados a través de la encuesta realizada a personas que interactúan con el Repositorio Institucional y la gestión de los datos que se almacenan en el mismo. Otros requisitos de información fueron seleccionados de acuerdo a la intención del conocimiento que se espera deducir de los datos. La descripción de los mismos se puede apreciar en la Tabla 1.

Tabla 1: Descripción de los requisitos de información.

No.	Atributos	Descripción
1	palabras_clave	Palabras que mejor identifican el contenido de la publicación.
2	resumen	Aborda los puntos centrales de la publicación.
3	serie_numeroReporte	Cuando el ítem forma parte de una serie numerada (reportes técnicos o documentos de trabajo) entonces el nombre de esa Serie es el que va en el casillero correspondiente al número asignado.
4	identificador	Código único que identifica el artículo en algún sistema determinado.
5	editorial	Nombre del editor del artículo en cuestión, si ha sido previamente publicado o distribuido.
6	idioma	Idioma principal del artículo en cuestión.
7	autor (género, región)	Puede ser una persona, institución o servicio responsable de la creación o contribución a la creación del artículo. De ser una persona, también son de interés el género y el lugar de procedencia
8	comunidad	El contenido del repositorio se organiza por comunidades, que contienen todos los trabajos de la producción científica de la universidad. Estas son: Tesis, UCIENCIA, Eventos y Revistas.
9	colección	Una comunidad puede tener un ilimitado número de colecciones en el repositorio. Entre las cuales esta: Trabajos de Diploma, Tesis de Maestría, Tesis de Doctorado,

		UCIENCIA 2012, Pedagogía 2013, JCE y UXI.
10	mes	Mes en que fue hecha la publicación.
11	anno	Año en que fue hecha la publicación.
12	trimestre	Trimestre al que pertenece la publicación
13	semestre	Semestre al que pertenece la publicación
14	área	Facultad o área de la Universidad a la que pertenece el autor de la publicación.

2.2.2 Requisitos funcionales

En la Tabla 2 se muestra la descripción detallada de los requisitos funcionales:

Tabla 2: Descripción de los requisitos funcionales del sistema.

No.	Requisito funcional	Descripción
4	Extraer datos	Permite extraer los datos del repositorio institucional y del servicio LDAP.
5	Transformar datos	Permite realizar la transformación de los datos del repositorio institucional.
6	Cargar datos	Permite realizar la carga de los datos del repositorio institucional hacia el mercado de datos a desarrollar.
7	Extraer datos para vista minable	Permite extraer los datos hacia una tabla auxiliar con el objetivo de que sean procesados en la misma para crear una vista minable.

2.2.3 Requisitos no funcionales

A continuación en la Tabla 3, se detallan algunos de los requisitos no funcionales

Tabla 3: Descripción de los requisitos no funcionales del sistema.

No.	Clasificación	Requisito no funcional	Descripción
1	Accesibilidad	Garantizar la disponibilidad.	El sistema debe estar disponible para los usuarios en cualquier momento.
2	Usabilidad	Garantizar la fácil integración de datos al sistema.	El sistema debe gestionar cada vez que se requiera, la integración de los datos de manera sencilla y ágil.
		Lograr la gestión de datos de forma sencilla.	El sistema debe ser capaz de realizar el procesamiento de los datos de

			manera fluida y ágil.
3	Confiabilidad	Garantizar la persistencia de la información.	Se debe realizar un respaldo total de los datos del mercado de datos anualmente.
4	Soporte	Garantizar que los datos sean homogéneos.	Lograr que los elementos definidos en el mercado tengan una nomenclatura homogénea
5	Software	Proporcionar características mínimas de software a las PC clientes	Requerimientos mínimos de software para las PC clientes: • Sistema operativo (Debian 6, Ubuntu Server 12.04, Microsoft Windows XP o superior) • PostgreSQL 9.1 • Máquina virtual de Java 1.8 • Navegador web Firefox versión 15
		Proporcionar características mínimas de software a los servidores.	Requerimientos mínimos de software para los Servidores: • Sistema operativo (Debian 6, Ubuntu 12.04 o superior) • PostgreSQL 9.1 o superior • Máquina virtual de Java 1.8 o superior • Navegador web Firefox versión 15 o superior
6	Hardware	Proporcionar características mínimas de hardware a las PC clientes	Requerimientos mínimos de hardware para las PC clientes: • 1 GB RAM DDR 2 o superior • 1 Procesador 2.66 GHz o superior
		Proporcionar características mínimas de hardware a los servidores.	Requerimientos mínimos de hardware para los Servidores:

			• 2 GB RAM o superior • 1 Procesador Dual Core 2.9 GHz o superior • 250 GB de disco o superior

2.2.4 Modelo de casos de uso del sistema

El modelo de caso de uso del sistema refleja gráficamente las metas y funciones que persigue el negocio. Se usa como una entrada esencial para identificar roles y entregables en la organización. (34)

Actores del sistema

Un actor del sistema es aquel que interactúa con él, ya sea interno o externo a la organización. (28) Se definieron dos actores, con el fin de separar responsabilidades y delimitar correctamente la función de cada actor. Los mismos se describen a continuación:

- **Administrador de ETL**: es el responsable del proceso ETL que requiere el sistema con respecto a la BD fuente. Es decir la extracción, transformación y carga de los datos.
- **Analista de Datos**: actor que controla los análisis a efectuar, es decir determina el punto de vista del análisis, que parámetros son los más convenientes para ello y la granularidad de la dimensión tiempo más adecuada.

Descripción Casos de uso del sistema

Un caso del uso del sistema representa un conjunto de tareas relacionadas que generan un resultado de valor para los actores del sistema. (28) Para la descripción de los casos de uso del sistema se siguió el modelo de tabla descrito por la metodología RUP (Proceso Unificado de Rational), porque en la metodología seleccionada no se describe una tabla para describir los casos de uso del sistema y porque los autores tienen experiencia previa con el modelo propuesto en RUP. En las Tablas 4 y 5, se describen los casos de uso del sistema:

Tabla 4: Descripción del Caso de Uso: Extraer, Transformar y Cargar datos.

Objetivo	Extraer, Transformar y Cargar datos del sistema fuente para el MD.
Actores	Administrador ETL
Resumen	El caso de uso inicia cuando el actor requiere de una nueva integración de los datos del Repositorio Institucional, para posteriormente cargarlos en el MD.
Complejidad	Alta.
Prioridad	Primario

Precondiciones	-
Postcondiciones	El MD posee los datos disponibles para el análisis.
Flujo de eventos	
Flujo básico Extraer, Transformar y Cargar datos	
Actor	**Sistema**
1. Selecciona el job general, creado con el objetivo de ejecutar cada una de las transformaciones pertinentes para todas tablas del MD, con la secuencia requerida.	
2. Selecciona la opción "*Run this transformation or Job*" destinada a ejecutar el job general.	
	3. Muestra una ventana con diferentes opciones para la ejecución del job general.
4. Seleccionar el botón "*Launch*"	
	5. Inicia el job, durante la ejecución del mismo se aprecia el flujo de los datos por cada uno de las transformaciones presentes en el job.
	6. Termina el job general y todas las tablas del MD cuentan con los datos procesados del Repositorio Institucional.
	7. Termina el caso de uso.
Flujos alternos:	
4ª El actor selecciona el botón "*Cancel*"	
	4.1 Cierra la ventana con diferentes opciones para la ejecución del job.
	4.2 Regresa al paso 1 del flujo básico.
5ª El sistema detecta un error durante la ejecución del job general en alguna de sus transformaciones.	

		5.1 Detiene la ejecución y muestra la ventana "*Loggin*" con los errores existentes.
		5.2 Regresa al paso 1 del flujo básico.

Tabla 5: Descripción del Caso de Uso: Extraer datos para vista minable.

Objetivo	Extraer datos del MD para crear la vista minable	
Actores	Analista de datos	
Resumen	El caso de uso inicia cuando el actor desea extraer los datos del MD para la creación de la Vista Minable, la cual reúne todos los datos a analizar para la posterior realización del agrupamiento como tarea de minería de datos.	
Complejidad	Baja	
Prioridad	Alta	
Precondiciones	Deben existir datos disponibles en el MD	
Postcondiciones	Se obtiene la Vista Minable con los datos necesarios para posteriormente realizar el agrupamiento.	
Flujo de eventos		
Flujo básico Generar vista minable		
Actor		**Sistema**
1. Selecciona la transformación previamente elaborada y asociada al llenado de la Vista Minable presente en la suite Pentaho BI.		
2. Selecciona la opción "*Run this transformation or Job*" destinada a ejecutar la transformación previamente seleccionada.		
		3. Muestra una ventana con diferentes opciones para la ejecución de la transformación.
4. Seleccionar el botón "*Launch*".		
		5. Inicia la transformación, durante la ejecución de la misma se aprecia el flujo de los datos por cada uno de los componentes presentes en la misma.

		6. Termina la transformación e inserta los datos en la tabla destino.
		7. Termina el caso de uso.
Flujos alternos:		
4ª El actor selecciona el botón "*Cancel*"		
		4.3 Cierra la ventana con diferentes opciones para la ejecución de la transformación.
		4.4 Regresa al paso 1 del flujo básico.
5ª El sistema detecta un error en la ejecución de la transformación.		
		5.1 Detiene la ejecución y muestra la ventana "*Loggin*" con los errores existentes.
		5.2 Regresa al paso 1 del flujo básico.

2.2.5 Diagrama de casos de uso del sistema

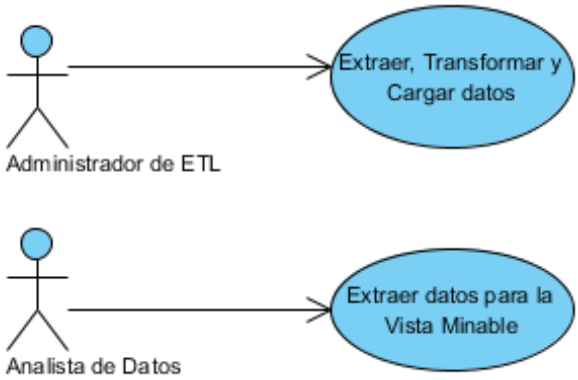

Figura 8: Diagrama de caso de uso

2.3 Arquitectura del sistema

La arquitectura del sistema se basa en los principios que referencia la metodología de la "base para la especificación de arquitecturas de software". Donde de las nueve vistas de la arquitectura según Lazo y Piñeiro (35), se entiende por **vista de sistema**, la perspectiva donde se definen los subsistemas, componentes y paquetes que conforman la solución.

La arquitectura del MD tal como se muestra en la Figura 9, cuenta con fuentes de datos internas y externas (repositorio institucional UCI y los servicios LDAP). Ambas fuentes nutren al subsistema de integración, quien a través de un adecuado subsistema de integración, implementado con la herramienta de Pentaho para procesos ETL "Kettle (Pentaho Data Integration)"; se encarga de que los datos sean insertados correctamente en el MD. Este último, es un subsistema de

almacenamiento administrado por el sistema gestor de base de datos PostgreSQL 9.3 y su herramienta PG Admin III e su versión 1.18:

Figura 9: Arquitectura del Mercado de Datos.

2.4 Subsistema de almacenamiento

Para materializar el MD es necesario contar con estándares de codificación, sin perder de vista la simpleza y sencillez del mismo, además también se debe diseñar una estructura física de almacenamiento. Ambos elementos tributan a una correcta y mejor práctica durante la implementación del MD.

2.4.1 Diseño del subsistema de almacenamiento

Después de un detenido estudio con el objetivo de establecer un primer contacto con el problema, comprender de los datos del repositorio, cuál es su arquitectura, determinar la calidad de los mismos y las relaciones existentes, se procede a extraer los datos necesarios.

Esta primera recolección de datos requeridos, parte de la necesidad de información para la toma decisiones, agregando la información relevante extraída del estudio y de la encuesta realizada a los usuarios para la generación de la vista minable.

Fueron creadas dos tablas de hechos, cuatro tablas dimensiones y tres tablas auxiliares con el fin de obtener la granularidad requerida de algunos datos. A continuación, en las Tablas 6, 7 y 8 se brindan detalles de las tablas identificadas:

Tabla 6: Descripción de las tablas hechos del almacén.

No.	Hechos	Descripción
1	hecho_publicación	Consiste en agrupar todos los detalles de una publicación, donde para materializarla correctamente se deben tener en

		cuenta los datos del artículo contemplados en las tablas: dim_ Instancia, dim_ Tiempo, dim_ Lugar_Area y dim_ Lugar_Centro, descritas posteriormente.
2	hecho_bibliografía	Posee los datos de las bibliografías consultadas en las publicaciones, de acuerdo a la norma "*ISO 690*" entre los que están: título, autor y anno, además tiene en cuenta datos como: los artículos a los que pertenece, y el tiempo en que fueron referenciadas.

Tabla 7 Descripción de las tablas dimensiones del mercado.

No.	Dimensiones	Descripción
1	dim_instancia	Agrupa todos los detalles desde el punto de vista bibliográfico y científico, entre los que se pueden incluir: indicador del autor, palabras claves, resumen, detalles si ha sido publicado anteriormente el artículo, comunidad y colección que corresponde.
2	dim_tiempo	Define la línea del tiempo mensual especifica de la publicación, con características como: mes, anno.
3	dim_lugar_area	Define el área a la que pertenece el autor del artículo.
4	dim_autor	Agrupa todos los detalles propios de los autores de publicaciones, entre los cuales se pueden incluir: nombre, región y género.
5	dim_autor_x_instancia	Posee la relación de autores con sus publicaciones
6	dim_palabras_claves	Concentra todas los palabras claves de los artículos publicados
7	dim_autor_bibliográfico	Agrupa todos los detalles propios de los autores de las bibliografías citadas en las diferentes publicaciones existentes en el repositorio, entre los cuales se pueden incluir: nombre, apellidos, si es corporativo o no.

Tabla 8: Descripción de la tabla de metadatos del mercado.

No.	***Tabla de Metadatos***	***Descripción***
1	met_gestion_carga_historica	Agrupa los detalles propios de los datos que han sido cargados históricamente en el mercado, es decir recoge de

		cada carga realizada a través del sistema ETL: fecha y hora, numero ip y el nombre de la transformación ejecutada.

El hecho bibliografía no fue llenado pues no se pudo acceder a los datos por falta de autorización para el acceso a los mismos, no obstante, se mantiene en el diseño para que cuando se logre el autorizo pertinente se carguen los datos al MD para que los datos para generar en la vista minable sean mayores y sea mejor el proceso de extracción de conocimiento a realizar sobre la vista minable en el futuro.

2.4.2 Matriz bus o de trazabilidad

La matriz bus o de trazabilidad muestra de manera legible la relación existente entre las tablas hechos y dimensiones, permitiendo la visualización del impacto que tendría la modificación de algunas de las tablas durante el desarrollo del sistema, como describe la Tabla 9.

Tabla 9: Matriz bus del Mercado de Datos.

Dimensión / Hecho	*hecho_publicacion*	*hecho_bibliografia*
dim_instancia	X	X
dim_tiempo	X	X
dim_lugar_area	X	
dim_autor		
dim_autor_x_instancia		
dim_palabras_claves		
dim_autor_bibliográfico		X

2.4.3 Modelo de datos del mercado de datos

El modelo utilizado en la propuesta de solución se basa en el esquema constelación o copo de estrella. Como eventos fundamentales resaltan entonces: las publicaciones existentes en el repositorio institucional, sus bibliografías y los registros de navegación conocidos como "*logs*". Dichos eventos serán las tablas de hechos del MD y sus características las tablas dimensiones. Adicionando también la tabla de metadatos encargada de almacenar los registros de cargas históricas, y la tabla dedicada directamente a la tarea de minería de datos, tal como se muestra en la Figura 10:

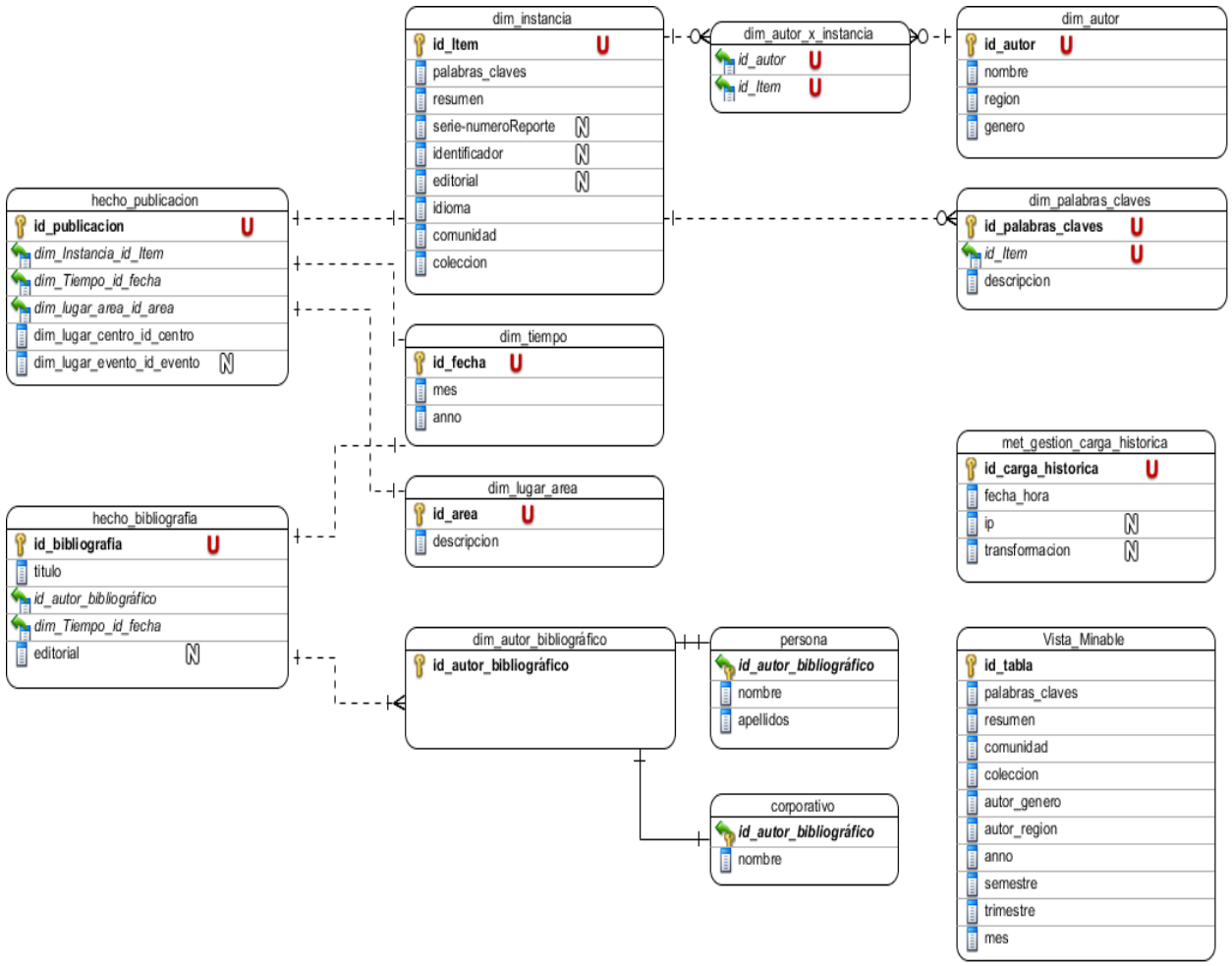

Figura 10 Modelo de Datos.

2.4.4 Estándares de Codificación

Tomando como base lo planteado por la especialista Marisel Santana en "*Estándares de codificación*" (36), **los estándares de codificación** se utilizan para lograr un entendimiento de las estructuras utilizadas durante el desarrollo de una aplicación. La Tabla 10 muestra los detalles de la codificación realizada:

Tabla 10: Descripción de la codificación empleada.

Tipo de objeto	Función	Nomenclatura	Descripción
Tablas	Dimensiones	dim_[nombre]	Tablas dimensiones
	Hechos	hecho_[nombre]	Tablas hechos
	Metadatos	met_[nombre tabla]	Tablas de datos históricos
Contrains	Llaves primarias y foráneas	id_[nombre tabla]	Identificadores de cada tabla

2.4.5 Implementación de estructura física de almacenamiento

Una de las formas más eficientes de implementar un mercado de datos multidimensional mediante bases de datos relacionales se basa en ignorar casi completamente, la estructura de los datos en las fuentes de origen. (1) Habitualmente se propone utilizar al menos uno de los tres esquemas descritos anteriormente (Estrella, Copo de Nieve y Constelación), para este tipo de soluciones.

El modelo de datos propuesto en la fase anterior toma como referencia el esquema constelación o copo de estrella. Entonces el patrón a seguir para la implementación será el mismo que en el diseño del mercado, obteniendo entonces como resultado: un MD que cuenta con trece tablas, de las cuales dos son hechos, además cuenta con las clásicas dimensiones de *tiempo* y *lugar*, añadiendo la dimensión *instancia*, relacionada con las también dimensiones dim_autor*, dim_palabras_claves* y *dim_autor_x_instancia* (esta última como resultado al tipo de relación entre *autor* e *instancia*). Por otra parte se tiene en cuenta la tabla de *metadatos* denominada *gestión_carga_historica.*

Para el esquema constelación, se deberán especificar las jerarquías que existirán dentro de cada tabla de dimensión, teniendo siempre presente cuál es el objetivo de las mismas. Una jerarquía representa una relación lógica entre dos o más atributos dentro de una misma dimensión. La principal ventaja de manejar jerarquías, reside en poder analizar los datos desde su nivel más general al más detallado y viceversa. (37) Como se muestra en la Figura 11, existen ciertas jerarquías entre los atributos existentes en las dimensiones:

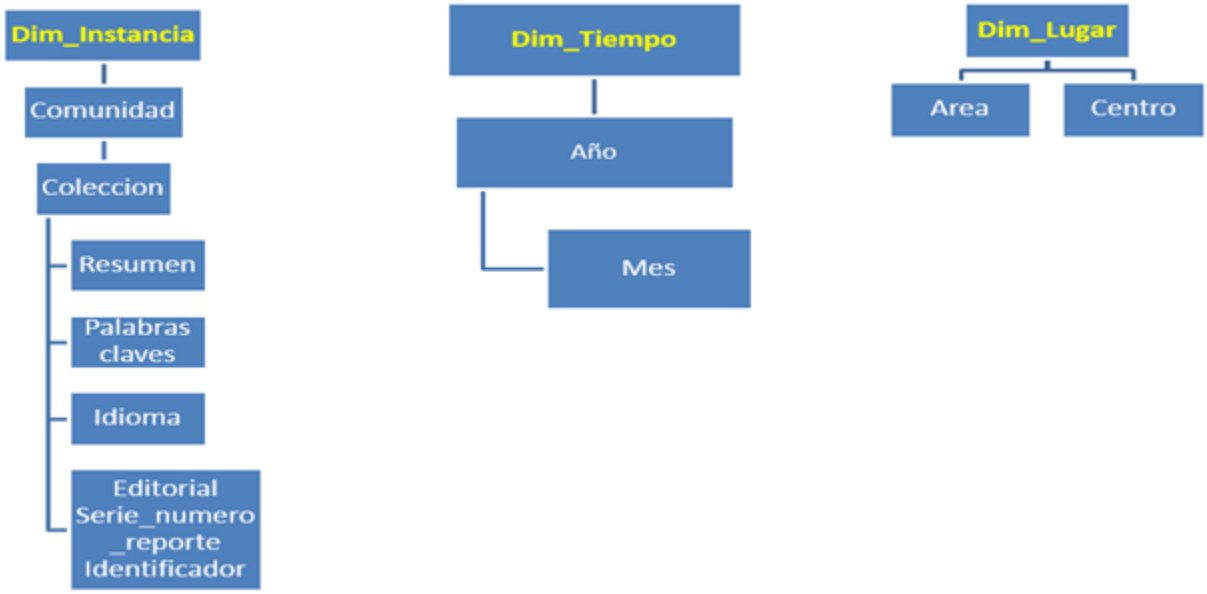

Figura 11 Jerarquía de las tablas dimensiones.

2.5 Subsistema de integración

La integración de los datos, involucra la creación de nuevas estructuras, a partir de los datos seleccionados. Por ejemplo, generación de nuevos campos a partir de otros existentes, creación de

nuevos registros, fusión de tablas, campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen. (23)

2.5.1 Diseño del subsistema de integración

Durante las fases iniciales "*Estudio preliminar del negocio*" y "*Levantamiento de requisitos*", se realizó un profundo estudio de los datos presentes en la fuente (Repositorio Institucional UCI), el cual fue implementado con "*DSpace*", software de código abierto diseñado para gestionar repositorios de ficheros (textuales, audio, vídeo, etc.) facilitando su depósito y organización.

La detección de valores atípicos, erróneos, o inválidos se puede realizar de maneras muy diversas dependiendo del formato, origen y posibles valores del campo.

A pesar de la garantía que brinda DSpace, antes de ser cargados se debe realizar la limpieza de datos para darle tratamiento a los valores nulos y codificarlos, en búsqueda de elevar la calidad de los mismos. Esta importante tarea se realiza mediante las seis reglas de extracción siguientes:

Reglas ETL

- Las publicaciones que previamente no han sido publicados sus artículos, el campo "editorial" se debe renombrar con: No publicado.
- Las publicaciones que no sean parte de una serie numerada tal como reportes técnicos o documentos de trabajo, el campo "serie_numero_reporte" se debe renombrar con: No posee.
- Las publicaciones que no tengan un número o código único que identifique su artículo en algún sistema específico, el campo "identificador" se debe renombrar con: No posee.
- Existen publicaciones que no tienen resumen, por lo que en el campo "resumen" se debe renombrar con: No posee.
- Existen publicaciones que no tienen idioma definido, por lo que en el campo "idioma" se debe renombrar con el idioma por defecto: es.
- Existen publicaciones que no tienen palabras clave, por lo que el campo con el mismo nombre será llenando con la frase: No posee.

2.5.2 Registros de sistemas fuentes

El Registro de Sistemas Fuentes detalla la localización de los datos (sistemas fuentes) y las técnicas que se deben emplear para su recolección. Seguidamente la Tabla 12 detalla los atributos requeridos por la dimensión instancia.

Tabla 11: Lista de atributos (Fuente) dim_Instancia.

Nombre de	***Sistema***	***Esquema origen***	***Tabla origen***	***Nombre del campo***	***Tipo de Dato***	***Reglas ETL***	***Comentarios***

columna	***origen***			***origen***	***origen***		
id_item	***Proceso ETL***						Llave primaria de la tabla
identificador		Public	metadatavalue	text_value	text		Verificar que el campo metadata_field_id tenga valor 22.
resumen		Public	metadatavalue	text_value	text		Verificar que el campo metadata_field_id tenga valor 27.
editorial		Public	metadatavalue	text_value	text		Verificar que el campo metadata_field_id tenga valor 39.
idioma		Public	metadatavalue	text_value	text		Verificar que el campo metadata_field_id tenga valor 38.
coleccion		Public	collection	name	text		
comunidad		Public	community	name	text		

fuente_id_item		Public	metadatavalue	item_id	serial		
serie_no_reporte		Public	metadatavalue	text_value	text		Verificar que el campo metadata_field_id tenga valor 43.

Nota 2: Los registros restantes de las fuentes, se encuentra en los anexos. (A.2)

2.5.3 Registros de atributos requeridos por cada tabla del mercado

Mediante la siguiente lista de los datos adquiridos o atributos de la tabla "*hecho publicación*" mostrados en la Tabla 13, se obtiene un informe con la ficha técnica de los mismos.

Tabla 12: Lista de atributos II (Tarjeta): dim_Instancia

Columna	Descripción	Tipo de Dato	Llave	FK de Tabla	Campo nulo	Valor x defecto	Ejemplo de Valores
id_Item	Id de la tabla dim_Instancia	int	PM		No		
palabras_claves	palabras que mejor identifican el contenido de la publicación	varchar			No		ciencia, mercado de datos, minería de datos
resumen	Aborda los puntos centrales del contenido	varchar			No		
serie_numero Reporte	Cuando el ítem forma parte de una serie numerada (reportes técnicos o	int			Si		1,2,3

	documentos de trabajo) entonces el nombre de esa Serie es el que va en el casillero correspondiente al número asignado.						
identificador	Código único que identifica el ítem en algún sistema	varchar			Si		A123.
editorial	Ingrese el nombre del editor del ítem	varchar			Si		Editorial Félix Varela
idioma	Idioma principal del ítem	varchar			No	esp	esp,ing
autor_id_autor	Id de la tabla autor	int	FK	autor	No		1,2,3
comunidad_id_ comunidad	Id de la tabla comunidad	int	FK	comunidad	No		1,2,3

Nota 3: Los registros restantes de las atributos, se encuentra en los anexos. (A.3)

2.5.4 Proceso general de Integración

Después de concluido el análisis de los datos de la fuente, se procede a realizar el proceso de integración de datos. En la Figura 12, se describe el proceso general de integración del MD, en el cual se define el flujo a seguir de cada proceso de integración mediante transformaciones, finalizando con la carga de los datos en el MD y proceder al análisis:

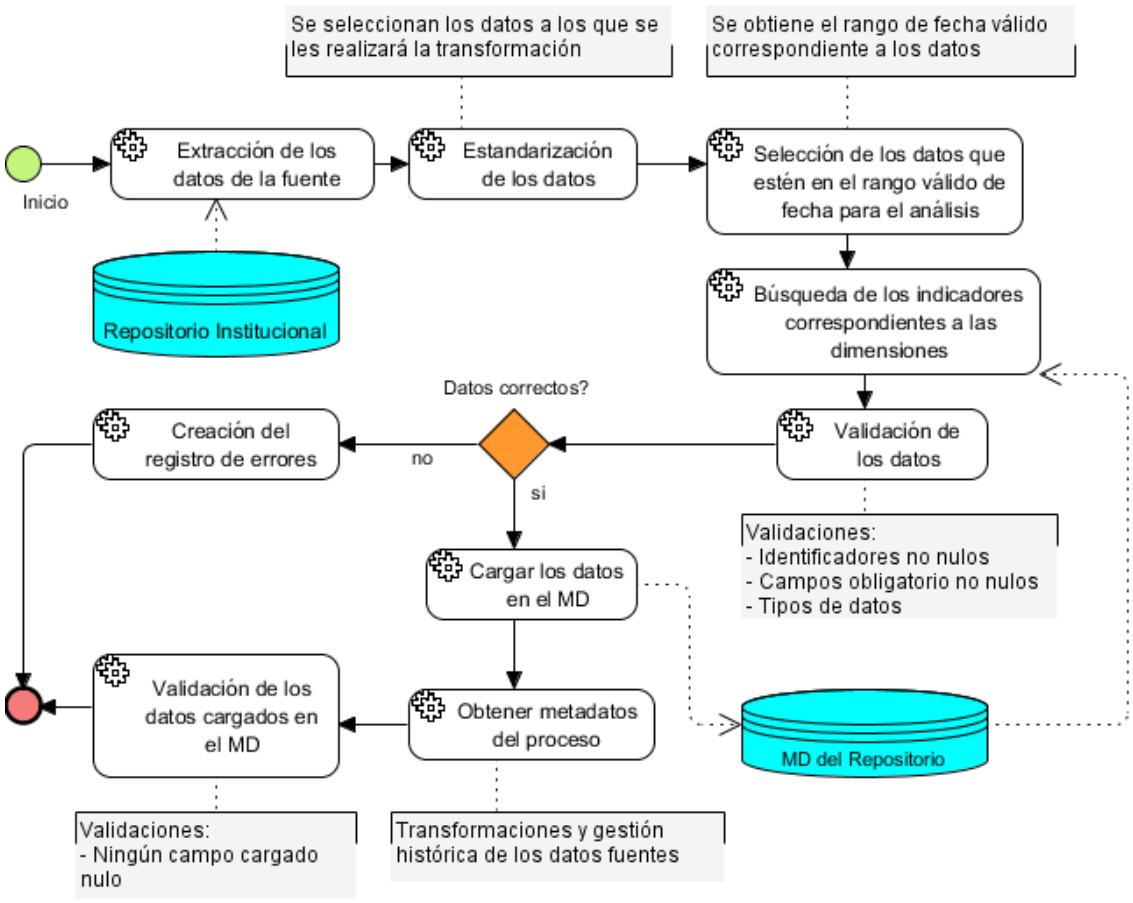

Figura 12 Proceso general de integración del mercado de datos

2.5.5 Implementación de las ETL

En la presente fase, se procede a ejecutar el sistema ETL. Lo que constituye; la extracción de los datos de las fuentes correspondientes, los mismos son almacenados temporalmente para su transformación y posteriormente se realiza la limpieza de los datos mediante las reglas ETL diseñadas previamente. Finalmente, se cargan los datos transformados en el MD. A continuación en las Figuras 13 y 14, la ejemplificación de las transformaciones ejecutadas para hecho publicación y la dimensión tiempo:

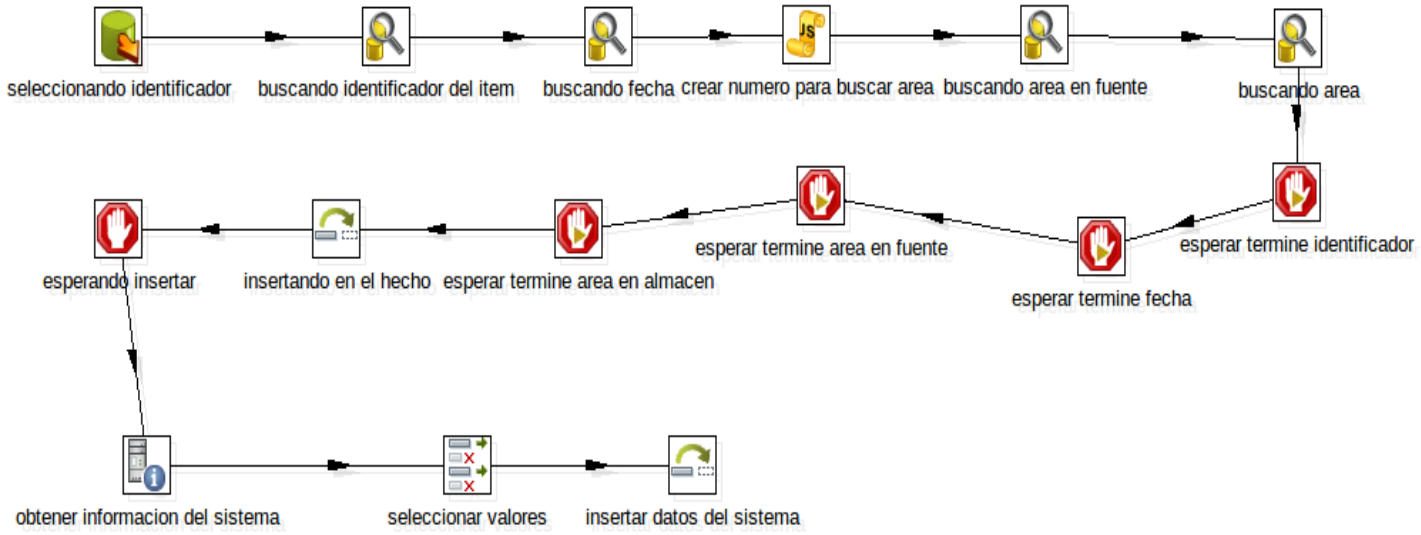

Figura 13: Transformación de la tabla "hecho_publicacion".

En la transformación mostrada:

Primero se buscan los identificadores de las publicaciones en la fuente (repositorio) para buscar con cual corresponde en el almacén. Seguido se busca el identificador correspondiente de cada instancia y su fecha en el almacén. Luego se crea un modificador de java script, con un dato con valor 72 para buscar el identificador de cada área correspondiente a cada publicación en la fuente y se compara con el almacenado en la dimensión: dim_lugar_area.

Los pasos de bloqueo garantizan el tiempo y el flujo necesario para que todos los datos se inserten correctamente en la tabla hecho. Posteriormente se recopilan todos los datos pertinentes de la transformación para agregarlos en la tabla hecho, junto a la ejecución del paso de bloqueo correspondiente.

De esta forma, una vez ejecutada la transformación los datos de la tabla "hecho_publicacion" quedarán completamente integrados.

Transformación de la dimensión "*tiempo*":

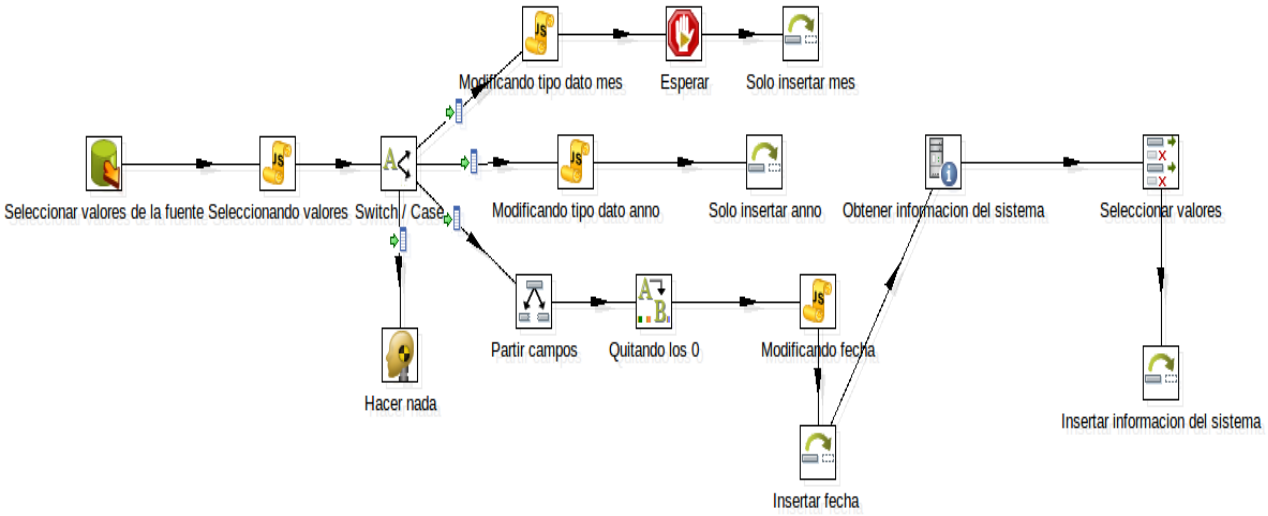

Figura 14: Diseño de la transformación tiempo.

Para la transformación de la dimensión tiempo, se parte de la condición de que los valores del campo fecha se pueden encontrar en diferentes formatos en la fuente de datos, estructurados de la siguiente forma: solamente el mes, solamente el año o una fecha que contiene año, mes, día y hora. Para poder tratar los mismos y unificarlos de una sola forma, se introduce en la transformación un switch/case, que realiza una acción diferente para cada caso presente:

1. Se seleccionan los valores de la fuente que corresponda con el dato fecha.
2. En el javascript se analiza la forma en que pueden aparecer los datos según las descritas anteriormente y se asigna un valor a una variable que será analizada por el switch/case, en el siguiente caso, determinará cual tratamiento se realizará con los datos.
3. En el switch/case se decide que tratamiento se hará con los datos de acorde a como se encuentren.
   a) En caso de que solo sea el mes:
      i. Se crean las variables trimestre y semestre para introducirlas en el almacén, teniendo en cuenta el mes con el que se trabaja y se modifica el valor para introducirlo de acuerdo al esperado en el MD.
      ii. El paso stop 1 esperará a que se inserten todos los años en el MD para luego permitir que se inserten los valores del mes, trimestre y semestre.
      iii. Se insertan los valores de mes, trimestre y semestre en el MD.
   b) En caso de que solo sea el año:
      i. Se modifica el valor para introducirlo de acuerdo al esperado en el MD.
      ii. Se inserta en el MD el valor del año.
   c) En caso de que sea una fecha compuesta por año, mes, día y hora:

i. Se separan los campos por el delimitador para que queden en tres variables los datos año, mes y día-hora, como no es necesario el dato del día y la hora se deja en una sola variable con la que no se realizará ningún trabajo o transformación.
ii. En el caso del dato del mes puede traer un cero delante por lo que se eliminan para insertarlo en el MD.
iii. Se crean las variables trimestre y semestre para introducirlas en el almacén, teniendo en cuenta el mes con el que se trabaja y se modifica el valor del año y el mes para introducirlo de acuerdo al esperado en el MD.
iv. Se insertan los valores de año, mes, semestre y trimestre en el MD.
v. Se recogen los datos referentes a la ejecución de la transformación, que son: ip de la pc desde donde se realiza la transformación, fecha en la que se realiza y el nombre de la transformación.
vi. Se seleccionan los valores a introducir en el MD del paso anterior.
vii. Se insertan en el MD los valores referentes a la ejecución de la transformación.

d) Cuando no es ninguno de los casos descritos anteriormente:
   i. En este paso no se realiza ninguna acción, se introduce para que no existan problemas durante la ejecución de la transformación y evitar la inserción de datos erróneos en el mercado.

Nota 4: Los diseños de las transformaciones de las tablas: dim_instancia, dim_autor, dim_autor_x_instancia y dim_palabras_claves se encuentran en los anexos (A.4).

Job para el mercado de datos

Un trabajo o job, es un conjunto de tareas donde se utilizan pasos específicos que son distintos a los disponibles en las transformaciones. Mediante los trabajos se define el horario y frecuencia de la carga, así como el orden en que van a ser ejecutadas las transformaciones para poder realizar la carga de los datos. (38)

Posteriormente en la Figura 15, se describe el trabajo general del cubo publicaciones, donde se observa que el orden de ejecución de las cargas comienza por dimensiones y tablas alternadamente, dependiendo de las relaciones de cada una de ellas, por ejemplo no se puede realizar la transformación instancia si los valores de los que ella depende del autor (transformación autor) no están en el mercado aún. Por último se procede a la carga del hecho "*publicaciones*":

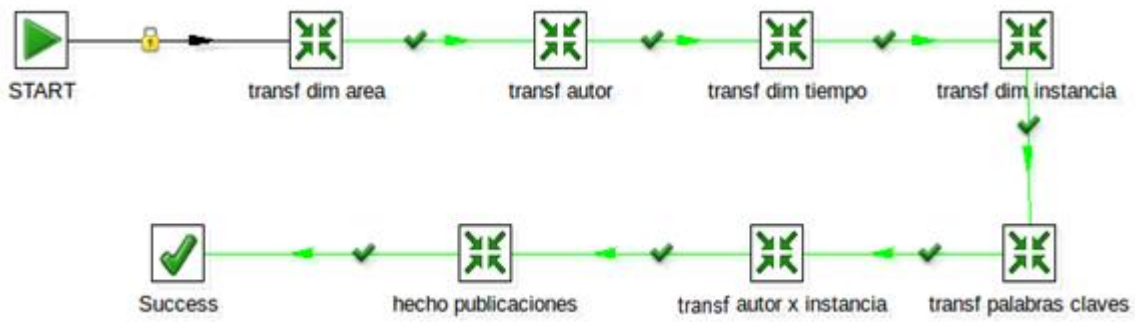

Figura 15 Trabajo o Job del cubo publicaciones.

Luego de la completa implementación del MD se puede concluir que:

1) La arquitectura del sistema está correctamente diseñada y enfocada al tipo de análisis que se pretende realizar.
2) El subsistema de almacenamiento está implementado a un 100%, estructurado por tres tipos de tablas, entre las que se encuentran tablas hechos, dimensiones y la de metadatos.
3) El subsistema de integración se realizó mediante el Job general, garantizando así el orden adecuado de las transformaciones y teniendo en cuenta las reglas ETL.
4) Todos los requisitos de información y los datos que requieren, están integrados y contemplados en el sistema.
5) El mercado de datos posee la información necesaria de acuerdo a los requisitos de información recopilados para la investigación y cuenta con funcionalidades que responden a los requisitos funcionales del mismo.
6) El mercado de datos posee todas las capacidades y requisitos para comenzar con la exploración y descripción de sus datos, pre procesarlos de ser necesario y generar la vista minable para futura realización de las tareas de minería de datos.

Capítulo 3: Generación de la Vista Minable.

En el presente capítulo se explica el proceso para generar la vista minable que se desea obtener como objetivo de la presente investigación; que pudiera describirse también como parte del proceso KDD. Precisamente todo el contexto que conlleva la aplicación de técnicas de minería de datos a un conjunto de datos específicos, los cuales requieren de una previa fase de preparación, en vista de presentar los datos de la manera más idónea para el descubrimiento de conocimiento. Este proceso está guiado por la metodología para proyectos de minería de datos "*CRISP-DM"*, por lo que cada uno de los epígrafes siguientes responde a las fases planteadas por la metodología.

3.1 Comprensión del negocio

Esta fase está completamente a fin, con la fase "*Estudio preliminar del negocio*" descrita en el capítulo anterior, la cual contempla un diagnóstico integral de la organización desde tres puntos de vistas: negocio, datos e infraestructura tecnológica.

3.2 Comprensión de los datos

La fase de comprensión de los datos, comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. (13)

3.2.1 Recopilación de los datos

Para esta etapa, al estar ya confeccionado el mercado de datos, solo se procede a determinar los datos que resulten relevantes para el análisis. Los mismos son descritos en la Tabla 13:

Tabla 13: Lista de datos relevantes para la vista minable.

No.	Atributos	Posibles análisis que puedan generar
1	palabras_claves	Análisis a partir de las palabras identificativas de las publicaciones, utilizando minería de texto.
2	resumen	Análisis a partir de los puntos centrales del contenido.
3	comunidad	Análisis en dependencia de la comunidad a la que pertenecen las publicaciones.
4	colección	Análisis en dependencia de la colección a la que pertenecen las publicaciones.
5	autor (género)	En caso de ser una persona, puede ser analizado su sexo.
6	autor (región)	En caso de ser una persona, puede ser analizado su sexo.
7	mes	Análisis de las publicaciones de algún mes determinado.
8	anno	Análisis de las publicaciones de algún año determinado.

9	semestre	Análisis de las publicaciones de algún semestre determinado.
10	trimestre	Análisis de las publicaciones de algún trimestre determinado.

3.2.2 Descripción y exploración de los datos

Después de adquiridos los datos iniciales, estos deben ser descritos y explorados. Este proceso involucra establecer volúmenes de datos (número de registros y campos por registro). Luego se procede a la aplicación de pruebas estadísticas básicas, que revelen propiedades en los datos recién adquiridos, se crean tablas de frecuencia y se construyen gráficos de distribución. (9)

A continuación en la Tabla 14 se muestra la descripción y distribución de los datos de interés para el análisis mediante la minería:

Tabla 14: Descripción y distribución de los datos necesarios para la minería de datos.

CANTIDAD DE PUBLICACIONES		6768	**CANTIDAD DE ATRIBUTOS**		**9**
9					
COLUMNA	**Total de Campos**	**Tipo de Dato**	**Vacíos modificados "No posee"**	**Erróneos**	**Nulos**
ID_TABLA	**6768**	**Serial**	**0**	0	0
RESUMEN	**6768**	varchar	989	0	0
COMUNIDAD	**6768**	Varchar	0	1	1
COLECCIÓN	**6768**	Varchar	0	1	1
PALABRAS CLAVES	**6768**	Varchar	2109	0	0
MES	**6768**	Varchar	0	2	2
ANNO	**6768**	Varchar	0	7	7
TRIMESTRE	**6768**	Varchar	0	2	2
SEMESTRE	**6768**	varchar	0	2	2

3.2.3 Calidad de los datos

Luego de una detallada exploración de los atributos de las 6 768 publicaciones registradas, se procede a determinar la calidad de las mismas. El número de valores nulos es prácticamente insignificante aunque requieren de tratamiento, los datos a estudiar se encuentran en un rango de tiempo de 13 años (2001-2014). También destaca que todos los atributos fueron convertidos al tipo de dato "*varchar*" pues la tarea de minería a la que se perfilan los datos solo admite atributos

nominales. Otro aspecto relevante es que las tesis publicadas que posean en los atributos: resumen (989) y palabras_claves (2 109) igual a “No poseen”, no se tendrán en cuenta, debido a que esto fue un tratamiento dado mediante las reglas ETL, para evitar integrar muchas publicaciones con datos vacíos. Como no aportan nada al análisis esta característica creada, las tesis con estas condiciones no se contemplarán en el mismo.

3.3 Preparación de los datos

En esta fase y una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se utilicen posteriormente. Incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. (9)

3.3.1 Construcción de los datos

En la mayoría de los análisis de minería de datos es conveniente realizar transformaciones sobre el conjunto de datos con el fin de mejorar la precisión de los modelos de aprendizaje que se desarrollarán posteriormente. (12)

Por ello, se concibe la creación de tres variables con el objetivo de obtener la granularidad adecuada para la dimensión tiempo, debido a que el análisis a realizar no tiene sentido ejecutarlo diariamente, y también se procede a agrupar el lugar de procedencia de los autores de cada publicación, ya que puede resultar de interés para la estimación de patrones.

A continuación se contempla la descripción de las mismas:

Variable trimestre: responde al trimestre en que fue hecha la publicación, se discretiza la variable trimestre para que contemple un rango de tres meses, y al existir 12 meses en el año, la misma cuenta con las siguientes etiquetas: primero (enero, febrero, marzo), segundo (abril, mayo, junio), tercero (julio, agosto, septiembre), cuarto (octubre, noviembre, diciembre).

Variable semestre: basado en la estructura del curso escolar, donde el mismo se divide en dos semestres de cinco meses cada uno, la variable se refiere a cual de ambos corresponde la publicación, contando con las siguientes etiquetas posibles: primero (septiembre - enero), segundo (febrero - junio).

Variable región: se agrupa la de procedencia de cada autor mediante las regiones: Oriente, Centro y Occidente, siendo solo estos los valores posibles.

3.3.2 Limpieza de los datos

Esta tarea complementa a la anterior, y es una de las que más tiempo y esfuerzo consume, debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos a objeto de prepararlos para la próxima fase. (9)

Pre procesamiento de los datos

Ante la presencia de datos erróneos, WEKA para enfrentar esta situación posee una amplia gama de filtros para el pre procesamiento de datos. Una característica clave en el porcentaje de datos con errores, es la presencia de campos vacíos. Para ello, se aplica el filtro "*Replace Missing Value*s", que se describe en el primer capítulo del documento de la investigación en curso.

Se muestra a continuación en las Figuras 16 y 17, la aplicación de este filtro a uno de los atributos afectados (semestre).

Figura 16: Atributo "semestre" con 1301 datos erróneos.

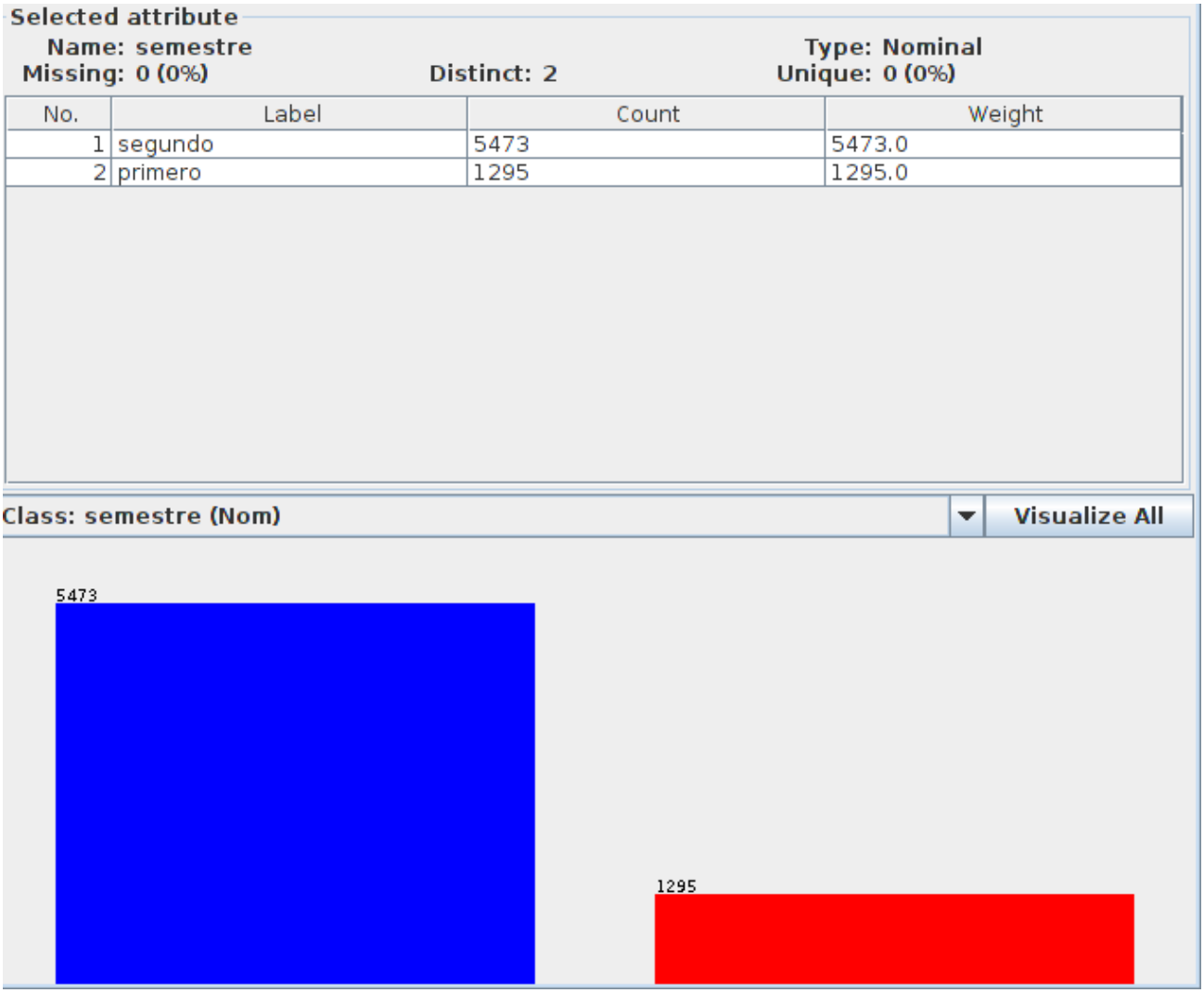

Figura 17: Atributo "semestre" después del filtro ReplaceMisingValues.

Como se observa, los datos erróneos de un 19% se redujeron a 0%, y el atributo *1 (segundo semestre)* de 4172 valores (el mayor inicialmente) incrementó a 5473. Por lo que ya el atributo "*semestre*", está listo para aplicarle las técnicas de minería de datos.

Nota 5: En los anexos se encuentran los restantes atributos a los que se les aplicó este filtro. (A.5)

3.3.3 Integración de datos

Terminados los pasos previos para la aplicación de la técnica minería de datos, donde se incluyen: análisis exploratorio, transformación y creación de datos necesarios, se obtuvo el conjunto de datos a estudiar que conformarán la vista minable. En la Tabla 15 se describe la misma:

Tabla 15: Vista Minable o Tabla de datos.

No.	Atributo	Tipo de dato	Número de valores	Ejemplo de Valores

1	palabras_claves	Nominal	2582	Inform.., metodo..
2	resumen	Nominal	5781	(El resumen inicial de cada tesis)
3	comunidad	Nominal	4	Tesis, Uciencia, Eventos, Revistas
4	colección	Nominal	7	Trabajos de Diploma, Tesis de maestría o doctorado
5	autor (género)	Nominal	2	M,F
6	autor (región)	Nominal	3	oriente, centro y occidente
8	mes	Nominal	12	1,2,3….12
9	trimestre	Nominal	3	primero, segundo, tercero y cuarto
10	semestre	Nominal	2	primero y segundo
11	anno	Nominal	10	2007, 2008, 2009… 2014

Luego de integrar los datos en la vista minable desde la herramienta WEKA se genera un fichero .arff que es el creado por la herramienta WEKA que permitirá luego cargarlo en la misma para poder aplicar en el futuro técnicas de minería de datos que permitirá continuar estudiando los datos existentes en el repositorio institucional.

De esta forma se cumple el objetivo general de la investigación que es crear una vista minable a partir del mercado de datos del Área de Informatización de la Biblioteca UCI, para contribuir al perfeccionamiento del proceso de toma de decisiones por los directivos del Centro de Informatización Científico Técnica y la Dirección de Investigación de la UCI.

Conclusiones

Con el cumplimiento del objetivo trazado en la siguiente investigación se arribó a las siguientes conclusiones:

1) El estudio del estado del arte y elaboración del marco teórico para describir la situación actual existente con respecto a la investigación llevada a cabo permitió el desarrollo de la solución informática.
2) El estudio de las herramientas y métodos a utilizar para la implementación del mercado de datos y la creación de la vista minable, permitió realizar la selección de los métodos: "Metodología de Desarrollo de Proyectos de Almacenes de Datos" para desarrollar el MD y "CRISP-DM" para llevar a cabo el proceso de KDD; así como la elección de las herramientas: PostgresSQL como sistema gestor de base de datos, "WEKA" para el pre procesado de datos y "Pentaho BI" para el análisis, integración y carga de los datos al MD.
3) El estudio y exploración de los datos almacenados en el repositorio y los relacionados con el acceso al mismo, permitió comprender la estructuración de los datos en el sistema fuente, con lo cual se pudo diseñar un apropiado sistema ETL (extracción, transformación y carga) y posteriormente la ejecución del mismo para la correcta integración de los datos al MD.
4) El estudio y selección de las técnicas de pre-procesamiento a emplear, permitió obtener un conjunto de datos a estudiar completamente heterogéneo, garantizando el estado óptimo de los mismos para su estudio.
5) Al aplicar las técnicas de pre-procesamiento sobre los datos del mercado de datos se obtuvo una vista minable que contribuirá al perfeccionamiento del proceso de toma de decisiones de los directivos del Centro de Informatización Científico Técnica y de la Dirección de Investigación de la universidad.

Recomendaciones

Se recomienda:

- Continuar con el proceso de extracción de conocimientos en bases de datos utilizando la vista minable resultante de la investigación desarrollada.
- Utilizar un diccionario (thesaurus en inglés) para reducir el universo de búsqueda existente de 2 582 palabras resultantes luego de aplicar el preprocesamiento para que se pueda descubrir patrones.
- Al aplicar la técnica de agrupamiento k-means, o alguna de sus variantes, sustituir la fórmula de distancia que se utiliza, la euclideana, por la similitud del coseno para mejorar el resultado final del agrupamiento.

Referencias Bibliográficas

1. **Hernández Orallo, José, Ramírez Quintana, José María y Ferri Ramírez, César**. *Introducción a la Minería de Datos.* Madrid : PEARSON EDUCACIÓN, 2004.
2. *Lineamientos de la política económica y social del Partido y la Revolución.* 2011.
3. **Hernández Sampieri, Roberto, Fernández-Collado, Carlos y Baptista Lucio, Pilar**. *Metodologia de la Investigacion.Cuarta Edicion.*
4. **Hernández León, Rolando Alfredo y Coello González, Sayda**. *El proceso de Investigación científica.* s.l. : Editorial Universitaria Cubana, 2011.
5. *Warehousing and OLAP Analysis of Bibliographic Data.* **Georgieva-Trifonova, Tsvetanka. Veliko Tarnovo**, Bulgaria : s.n., 2011.
6. *DBPubs: Multidimensional Exploration of Database Publications.* **Baid, Akanksha**, y otros.
7. **Brown La O., Reina y Mesa Prada, Luis Eduardo** . *Desarrollo de un mercado de datos para el análisis de datos bibliográficos.* La Habana, Cuba : s.n., 2013.
8. *Practical Machine Leraning Tools and Techniques with Java Implementations.* **Witten Clark, P y Frank Boswell, R**. s.l. : Morgan Kaufmann Publishers, 2000.
9. **Gallardo, Jose Alberto Arencibia**. *Metodología para el Desarrollo de Proyectos en minería de datos CRISP-DM. Sistemas del Conocimiento.EPB603.*
10. **H. Witten, Ian, Frank, Eibe y A. Hall, Mark**. *Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edition.* s.l. : Morgan Kaufmann Publishers , 2005.
11. **Pascual, D., Pla, F. y Sánchez**, S. *Algoritmos de agrupamiento.* Santiago de Cuba,Castellon : s.n.
12. **Boada, Lic. Darian Horacio Grass**. *Tesis de Maestria: Minería de Datos aplicada a los registros de navegación por Internet en la Universidad de Ciencias Informáticas.* La Habana, Cuba : s.n., 2011.
13. **Moine, Ing. Juan Miguel , Haedo** , Dra. Ana Silvia y Gordillo , Dra. Silvia . *Estudio comparativo de metodologías para minería de datos.* Buenos Aires, Argentina : s.n., 2011.
14. **Inmon, William H**. *Building the Data Warehouse.* s.l. : Indianapolis : Wiley Publishing, 2005.
15. **Inmon, William H**. *Building the Data Warehouse. Fourth Edition.* s.l. : Wiley Publishing, 2005.
16. *Revista Ingeniería Informática: revista electrónica del DIICC.* **Wolff, Carmen**. 2000, Vol. Vol. 3.
17. **Sánchez, L. Zepeda**. *Metodología para el Diseño Conceptual de Almacenes de Datos.* Valencia, España : s.n., 2008.
18. **Bernabeu, Ricardo Dario**. *Hefesto: Metodología propia para la construcción de un Data Warehouse.* Córdoba, Argentina : s.n., 2007.
19. **Kimball, Ralph y Ross, Margy**. *The Data Warehouse Toolkit.* New York : Wiley Computer Publishing, 2002.
20. **Sen, Arun y P. Sinha, Atish**. *A Comparison of Datawarehousing Methodologies.* 2005.
21. *Mercado de datos: conceptos y metodologías de desarrollo.* **Arley, Ricardo Chinchilla**. 3, s.l. : Tecnología en Marcha, 2011, Vol. 24.
22. **Huallpa, Rodrigo Anccasi**. [En línea] [Citado el: enero 19, 2015.] www.oocities.org/syscon_sac/r5.html.
23. **Hernández, Ms Yanisbel González**. *Tesis de Maestria: Metodologia de Desarrollo para Proyectos de Almacenes de Datos.* 2013.
24. **García, Gerardo Clemente**. *Tesis doctoral. Un Sistema para el Mantenimiento de Almacenes de Datos.* Valencia : s.n., 2008.

25. **Velasco, Roberto Hernando**. El SGBDR Oracle. [En línea] [Citado el: enero 19, 2015.] http://www2.rhernando.net/modules/tutorials/doc/bd/oracle.html.
26. **Martínez, Daniel Pecos**. geekWare. [En línea] [Citado el: enero 19, 2015.] http://danielpecos.com/documents/postgresql-vs-mysql/#AEN12.
27. **E. Kendall, Kenneth y E. Kendall, Julie**. *Analisis y Diseño de Sistema 3era edic.* 2003.
28. **Mendoza M., Luis Eduardo, Pérez de Ovalles , María Angélica y Grimán P., Anna Cecilia** . Reingenieria de los procesos del negocio. Modelado del Negocio con UML.
29. *El Lenguaje Unificado de Modelado.* **Booch, Grady, Rumbaugh, Jim y Jacobson, Ivar**.
30. *Business Process Modeling Notation.* **BizAgi**. Bogotá, Colombia : s.n.
31. **Kafati, Elizabeth Gutiérrez**. *La plataforma Pentaho Open Source Business Inteligence.*
32. **R. Bouckaert, Remco, y otros**. *WEKA Manual for Version 3-7-10.* Hamilton, Nueva Zelanda : s.n., 2013.
33. **Gimeno, Juan Manuel y Gonzáles, Jose Luis**. *Introducción a Netbeans.* 2011.
34. **S. Pressman, Roger**. *Ingeniería del Software. Un enfoque práctico.* s.l. : Mc Graw Hill, 2010.
35. **Lazo y Piñero** . *Guía base para la especificación de arquitecturas de software.* 2010.
36. **Rodríguez, Marisel Santana**. *Estándares de codificación.* La Habana. Cuba : Departamento Almacenes de datos, 2012.
37. **Martínez Docal, Adrian y Peña Caballero, Jesús Eiseel**. *Mercado de Datos para el Sistema de Gestión Estomatológica alasSiGEST.* La Habana : s.n., 2013.
38. **Melendez Luis, Maylen y Gondres Cobas, Adriana**. *Mercado de datos para el área de Capital Humano de la Universidad de las Ciencias Informáticas.* La Habana : s.n., 2012.

Bibliografía

1. **Hernández Orallo, José, Ramírez Quintana, José María y Ferri Ramírez, César**. I*ntroducción a la Minería de Datos.* Madrid : PEARSON EDUCACIÓN, 2004.

2. *Lineamientos de la política económica y social del Partido y la Revolución.* 2011.

3. **Hernández Sampieri, Roberto, Fernández-Collado, Carlos y Baptista Lucio, Pilar**. *Metodologia de la Investigacion*.Cuarta Edicion.

4. **Hernández León, Rolando Alfredo y Coello González, Sayda**. *El proceso de Investigación científica.* s.l. : Editorial Universitaria Cubana, 2011.

5. *Warehousing and OLAP Analysis of Bibliographic Data.* **Georgieva-Trifonova, Tsvetanka. Veliko Tarnovo**, Bulgaria : s.n., 2011.

6. *DBPubs: Multidimensional Exploration of Database Publications*. **Baid, Akanksha, y otros.**

7. **Brown La O., Reina y Mesa Prada, Luis Eduardo** . *Desarrollo de un mercado de datos para el análisis de datos bibliográficos.* La Habana, Cuba : s.n., 2013.

8. *Practical Machine Leraning Tools and Techniques with Java Implementations.* **Witten Clark, P y Frank Boswell, R.** s.l. : Morgan Kaufmann Publishers, 2000.

9. **Gallardo, Jose Alberto Arencibia**. *Metodología para el Desarrollo de Proyectos en minería de datos CRISP-DM.* Sistemas del Conocimiento.EPB603.

10. **H. Witten, Ian, Frank, Eibe y A. Hall, Mark**. *Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edition.* s.l. : Morgan Kaufmann Publishers , 2005.

11. **Pascual, D., Pla, F. y Sánchez, S**. *Algoritmos de agrupamiento.* Santiago de Cuba,Castellon : s.n.

12. **Boada, Lic. Darian Horacio Grass**. *Tesis de Maestria: Minería de Datos aplicada a los registros de navegación por Internet en la Universidad de Ciencias Informáticas.* La Habana, Cuba : s.n., 2011.

13. **Moine, Ing. Juan Miguel , Haedo , Dra. Ana Silvia y Gordillo , Dra. Silvia** . *Estudio comparativo de metodologías para minería de datos .* Buenos Aires, Argentina : s.n., 2011.

14. **Inmon, William H.** *Building the Data Warehouse*. s.l. : Indianapolis : Wiley Publishing, 2005.

15. **Inmon, William H.** *Building the Data Warehouse.* Fourth Edition. s.l. : Wiley Publishing, 2005.

16. *Revista Ingeniería Informática: revista electrónica del DIICC*. **Wolff, Carmen**. 2000, Vol. Vol. 3.

17. **Sánchez, L. Zepeda**. *Metodología para el Diseño Conceptual de Almacenes de Datos.* Valencia, España : s.n., 2008.

18. **Bernabeu, Ricardo Dario**. *Hefesto: Metodología propia para la construcción de un Data Warehouse.* Córdoba, Argentina : s.n., 2007.

19. **Kimball, Ralph y Ross, Margy**. *The Data Warehouse Toolkit.* New York : Wiley Computer Publishing, 2002.

20. **Sen, Arun y P. Sinha, Atish**. *A Comparison of Datawarehousing Methodologies.* 2005.

21. *Mercado de datos: conceptos y metodologías de desarrollo.* **Arley, Ricardo Chinchilla.** 3, s.l. : Tecnología en Marcha, 2011, Vol. 24.

22. **Huallpa, Rodrigo Anccasi**. [En línea] [Citado el: enero 19, 2015.] www.oocities.org/syscon_sac/r5.html.

23. **Hernández, Ms Yanisbel González**. *Tesis de Maestria: Metodologia de Desarrollo para Proyectos de Almacenes de Datos.* 2013.

24. **García, Gerardo Clemente**. *Tesis doctoral. Un Sistema para el Mantenimiento de Almacenes de Datos.* Valencia : s.n., 2008.

25. **Velasco, Roberto Hernando**. *El SGBDR Oracle.* [En línea] [Citado el: enero 19, 2015.] http://www2.rhernando.net/modules/tutorials/doc/bd/oracle.html.
26. **Martínez, Daniel Pecos**. *geekWare.* [En línea] [Citado el: enero 19, 2015.] http://danielpecos.com/documents/postgresql-vs-mysql/#AEN12.
27. **E. Kendall, Kenneth y E. Kendall, Julie**. *Analisis y Diseño de Sistema 3era edic.* 2003.
28. **Mendoza M., Luis Eduardo, Pérez de Ovalles , María Angélica y Grimán P., Anna Cecilia** . *Reingenieria de los procesos del negocio. Modelado del Negocio con UML.*
29. *El Lenguaje Unificado de Modelado.* **Booch, Grady, Rumbaugh, Jim y Jacobson, Ivar**.
30. *Business Process Modeling Notation.* **BizAgi**. Bogotá, Colombia : s.n.
31. **Kafati, Elizabeth Gutiérrez**. *La plataforma Pentaho Open Source Business Inteligence.*
32. **R. Bouckaert, Remco, y otros**. *WEKA Manual for Version 3-7-10.* Hamilton, Nueva Zelanda : s.n., 2013.
33. **Gimeno, Juan Manuel y Gonzáles, Jose Luis**. *Introducción a Netbeans*. 2011.
34. **S. Pressman, Roger**. *Ingeniería del Software. Un enfoque práctico.* s.l. : Mc Graw Hill, 2010.
35. **Lazo y Piñero** . *Guía base para la especificación de arquitecturas de software.* 2010.
36. **Rodríguez, Marisel Santana**. *Estándares de codificación.* La Habana. Cuba : Departamento Almacenes de datos, 2012.
37. **Martínez Docal, Adrian y Peña Caballero, Jesús Eiseel**. *Mercado de Datos para el Sistema de Gestión Estomatológica alasSiGEST.* La Habana : s.n., 2013.
38. **Melendez Luis, Maylen y Gondres Cobas, Adriana**. *Mercado de datos para el área de Capital Humano de la Universidad de las Ciencias Informáticas.* La Habana : s.n., 2012.
39. *Procesamiento de variantes morfológicas en búsqueda de textos en castellanos.* **Alfredo Bordignon, Fernado Raúl y Panessi, Walter**. 1, 2001, Vol. 24.
40. **Morera Leirado, Mariza, y otros**. *Algoritmo de stemming para el gallego.* España : s.n., 2002.
41. **Rubio Liniers, María Cruz**. *Análisis documental: Indización y resumen en bases de datos especializadas.*
42. *Mercado de datos para la empresa de mantenimiento de grupos electrógenos.* **Brizuela Figueredo, Clara Elena, García Suárez del Villar, Claudia y Brito Rodríguez, Julio Cesar**. La Habana, Cuba : s.n., 2013.
43. *Ayuda del Repositorio Institucional de la UCI.* **Universidad de la Ciencias Informáticas**. La Habana, Cuba : s.n., 2014.
44. **Cuello, Gabriel**. *Técnicas de Minería de Datos dentro de contextos metodológicos y de empresa.* Buenos Aires, Argentina : s.n., 2006.
45. *Descubrimiento de conocimiento en repositorios documentales mediante técnicas de minería de texto y swarm intelligenct.* **Cobo Ortega, Angel, Rocha Blanco, Rocio y Alonso Martínez, Margarita**. 2009, Vol. 10.
46. *Técnicas de Preprocesamiento para mejorar la calidad de los datos en un estudio de caracterización de ingresantes universitarios.* **Dapozo, Gladis, y otros.** Corrientes, Argentina : s.n.
47. *Preprocesamiento de datos estructurados.* **Hernández G., Claudia L. y Rodríguez R., Jorge E.** 2, 2008, Vol. 4.

ANEXOS

A.1 Levantamiento de Información: Captura de requisitos de información

Técnica: Entrevista.

Dirigida a: Usuario de la Aplicación del RI, Usuarios del proyecto de MD, Directora de la Biblioteca y trabajadores, Desarrolladores de la Capa de Almacenamiento en el RI.

1. Aspectos Generales.

(Entrevista a realizar al director(s) de la institución (compañía)).

Institución (compañía):

Ministerio:

Nombre(s) y Apellidos del Entrevistado:

Ocupación:

1. Organigrama de la Institución (compañía):
2. Misión de la Institución: ¿Cuál es la función de la institución o departamento?
3. Visión de la Institución:
4. ¿Cuáles son los procesos (funciones) de la institución? ¿Cuáles son las principales actividades que se realizan, cómo se relacionan estas actividades con las del resto de la organización?
5. ¿Cómo funciona la gestión del Repositorio Institucional? ¿Cuál es la estructura inicial de la documentación que maneja?)
6. ¿Cómo tienen organizada esta documentación? Tipo de cuadro de clasificación.
7. ¿Se realiza alguna toma de decisión a partir de los datos del Repositorio Institucional?
8. ¿Qué tipo y cuales decisiones de realizan?
9. ¿Cuentan con alguna norma y\o resolución que incida en la gestión de los documentos y el Repositorio?
10. ¿Qué datos son almacenados en el Repositorio?
11. ¿Cuál es el dominio o negocio?
12. ¿Qué parte de los datos es pertinente analizar (usuarios, citas bibliográficas, artículos científicos)?
13. ¿Qué tipo de conocimiento se desea extraer y cómo se debe presentar?
14. ¿Qué conocimiento puede ser válido, novedoso e interesante?
15. ¿Qué conocimiento previo me hace falta para realizar esta tarea?
16. ¿Qué aspectos son cruciales en su negocio?
17. ¿Qué reglas o modelos de decisión están utilizando? ¿Se pueden mejorar dichas reglas?

18. ¿Qué base tienen .dichas reglas?

19. ¿Existen decisiones que se toman de una manera arbitraria o basándose en reflexiones personales no explícitas?

20. ¿Existe documentación sobre decisiones anteriores?

21. ¿Quiénes toman las decisiones? ¿Qué decisiones son críticas?

22. ¿Los modelos deben ser comprendidos y validados por expertos?

23. ¿Qué otros requerimientos exigiríamos a los patrones extraídos?

24. ¿Qué conocimiento previo suele utilizar para apoyarse en sus decisiones?

25. ¿Utiliza otras fuentes de datos externas para fundamentarse en sus decisiones?

26. ¿Cuáles son las necesidades y expectativas?

A.2 Registro de Sistemas Fuentes

Tabla 16: Lista de atributos (Fuente) Hecho_publicacion.

Nombre de columna	***Sistema origen***	***Esquema origen***	***Tabla origen***	***Nombre del campo origen***	***Tipo de Dato origen***	***Reglas ETL***	***Comentarios***
id_hec_publicacion	***Proceso ETL***						Llave primaria de la tabla
id_item		Public	dim_Instancia	id_Item	serial	Llave foránea	
id_fecha		Public	dim_Tiempo	id_fecha	serial	Llave foránea	
id_area		Public	dim_lugar_area	id_area	serial	Llave foránea	

Tabla 17: Lista de atributos (Fuente) dim_lugar_area.

Nombre de columna	***Sistema origen***	***Esquema origen***	***Tabla origen***	***Nombre del campo origen***	***Tipo de Dato origen***	***Reglas ETL***	***Comentarios***
id_area	*Proceso ETL*						Llave primaria de la tabla

descripcion		Public	metadatavalue	text_value	text		Verificar que el campo metadata_field_id tenga valor 72

Tabla 18: Lista de atributos (Fuente) dim_tiempo

Nombre de columna	***Sistema origen***	***Esquema origen***	***Tabla origen***	***Nombre del campo origen***	***Tipo de Dato origen***	***Reglas ETL***	***Comentarios***
id_fecha	***Proceso ETL***						Llave primaria de la tabla
anno		Public	metadatavalue	text_value	text		Verificar que el campo metadata_field_id tenga valor 15.
mes		Public	metadatavalue	text_value	text		Verificar que el campo metadata_field_id tenga valor 15.
fuente_id_item		Public	metadatavalue	item_id	serial		
trimestre	*Proceso ETL*						Este dato se crea en el proceso ETL.
semestre	*Proceso ETL*						Este dato se crea en

							el proceso ETL.

Tabla 19: Lista de atributos (Fuente) autor

Nombre de columna	***Sistema origen***	***Esquema origen***	***Tabla origen***	***Nombre del campo origen***	***Tipo de Dato origen***	***Reglas ETL***	***Comentarios***
id_autor	***Proceso ETL***						Llave primaria de la tabla
nombre		Public	metadatavalue	text_value	text		Verificar que el campo metadata_field_id tenga valor 3.

Tabla 20: Lista de atributos (Fuente) autor_x_instancia

Nombre de columna	***Sistema origen***	***Esquema origen***	***Tabla origen***	***Nombre del campo origen***	***Tipo de Dato origen***	***Reglas ETL***	***Comentarios***
id_autor_x_instancia	***Proceso ETL***						Llave primaria de la tabla
id_item		Public	dim_instancia	id_item	serial		
id_autor		Public	autor	id_autor	serial		

Tabla 21: Lista de atributos (Fuente) palabras_claves.

Nombre de columna	***Sistema origen***	***Esquema origen***	***Tabla origen***	***Nombre del campo origen***	***Tipo de Dato origen***	***Reglas ETL***	***Comentarios***

id_palabras_claves	***Proceso ETL***						Llave primaria de la tabla
id_item		Almacen	dim_instancia	id_item	serial		
descripcion		Almacen	metadatavalue	text_value	text		Verificar que el campo metadata_field_id tenga valor 61.

Tabla 22: Lista de atributos (Fuente) gestion_carga_historica.

Nombre de columna	***Sistema origen***	***Esquema origen***	***Tabla origen***	***Nombre del campo origen***	***Tipo de Dato origen***	***Reglas ETL***	***Comentarios***
id_carga_historica	***Proceso ETL***						Llave primaria de la tabla
fecha_hora	***Proceso ETL***						Estos datos se extraen durante el proceso ETL
ip	***Proceso ETL***						Estos datos se extraen durante el proceso ETL
transformacion	***Proceso ETL***						Estos datos se extraen durante el proceso ETL

A.3 Registro de Atributos

Tabla 23: Lista de atributos (Tarjeta) hec_publicacion.

Columna	Descripción	Tipo de Dato	Llave	FK de Tabla	Campo nulo	Ejemplo de Valores
id_publicacion	Id de la tabla publicación	int	PM		No	1,2,3
dim_Instancia_id_Item	Id de la tabla dim_Instancia	int	FK	dim_Instancia	No	1,2,3
dim_Tiempo_id_fecha	Id de la tabla dim_Tiempo	int	FK	dim_Tiempo	No	1,2,3
dim_lugar_area_id_area	Id de la tabla dim_lugar_area	int	FK	dim_lugar_area	No	1,2,3
dim_lugar_centro_id_centro	Id de la tabla dim_lugar_centro	int	FK	dim_lugar_centro	No	1,2,3

Tabla 24: Lista de atributos III (Tarjeta): dim_Tiempo.

Columna	Descripción	Tipo de Dato	Llave	FK de Tabla	Campo nulo	Valor x defecto	Ejemplo de Valores
id_fecha	Id de la tabla publicación	int	PM		No		1,2,3
mes_ id_ mes	Id de la tabla mes	int	FK	mes	No		1,2,3
anno _id_ anno	Id de la tabla anno	int	FK	anno	No		1,2,3
trimestre	Trimestre al que pertenece la publicación	varchar			No		primero, segundo, tercero, cuarto
semestre	Semestre al que pertenece la publicación	varchar			No		primero, segundo

Tabla 25: Lista de atributos IV (Tarjeta): dim_lugar_area.

Columna	Descripción	Tipo de Dato	Llave	FK de Tabla	Campo nulo	Valor x defecto	Ejemplo de Valores
id_area	Id de la tabla dim_lugar_area	int	PM		No		1,2,3
descripción	Facultad o Área de la Universidad a la que pertenece	varchar			No		

Tabla 26: Lista de atributos V (Tarjeta): dim_lugar_centro.

Columna	Descripción	Tipo de Dato	Llave	FK de Tabla	Campo nulo	Valor x defecto	Ejemplo de Valores
id_centro	Id de la tabla dim_lugar_centro	int	PM		No		1,2,3
descripción	Centro al que pertenece	varchar			No		

A.4 Diseño de las transformaciones restantes

Transformación de la tabla “*autor*”:

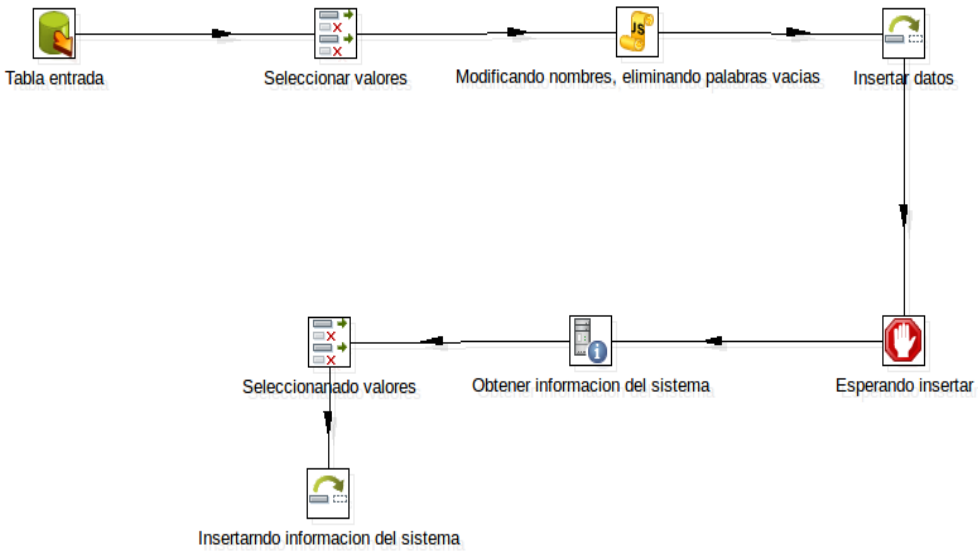

Figura 18: Diseño de la transformación autor.

Transformación de la tabla “*autor por instancia*”:

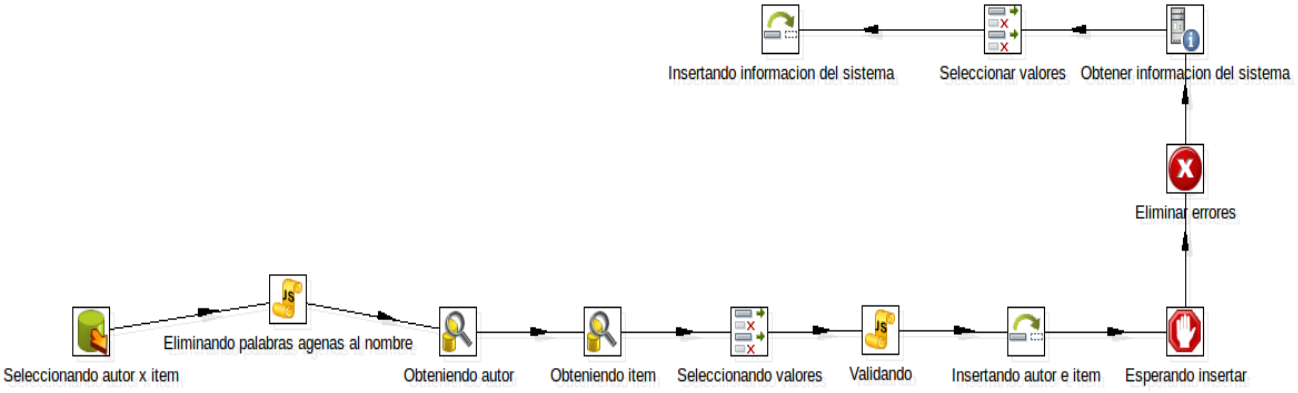

Figura 19: Diseño de la transformación autor_instancia.

Transformación de la dimensión “*instancia*”:

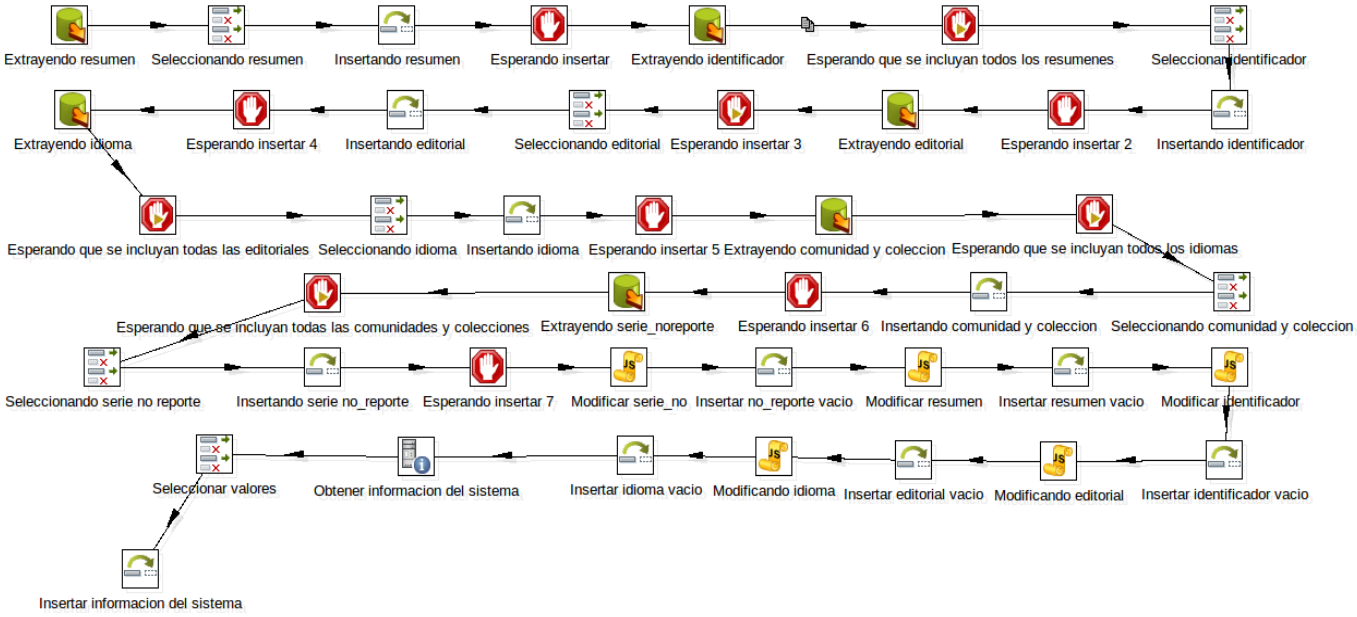

Figura 21 Diseño de la transformación instancia.

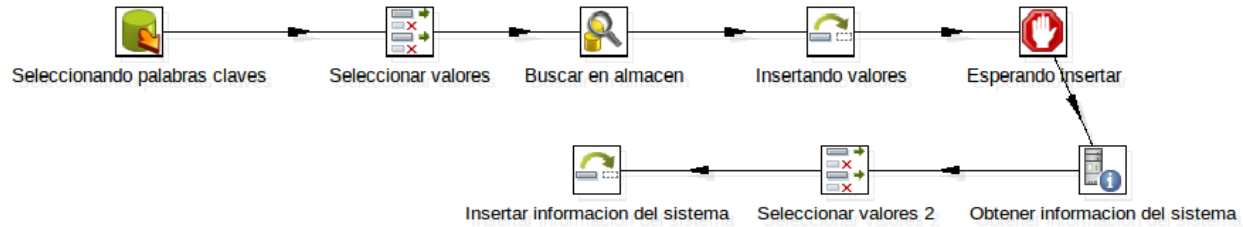

Figura 20: Diseño de la transformación palabras claves.

Transformación de la tabla “*palabras claves*”:

A.5 Atributos a los que se les aplico el filtro “*Replace Mising Values*”

Atributo “anno”

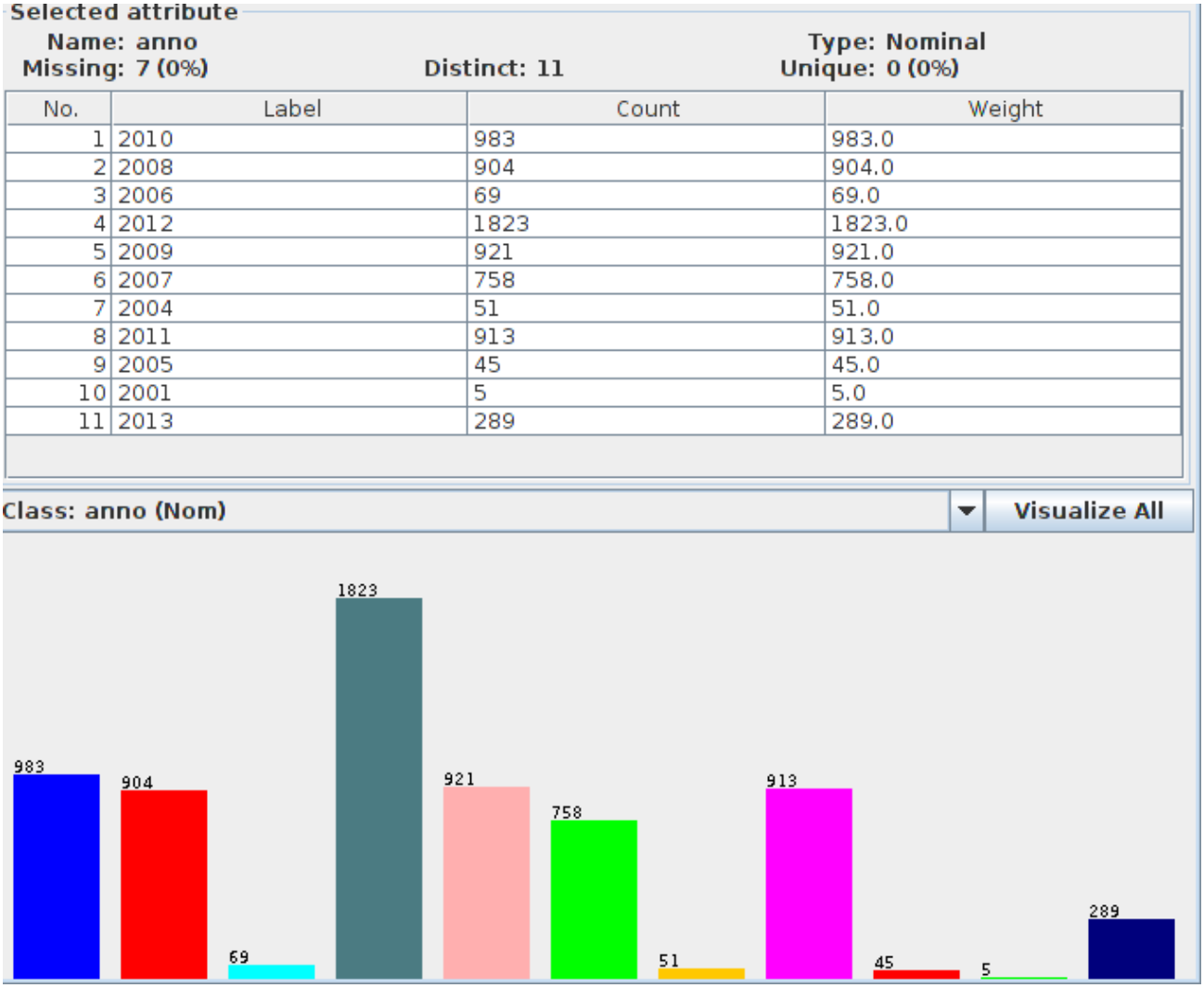

Figura 22: Atributo "anno" con 7 datos erróneos.

Selected attribute

Name: anno Type: Nominal
Missing: 0 (0%) Distinct: 11 Unique: 0 (0%)

No.	Label	Count	Weight
1	2010	983	983.0
2	2008	904	904.0
3	2006	69	69.0
4	2012	1830	1830.0
5	2009	921	921.0
6	2007	758	758.0
7	2004	51	51.0
8	2011	913	913.0
9	2005	45	45.0
10	2001	5	5.0
11	2013	289	289.0

Class: anno (Nom) Visualize All

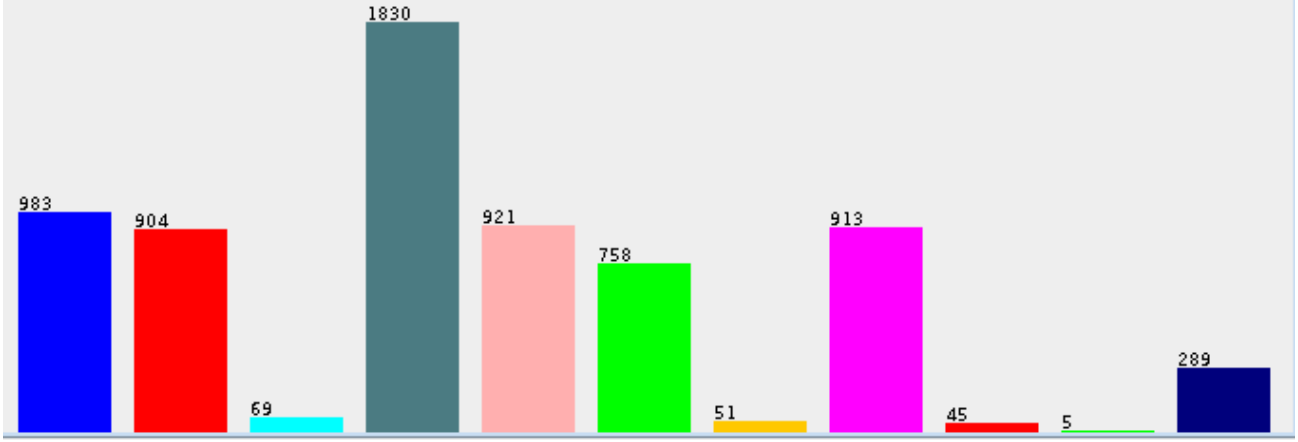

Figura 23: Atributo "anno" después del filtro ReplaceMisingValues