Universidad de las Ciencias Informáticas

Facultad 2

Título: Desarrollo de un mercado de datos para el análisis de datos bibliográficos.

Autores:

Reina Brown La O.

Luis Eduardo Mesa Prada.

Tutor: Ing. Vladimir Milián Núñez.

La Habana, Junio de 2013.

DECLARACIÓN DE AUTORÍA

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmamos la presente a los _____ días del mes de _____ del año 2013.

Reina Brown La O

_____

Firma del Autor

Luis Eduardo Mesa Prada

_____

Firma del Autor

Ing. Vladimir Milián Núñez

_____

Firma del Tutor

PENSAMIENTO

Cada día los pueblos van tomando mayor conciencia… y los resultados están a la vista.

Hugo Rafael Chávez Frías.

AGRADECIMIENTOS

De Reina:

Le agradezco a mi mamá por siempre apoyarme en todas las decisiones que tomo y estar siempre preocupada y atenta a mis estudios.

A mi papá a pesar de no estar presente ya que sé que desde donde esté me está mandando muchas bendiciones y prosperidad para mi vida.

A mis hermanos por ser tan buenos y apoyarme.

A mis abuelas Reina y Ana por ayudarme en todo lo que me ha hecho falta.

A la amiga de mi mamá Aloyda por ser como una tía para mí y a mi tío Alcides.

A mis amigos Aliannis, Leo y Rosalina ya que siempre puedo contar con ellos para lo que necesite.

A mis amistades de fiestas y a mis compañeros de estudios durante estos 5 años.

A mi compañero de tesis por su entrega y dedicación con la tesis.

Al tutor Vladimir Milián por siempre ayudarnos cada vez que teníamos algunas dudas y por el tiempo dedicado.

A la Revolución por darme esta gran oportunidad de estudiar en un sistema educacional gratis.

Y finalmente me agradezco a mí misma por el gran espíritu que tengo y por siempre poner mi mayor esfuerzo en triunfar en todo lo que hago a pesar de los obstáculos que aparezcan en el camino.

De Luis Eduardo:

Quiero empezar por mi madre, ya que no tengo forma de agradecerle su amor, su comprensión y todo el apoyo que siempre me ha brindado. Gracias por confiar en mí y por ayudarme a llegar hoy aquí.

Agradecer especialmente a mi abuela Ana, a mi padre Raúl, a mi novia Sady y a mi hermana Dailin, por la ayuda y el apoyo brindado en los momentos difíciles.

A mis primos, mis tías y tío por estar siempre de mi lado, gracias también por sus ánimos y recomendaciones que me ayudaron a seguir adelante.

Agradecer también a la familia de mi novia, por tanta preocupación hacia mí y mis estudios.

A mi compañera de tesis, que solo ella sabe tantos malos momentos que pasamos durante la confección de nuestro trabajo de diploma.

A mis grandes amigos Luis Manuel y Ángel, gracias por toda la ayuda brindada.

A mi piquete del día a día, el Fernand, Jose, Lizy, Maday y otros más que lamentablemente tuvieron que quedar atrás.

A mi tutor Vladimir, gracias por toda la ayuda y consejos brindados.

A Alejandro y Danisey, que fue decisiva la ayuda brindada por ellos.

DEDICATORIA

De Reina:

A mi papá ya que es su sueño unido al mío hecho Realidad.

A mi mamá principalmente por siempre estar a mi lado apoyándome y ser madre y padre a la vez para mí y mis hermanos.

A todas las personas que me han apoyado y brindado su amistad.

De Luis Eduardo:

A mi mamá querida, que ya nos graduamos, ella y yo.

A mi abuela Ana, a mi padre Raúl, a mi novia Sady, a mi hermana Dailin, mis tías Edivia y Rosa, mi tío Eduardo, así como a mis primos Dayana y Tito y al gran César. A mi suegra, mi mamá suegra y a mi cuñada, así como a mi suegro y papá suegro.

Dedico este trabajo a toda mi familia, a mis buenos amigos y a los que me quieren y que me brindaron su ayuda.

Resumen

El presente trabajo tiene como objetivo desarrollar un mercado de datos para apoyar al proceso de toma de decisiones que realizan los especialistas en Ciencia de la Información del Instituto de Cibernética, Matemática y Física de Cuba (ICIMAF) que pertenecen al grupo de trabajo de los Datos No Estructurados o Semi-Estructurados, para el área de los datos bibliográficos. La construcción del mercado de datos facilita un acceso a los datos almacenados y permite que exista información relevante para la toma de decisiones. El análisis, diseño, modelo dimensional y de datos e implementación está basado en la metodología de Kimball.

La fase más relevante del proceso de construcción del mercado de datos es el proceso de extracción, trasformación y carga de los datos, junto al procesamiento analítico relacional en línea. Se utilizó para el desarrollo la arquitectura Bottom-up y la arquitectura de datos de 3 capas. Para la implementación la Suite de Inteligencia de Negocio Pentaho, para el modelo la herramienta SQL-Power Architect, como gestor de base de datos PostgreSQL y como herramienta de administración el Administration Console.

PALABRAS CLAVES:

Bibliometría, datos bibliográficos, almacén de datos, mercado de datos, OLAP.

INTRODUCCIÓN

El desarrollo de la ciencia y la tecnología proporcionó el surgimiento de una nueva ciencia: la Informática, que se define como *"…conjunto de conocimientos científicos y técnicas que hacen posible el tratamiento automático de la información por medio de ordenadores..."*; (1) ayudando a mejorar la calidad de vida del hombre.

Dentro de las aplicaciones importantes que se le confieren a la Informática se encuentra el ser capaz de proveer gran cantidad de información efectiva en un mínimo de tiempo, lo cual contribuye en gran medida a la toma de decisiones por parte de los usuarios.

El hombre trabaja en diferentes ramas de la ciencia con grandes volúmenes de información, esto se manifiesta de modo considerable en la rama de la comunicación impresa donde el trabajo con documentación es excesivo haciendo primordial el uso de técnicas bibliométricas para solventar la situación. La bibliometría *"…es una disciplina con alcance multidisciplinario, analiza uno de los aspectos más relevantes y objetivos de la comunicación impresa…"* (2)

El Instituto de Cibernética, Matemática y Física de Cuba (**ICIMAF**) es un centro de investigación, el cual tiene como misión gestionar y ejecutar proyectos de investigación, desarrollo e innovación con personal motivado y de competencia reconocida. Además de realizar actividades de formación posgraduada, asesorías y servicios científicos y tecnológicos que brinden soluciones de alto valor agregado. (3)

Actualmente el ICIMAF se encuentra inmerso en el proyecto de gran envergadura "Vigilancia Tecnológica para la Investigación y la Innovación Científica" el cual es desarrollado para el Centro de Gestión de la Información y Desarrollo de la Energía (CUBAENERGÍA). Con el propósito de garantizar y facilitar el desarrollo de este proyecto el ICIMAF decidió crear cuatro grupos de trabajo:

1. Grupo de trabajo de tratamiento de datos estructurados (numéricos).
2. Grupo de trabajo de tratamientos de datos no estructurados o semi-estructurados (patentes, referencias de artículos, documentos a textos completos).
3. Grupo de trabajo de datos semi-estructurados (procedentes de internet e intranet).
4. Grupo de trabajo de desarrollo de biblioteca digital y repositorios de la organización.

El segundo grupo de trabajo de los antes mencionados, tiene como principal misión el crear un “Sistema para el Tratamiento de Datos No Estructurados o Semi-Estructurados con Técnicas de Minería de Datos” (Sistrada NOYSE), en el cual es necesario contar con una herramienta que permita realizar análisis de los datos bibliográficos que se obtienen por los investigadores. De analizar los metadatos bibliográficos (autor(es), palabras claves, institución, país, tiempo, entre otros), nos da la posibilidad de conocer por ejemplo el(los) autor(es) que más publica(n), áreas donde más se ha publicado, fechas de mayor producción científica, instituciones que más publican, áreas emergentes de investigación, producción científicas de investigadores, departamentos, instituciones o países en determinado espacio de tiempo, entre otros.

Para realizar este análisis es necesario contar con un cúmulo grande de datos, los cuales deben almacenarse durante cierto tiempo para realizar estudios y tomar decisiones a partir del análisis de la evolución de los indicadores mencionados anteriormente. Estos datos, por lo general, no se encuentran almacenados en un sistema central, con un formato común para todos y ocurre que muchas veces se encuentren duplicados. Por eso contar con un sistema que sea capaz de almacenar, gestionar y realizar un análisis histórico de la información almacenada nos posibilitará obtener información relevante de los datos bibliográficos de las publicaciones utilizadas por nuestros investigadores en sus trabajos, así como de las publicaciones realizadas por ellos.

Una solución a estos inconvenientes son los almacenes de datos, los cuales ofrecen el acceso a los datos integrados e históricos de las fuentes de información y proporcionan un apoyo a los directivos para el proceso de toma de decisiones.

Considerando los elementos expuestos anteriormente puede afirmarse que con un adecuado análisis a los almacenes de datos se adquiere un acceso sencillo y rápido a la información. Por lo que se considera como **problema a resolver:** ¿Cómo facilitar el proceso de análisis de datos bibliográficos para ayudar en la toma de decisiones que realizan los especialistas del **ICIMAF**?

El **objeto de estudio** es el proceso de análisis bibliométricos; el **campo de acción** está centrado en los indicadores bibliográficos.

Para darle respuesta al problema a resolver se define como **objetivo general** desarrollar un mercado de datos para el análisis de datos bibliográficos.

Derivándose de este los siguientes **objetivos específicos**:

- Determinar los requisitos del mercado de datos para el análisis de datos bibliográficos.
- Realizar el modelado dimensional y físico del mercado de datos.
- Implementar el proceso de extracción, transformación y carga del mercado de datos.
- Implementar el mercado de datos para el análisis de datos bibliográficos.
- Diseñar e implementar el o los cubos necesarios para el análisis de la información.
- Realizar pruebas para validar el mercado de datos realizado.

Idea a defender: Con el desarrollo de un mercado de datos para el análisis de los datos bibliográficos, se facilita el proceso de análisis de los mismos, apoyando así la toma de decisiones en el área de los Datos No Estructurados o Semi-Estructurados del proyecto "Vigilancia Tecnológica para la Investigación y la Innovación Científica" que desarrolla el **ICIMAF**.

Para dar cumplimiento a los objetivos se plantearon las siguientes **tareas investigativas**:

- Análisis de soluciones similares a las que se quiere implementar.
- Descripción de las herramientas y metodología a utilizar para el desarrollo del mercado de datos.
- Identificación de temas analíticos del mercado de datos para el análisis de datos bibliográficos.
- Identificación de los procesos del negocio.
- Identificación de los requisitos del mercado de datos.
- Identificación de los niveles de granularidad del mercado de datos.
- Identificación de las dimensiones, tablas de hechos y medidas del mercado de datos.
- Identificación de atributos de dimensiones y tablas de hechos del mercado de datos.
- Diseño del modelo dimensional detallado del mercado de datos.
- Diseño del modelo datos del mercado de datos.
- Revisión y validación del modelo dimensional del mercado de datos.
- Diseño del proceso ETL del mercado de datos.
- Realización del proceso de ETL para cargar las dimensiones y tablas de hechos del mercado de datos.

- Configuración de la herramienta de administración del mercado de datos.
- Diseño de los cubos OLAP necesarios para el análisis de datos bibliográficos.
- Implementación de los cubos OLAP necesarios para el análisis de datos bibliográficos.
- Implementación de los reportes del mercado de datos.
- Desarrollo del tablero de mando del mercado de datos.
- Selección de las técnicas de pruebas a realizar del mercado de datos.
- Aplicación de las pruebas para el mercado de datos.
- Recolección y clasificación de las no conformidades.

Métodos de Investigación

Métodos teóricos

- **Análisis Histórico – lógico:** se realiza un estudio del estado del arte sobre las aplicaciones informáticas que han utilizado los almacenes de datos para el análisis de datos bibliográficos, con el objetivo de conocer como estas aplicaciones manejan la información para la toma de decisiones.
- **Analítico – Sintético:** se llevó a cabo 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 del presente trabajo.
- **Inductivo – Deductivo:** la utilización de este método permitió establecer conclusiones y obtener una idea del funcionamiento del sistema, partiendo de la información consultada para el mismo.
- **Modelación:** se logra una relación entre el modelo y el objeto, para ello se definen las dimensiones del mercado de datos, el hecho asociado a las dimensiones definidas además se estructura el modelo dimensional y se transforma al diseño físico.

Métodos Empíricos

- **Entrevista:** se entrevista a los especialistas en Ciencia de la Información del ICIMAF para el levantamiento de los requisitos, a partir de ellos se definen los requisitos informacionales.

La tesis está estructurada de la siguiente manera: resumen, introducción, tres capítulos, conclusiones, recomendaciones, glosario de términos, referencias bibliográficas, bibliografía y anexos.

Capítulo 1: Fundamentación teórica.

En este capítulo se abordan definiciones y conceptos de vital importancia relacionados con el tema del mercado de datos. Se realiza un estudio, comparación y selección de la metodología a seguir para el proceso de desarrollo del mercado de datos. Se fundamenta el uso de las distintas herramientas y se selecciona el gestor de bases de datos a utilizar.

Capítulo 2: Análisis y diseño del mercado de datos.

En este capítulo se presenta un procedimiento a partir de las características del negocio y los aspectos más significativos de la metodología de Kimball. Además se presentan los resultados del diseño del mercado de datos para el análisis de datos bibliográficos. Se determinan los indicadores necesarios para la construcción del mercado de datos.

Capítulo 3: Evaluación del proceso de implementación del mercado de datos.

En este capítulo se realiza la implementación del proceso de Extracción, Transformación y Carga (por sus siglas en inglés ETL) para el cubo del mercado de datos. Se generan los reportes del mercado de datos y se desarrollara el tablero de mando. Además, se seleccionan y aplican las pruebas al mercado de datos con el objetivo de validar la solución propuesta.

CAPÍTULO 1: Fundamentación teórica.

1.1 Introducción

En este capítulo se abordan definiciones y conceptos relacionados con el tema del mercado de datos. Se realiza un estudio, comparación y selección de la metodología a seguir para el proceso de desarrollo del mercado de datos. Se fundamenta el uso de las distintas herramientas y se selecciona el gestor de base de datos a utilizar.

1.2 Análisis de soluciones similares (estado del arte)

Desde los años ochenta se ha experimentado una invasión del mundo de la Informática en el campo de la investigación en todos los niveles, específicamente, en las bases bibliográficas, especialmente en el análisis de los datos contenidos en estas. Las aplicaciones y aportes de la Informática a esta investigación han sido catalogados como importantes, teniendo en cuenta que la cantidad de datos que se manejan y la variedad de análisis que se le realizan a estos pueden rebasan la capacidad del cálculo manual.

El uso de los almacenes de datos con técnicas OLAP en el análisis de datos bibliográficos es reciente. Uno de los primeros sistemas que usó esta tecnología fue la base de datos nacional de biomedicina de la república de Eslovenia "Biomedicina Slovenica". (15) Este demuestra su utilidad a través de la extracción de datos para el estudio de temas cienciométricos como los cambios en los números de artículos publicados y de citas así como el número de autores.

Otro sistema implementado fue DBPubs, este utiliza el análisis y exploración de los contenidos de bases de datos de publicaciones mediante técnicas que combinan la búsqueda de palabras claves con operaciones OLAP. (16)

También existe, el bgMath/OLAP, el cual propone el uso de técnicas OLAP sobre un almacén de datos que se implementa para el análisis de literatura científica relacionada con el campo de las matemáticas.

Recientemente se desarrolló en la universidad de Sao Pablo una herramienta inteligente de apoyo a la investigación en la que se utiliza un almacén de datos para la indexación automática de tópicos de investigaciones.

Producto de las necesidades del estudio de los datos almacenados, Cuba también ha desarrollado programas en este sentido sobresaliendo la recopilación de la bibliografía

médica cubana CUMED generada por la Biblioteca Médica Nacional siendo su principal función el control bibliográfico de la producción documentaria de las revistas editadas por la Ecimed y constituye un depósito legal de las investigaciones cubanas.

También es usada en Cuba LILACS la base de datos de la Literatura Latinoamericana y del Caribe en Ciencias de la Salud, coordinada por BIREME (Centro Latinoamericano y del Caribe de Información en Ciencias de la Salud), a la que los diferentes países comprendidos en el área geográfica de Brasil contribuyen con el envío de registros bibliográficos correspondientes a la producción literaria nacional en el área de la salud. Se analizan y procesan documentos como tesis, libros, capítulos de libros, anales de congresos o conferencias, informes técnico-científicos, publicaciones gubernamentales y artículos seleccionados en el procesamiento de aproximadamente 550 títulos de revistas del área continental.

Luego de valorados los sistemas antes mencionados se puede llegar a la siguiente conclusión:

Aunque algunos de los sistemas antes mencionados (Biomedicina Slovenica y bgMath/OLAP) presentan las características deseadas para darle respuesta a la problemática que dio origen a este trabajo de tesis, los mismos son soluciones para un tema (área) específico, desarrollados con tecnologías privativas por entidades que no ponen a disposición de la comunidad ni la solución de software realizada, ni el código fuente de las mismas. Las restantes, aunque trabajan con datos bibliográficos y tecnologías de almacenes de datos u OLAP, no tienen como objetivo final darle respuesta a la problemática planteada en la presente investigación. Es por ello que se hace necesaria la realización de un mercado de datos para facilitar el proceso de análisis de datos bibliográficos para apoyar el proceso de toma de decisiones por parte de los especialistas del **ICIMAF**.

1.3 La bibliometría para el trabajo con los datos bibliográficos

En 1969 Alan Pritchard definió el término Bibliometría como la aplicación de los métodos estadísticos y matemáticos dispuestos para definir los procesos de la comunicación escrita y la naturaleza y el desarrollo de las disciplinas científicas mediante técnicas de recuento y análisis de dicha comunicación. (4)

El tratamiento y manejo de la literatura científica por medios cuantitativos de recuento y análisis sirve no solo para analizar el volumen de publicaciones, la productividad de

autores, revistas o materias, sino también en un sentido más amplio, para el conocimiento de los procesos y la naturaleza de las Ciencias. (4)

La bibliometría es una disciplina con alcance multidisciplinario, analiza uno de los aspectos más relevantes y objetivos de esa comunidad, la comunicación impresa. Comprende la aplicación de análisis estadísticos para estudiar las características del uso y creación de documentos, estudio cuantitativo de la producción de documentos y de las unidades físicas publicadas, o de las unidades bibliográficas. (2)

La cienciometría aplica técnicas bibliométricas a la ciencia, esta va más allá de las técnicas bibliométricas pues también examina el desarrollo y las políticas científicas; usa técnicas matemáticas y el análisis estadístico para investigar las características de la investigación científica. Se encarga de la evaluación de la producción científica mediante indicadores numéricos de publicaciones, patentes, etc. (2)

Utilizando técnicas bibliométricas y cienciométricas es posible:

- Identificar las tendencias y el crecimiento del conocimiento en las distintas disciplinas.
- Estimar la cobertura de las revistas secundarias.
- Identificar los usuarios de las distintas disciplinas.
- Identificar autores y tendencias en distintas disciplinas.
- Medir la utilidad de los servicios de diseminación selectiva de información.
- Predecir las tendencias de publicación.
- Identificar las revistas del núcleo de cada disciplina.
- Formular políticas de adquisiciones ajustadas al presupuesto.
- Adaptar políticas de descarte de publicaciones.
- Estudiar la dispersión y la obsolescencia de la literatura científica.

Los análisis bibliométricos se enmarcan en el análisis de datos bibliográficos y en el análisis de citas. La presente tesis se centra en los indicadores bibliográficos.

1.4 Almacén de datos

Un almacén de datos (conocido en inglés como data warehouse) proporciona un ambiente para que las organizaciones hagan un mejor uso de la información que está siendo administrada por diversas aplicaciones operacionales. Posibilita la extracción de datos de sistemas operacionales y fuentes externas, la integración y homogenización de

los datos de toda la empresa, provee información que ha sido transformada y resumida, para que ayude en el proceso de toma de decisiones estratégicas y tácticas. (5)

Tiene como objetivo almacenar y proveer a una organización de información relevante y a tiempo. Su utilización tiene gran importancia ya que mejora la entrega de información, mejora el proceso de toma de decisiones y tienen un impacto positivo sobre los procesos empresariales.

Se características por ser:

- **Orientado al tema:** En el almacén de datos, la información se clasifica en base a los aspectos que son de interés para la organización, sin embargo en el ambiente operacional la estructura de los datos se diseña en torno a las aplicaciones. Las diferencias entre las funciones de las aplicaciones y la orientación a temas, radican en el contenido de los datos a nivel detallado. En el almacén de datos se excluye la información que no será usada por el proceso de sistemas de soporte de decisiones, mientras que las orientadas a las aplicaciones contienen datos para satisfacer de inmediato los requisitos funcionales, que pueden ser usados o no por el analista de soporte de decisiones. (5)
- **Integrado:** La integración implica que los datos de diversas fuentes que son generados por distintos departamentos, secciones y aplicaciones, tanto internas como externas se consoliden en una instancia antes de ser agregados al almacén de datos, y deben por lo tanto ser analizados para asegurar su calidad y limpieza. (5)
- **De tiempo variable:** Un almacén de datos es de tiempo variante porque toda la información es requerida en algún momento, donde puede ser consultada con el objetivo de ser utilizada en comparaciones, tendencias y previsiones. Debido a que los datos pueden ser consultados en un corto, largo o mediano intervalo de tiempo. Los datos almacenados en estos, no pueden ser actualizados. Toda estructura clave de un almacén contiene implícita o explícitamente un elemento de tiempo, constituyendo una de las principales ventajas del almacén de datos, pues los datos son almacenados junto a sus respectivos históricos. (5)
- **No volátil:** La información de un almacén de datos existe para ser leída, pero no modificada. La información es por tanto permanente, considerando como actualización del almacén de datos, la incorporación de los últimos valores que

tomaron las distintas variables contenidas en él, sin ningún tipo de acción sobre los que ya existían. (5)

Los almacenes de datos se conforman de varios mercados de datos, el presente trabajo se centra precisamente en la construcción de un mercado de datos.

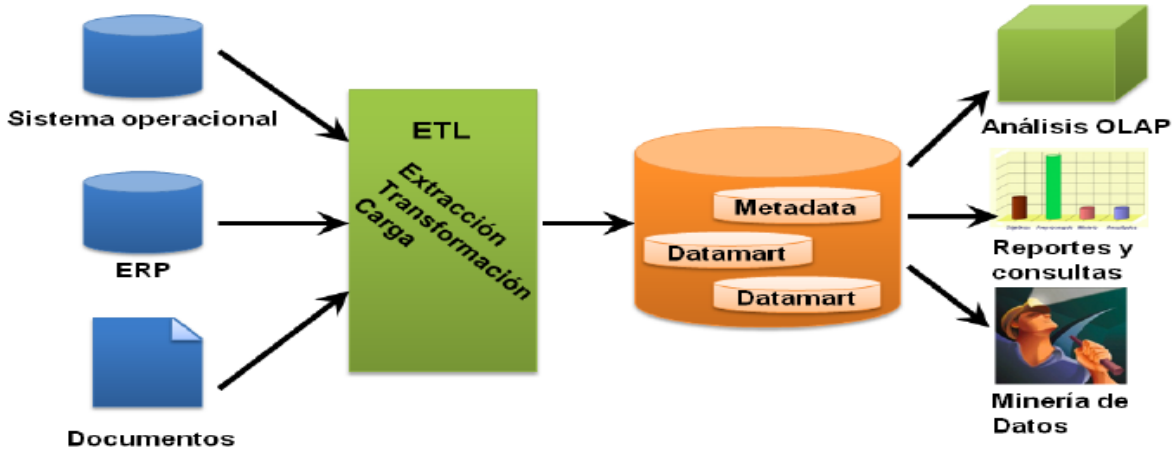


Figura 1: Arquitectura de un almacén de datos.

1.4.1 Procesos básicos de un almacén de datos. ETL (Extracción, Transformación y Carga)

Es la base sobre la cual se alimenta el almacén de datos. Si el sistema ETL se diseña adecuadamente, permite extraer los datos de los sistemas de origen de datos, aplicar diferentes reglas para aumentar la calidad y consistencia de los mismos, consolidar la información proveniente de distintos sistemas, y finalmente cargar (grabar) la información en el almacén de datos en un formato acorde para la utilización por parte de las herramientas de análisis. El proceso ETL se conforma de tres funciones que se enuncian a continuación:

Extracción: consiste en estudiar y entender los datos fuente, tomando aquellos que son de utilidad para el almacén de datos.

Transformación: una vez que los datos son extraídos, éstos se transforman. Este proceso incluye corrección de errores, resolución de problemas de dominio, borrado de campos que no son de interés, generación de claves, agregación de información, etc.

Carga: al terminar el proceso de transformación, se cargan los datos en el almacén de datos. Las formas más básicas para desarrollar el proceso de carga son dos:

- **Acumulación simple:** consiste en realiza un resumen de todas las transacciones comprendida en el periodo de tiempo seleccionado y transportar el resultado como una única transacción hacia el almacén de datos para su almacenamiento.
- **Rolling:** almacena información resumida a distintos niveles, correspondientes a distintas agrupaciones de la unidad de tiempo o diferentes niveles jerárquicos.

1.4.2 Definiciones relacionadas con el proceso ETL

Staging area: es un área temporal donde se recogen los datos que se necesitan de los sistemas origen. Se recogen los datos estrictamente necesarios para las cargas, y se aplica el mínimo de transformaciones a los mismos. No se aplican restricciones de integridad ni se utilizan claves, los datos se tratan como si las tablas fueran ficheros planos. De esta manera se minimiza la afectación a los sistemas origen, la carga es lo más rápida posible para minimizar la ventana horaria necesaria, y se reduce también al mínimo la posibilidad de error. Una vez que los datos están traspasados, el almacén de datos se independiza de los sistemas origen hasta la siguiente carga. Lo único que se suele añadir es algún campo que almacene la fecha de la carga. (7)

1.5 Mercado de datos

Un mercado de datos (conocido en inglés como data mart) es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento. Puede ser alimentado desde los datos de un almacén de datos, o integrar por sí mismo un compendio de distintas fuentes de información. (5)

Ventajas del mercado de datos:

- Fácil acceso a los datos que se utilizan con frecuencia.
- Mejora los tiempos de respuesta hacia el usuario final.
- El costo es inferior al de aplicar un almacén de datos.
- Los usuarios potenciales son identificados con más claridad.
- Se pueden crear vistas colectivas.

1.5.1 Mercado de datos OLAP

Se basan en los cubos OLAP, que se construyen agregando, según los requisitos de cada área o departamento, las dimensiones y los indicadores necesarios de cada cubo relacional. El modo de creación, explotación y mantenimiento de los cubos OLAP es muy heterogéneo, en función de la herramienta final que se utilice. (8)

1.5.2 Metadatos

Conjunto estructurado de datos que describen a otros datos, a su estructura interna y a sus servicios, cuyo propósito es incrementar el conocimiento y contestar a preguntas del tipo “qué”, “quién”, “dónde”, “cuándo”, “cuánto” y “cómo”. También pueden considerarse productos autónomos asociados a los datos que permiten mantener un inventario de los mismos, facilitar su publicación y consulta a través de catálogos en las Infraestructuras de datos espaciales (IDE) y facilitar la reutilización de los datos y la explotación de los servicios. (9)

1.6 Modelado multidimensional

Para la creación del mercado de datos uno de los modelos más utilizados son los modelos multidimensionales, los cuales dieron surgimiento a las bases de datos multidimensionales, garantizando una búsqueda rápida de los datos.

Existen tres variantes de modelado de las bases de datos dimensionales:

- **Esquema en estrella:** Este esquema consta de una tabla de hechos central que se relaciona con varias tablas de dimensiones. Debe estar totalmente desnormalizado, siendo innecesarias las uniones entre las tablas cuando se realiza una consulta, garantizando mayor rapidez en las consultas. Entre sus ventajas más significativas está que es el esquema más simple de interpretar, posee los mejores tiempos de respuesta, su diseño es fácil de modificar y simplifica el análisis. (10)
- **Esquema copo de nieve:** Constituye una extensión del esquema en estrella cuando las tablas de dimensiones se organizan en jerarquías de dimensiones. Consta de una tabla de hechos central que está relacionada con una o más tablas de dimensiones, quienes a su vez pueden relacionarse o no con nuevas dimensiones. Es muy flexible y puede ser implementado luego de haber desarrollado un esquema en estrella, siendo muy útil en las tablas de dimensiones de muchas tuplas. Tiene como desventaja que de existir muchas tablas de

dimensiones, cada una de ellas con varias jerarquías, pueden crearse abundantes tablas llegando a ser inmanejables. (10)

- **Esquema constelación:** Este modelo está compuesto por una serie de esquemas en estrella, conformado por una tabla de hechos principal y una o más tablas de hechos auxiliares. Estas tablas hacen en el centro del modelo y están relacionadas con sus respectivas tablas de dimensiones, vinculándose las tablas de hechos auxiliares con algunas dimensiones asignadas a la tabla de hecho principal y también con nuevas tablas de dimensiones. (10)

Luego de haber analizado todos los esquemas, el esquema a utilizar será el esquema en estrella.

1.7 Arquitectura

La estructura que reúne todos los componentes de un almacén de datos es conocido como arquitectura. Es la forma de representar la organización total de los datos, comunicación, procesamiento y presentación.

"…la arquitectura incluye todo lo que se necesita para preparar y guardar los datos. Por otro lado, también contiene todos los recursos para distribuir la información desde el almacén de datos. Está compuesta más allá de reglas, procedimientos y funciones que permiten al almacén de datos trabajar y cumplir los requisitos de la empresa… Define las normas, medidas, diseño general, y técnicas de apoyo". (11)

Los almacenes de datos pueden adoptar las siguientes arquitecturas: Top-Down y Bottom-up.

Top-Down: en esta arquitectura primero se define el almacén de datos y luego se desarrollan, construyen y cargan los mercados de datos a partir del mismo.

Figura 2: Arquitectura Top-Down.

El almacén de datos es cargado a través de procesos ETL y luego este alimenta los diferentes mercados de datos.

Bottom-up: en esta arquitectura se definen los mercados de datos y luego se integran en un almacén de datos centralizado.

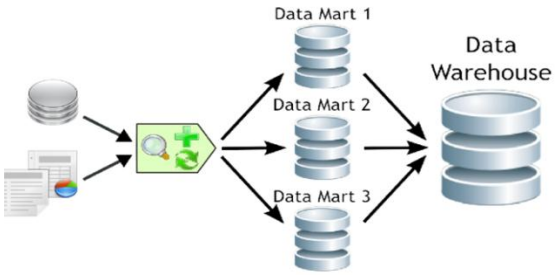


Figura 3: Arquitectura Bottom-up.

Los mercados de datos se cargan a través del proceso ETL, el que suministra la información adecuada a cada uno de ellos. En muchas ocasiones los mercados de datos son implementados sin que exista un almacén de datos, ya que tienen sus mismas características pero con la particularidad de que están enfocados en un tema específico. Luego de que hayan sido creados y cargados todos los mercados de datos, se procederá a su integración con el depósito.

Luego de estudio realizado se propone la utilización de la arquitectura Bottom-up para la creación del mercado de datos.

1.7.1 Arquitectura de datos

Existen tres tipos de arquitectura de datos, las cuales son: arquitectura de una capa, arquitectura de dos capas y arquitectura de tres capas. La arquitectura a utilizar será la de tres capas.

En la arquitectura de tres capas para realizar la transformación de los datos de tiempo real a datos derivados, es necesario utilizar una capa intermedia, conocida como capa de datos reconciliados, en la cual se solucionan los problemas de inconsistencias y se realiza el procesamiento de los distintos conjuntos de datos de tiempo real adecuadamente.

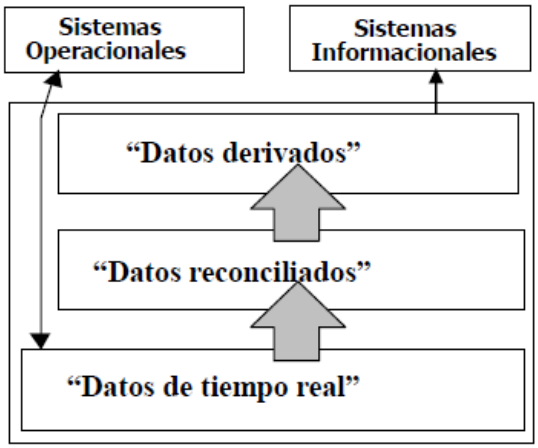


Figura 4: Arquitectura de tres capas.

1.8 Tecnología OLAP

Un sistema de Procesamiento Analítico en Línea (OLAP, por sus siglas del inglés *Online Analytical Process*) es una proyección multidimensional redundante de una relación. Al computar todas las consultas del tipo "agrupar por" (group by), realiza una agregación de sus resultados en un espacio N-dimensional para responder consultas. (12) Un ejemplo clásico es en el análisis de ventas de determinada empresa, donde se pueden agrupar por producto, región y año, para responder preguntas sobre el comportamiento de los montos totales de las ventas en dependencia de estos factores (dimensiones).

OLAP proporciona varias características a los usuarios que realizan análisis, como por ejemplo:

- Provee análisis multidimensional dinámico, permitiendo a los usuarios finales realizar actividades analíticas y navegacionales, que incluyen cálculo de dimensiones, análisis en periodos de tiempo, visualización de subconjuntos de datos, subir o bajar niveles, comparaciones de varias dimensiones en el área de visualización, etc.
- Está basado en una modalidad cliente/servidor multiusuario, que ofrece respuestas rápidas, independientemente del tamaño y la complejidad de la base de datos. (13)

Para el procesamiento analítico en línea existen actualmente tres modelos, **MOLAP**, **ROLAP** y **HOLAP**. El proceso de análisis es realizado de la misma forma pero en uno y otro caso varia la metodología de almacenamiento. Esta forma de almacenamiento es

crítica para garantizar la velocidad de recuperación de la información y el procesamiento de los datos en general.

- **ROLAP (Procesamiento analítico relacional en línea):** Este tipo de organización física se implementa sobre tecnología relacional, pero disponen de algunas facilidades para mejorar el rendimiento. ROLAP, *Relational On Line Analytic Processing*, cuenta con todos los beneficios de una Sistema Gestor de Base de Datos (SGBD) Relacional a los cuales se les provee extensiones y herramientas para poder utilizarlo como un sistema gestor de almacén de datos. (12) (14)
- **MOLAP (Procesamiento analítico multidimensional en línea):** Usa unas bases de datos multidimensionales para proporcionar el análisis. Su principal premisa es que el OLAP está mejor implantado almacenando los datos multidimensionalmente. El almacenaje de MOLAP, Multidimensional Online *Analytical Processing*, provee excelente rendimiento y compresión de datos. Tiene el mejor tiempo de respuesta, dependiendo solo en el porcentaje y diseño de las agregaciones del cubo. En general este método, es muy apropiado para cubos con uso frecuente por su rápida respuesta. (14)
- **HOLAP (Procesamiento analítico híbrido en línea):** El desarrollo más reciente ha sido la solución OLAP híbrida HOLAP, *Hybrid Online Analytical Process*, la cual combina las arquitecturas ROLAP y MOLAP para brindar una solución con las mejores características de ambas: desempeño superior y gran escalabilidad. Un tipo de HOLAP mantiene los registros de detalle (los volúmenes más grandes) en la base de datos relacional, mientras que mantiene las agregaciones en un almacén MOLAP separado. Los cubos almacenados como HOLAP, son más pequeños que los MOLAP y responden más rápidos que los ROLAP. HOLAP es generalmente usado para cubos que requieren rápida respuesta, para sumarizaciones basadas en una gran cantidad de datos. (14)

Se implementará el proceso analítico en línea utilizando ROLAP dado que la base de datos que se construirá como parte del mercado de datos está soportada por el gestor de base de datos PostgreSQL, el cual es un sistema gestor de base de datos relacional.

1.8.1 Herramientas OLAP

Las herramientas OLAP son las aplicaciones que se encargan de formar los cubos multidimensionales de este tipo de soporte y de analizarlos con el objetivo de producir y obtener la información más completa posible. Gracias a estas herramientas los usuarios

corporativos tienen la oportunidad de sacar el máximo partido a las bases de datos de información. Su principal finalidad es beneficiar a la hora de llevar a cabo análisis completos del estado de la empresa y de obtener las estadísticas más completas y detalladas de todo lo relacionado con los resultados corporativos. En base a estos resultados proporcionados por OLAP, las empresas tienen la posibilidad de poner en marcha procesos de optimización y de tomar decisiones en base a la situación de cada momento.

Cubos OLAP

Los cubos son elementos clave en OLAP, suministran un mecanismo para buscar datos con rapidez y tiempo de respuesta uniforme, independientemente de la cantidad de datos en el cubo o la complejidad del procedimiento de búsqueda. Los cubos son subconjuntos de datos de un almacén de datos, organizados y resumidos dentro de una estructura multidimensional. Los datos se resumen de acuerdo a factores de negocio seleccionados, proveyendo el mecanismo para un rápido y uniforme tiempo de respuesta a complejas consultas formuladas.

Para crear un cubo primeramente hay que definirlo, para lo cual se selecciona una tabla objetivo y las medidas (columnas numéricas de interés a los usuarios del cubo), dentro de esta tabla. También se seleccionan las dimensiones, compuestas de una o más columnas de otra tabla. Las dimensiones proveen la descripción categórica, por el cual las medidas son separadas para su análisis por los usuarios del cubo; luego se debe especificar la estrategia de resumen a utilizar, que puede definirse como la técnica que permite obtener partes de información clave a partir de una o más fuentes de información, lo cual constituye en una técnica de especial relevancia para tomar decisiones en un tiempo mínimo, diseñando las agregaciones (elementos precalculados de datos); y finalmente se procede a la carga del cubo para procesarlo; para esto último se hace uso de la herramienta seleccionada para la carga, consulta, filtrado y graficado de los cubos OLAP.
En un modelo de datos OLAP, la información es vista como cubos, los cuales consisten de categorías descriptivas (**dimensiones**) y valores cuantitativos (**medidas**). El modelo de datos multidimensional simplifica formular consultas complejas; arreglar datos en un reporte; cambiar de datos resumidos a datos detallados y filtrar o rebanar los datos en subconjuntos significativos a todos los usuarios. A continuación se explican algunos de estos conceptos importantes que se relacionan con las técnicas OLAP.

Las **dimensiones** son categorías descriptivas por los cuales los datos numéricos en un cubo, son separados para su análisis. El cubo puede expandirse para incluir otra dimensión; y también soporta la aritmética de matrices. Una dimensión puede ser creada para usarse en un cubo individual o en múltiples cubos. Una dimensión creada para un cubo individual es llamada dimensión privada. Por el contrario, si esta puede ser usada por múltiples cubos se le llama dimensión compartida; estas podrán ser usadas dentro de todo cubo en la base de datos; así se optimiza el tiempo y se evita el andar duplicando dimensiones privadas. Las dimensiones compartidas, también habilitan la estandarización de las métricas de negocios entre cubos. Por ejemplo, el estandarizar las dimensiones compartidas para el tiempo y localización geográfica, aseguran que los datos analizados, desde diferentes cubos, estén organizados similarmente.

Las **medidas** son datos numéricos de interés primario para los usuarios del cubo, estas son usadas por el procedimiento de agregación de los servicios de OLAP; y almacenadas para su rápida respuesta a las peticiones de los usuarios. Se puede crear una medida calculada y computar miembros de dimensiones, combinando expresiones multidimensionales (Multidimensional *Query eXpression*, MDX por sus siglas del inglés), fórmulas matemáticas y funciones definidas por el usuario. Se pueden registrar bibliotecas adicionales de estas funciones, para utilizarse en la definición de miembros calculados.

1.9 Modelo de datos

En las bases de datos tradicionales como en el almacén de datos existen tres niveles de modelado de datos: conceptual, lógico y físico.

Modelo de datos conceptual

El modelo conceptual brinda una información general del negocio, representando las entidades del dominio del problema y sus relaciones. Refleja más el espacio del problema, que el espacio de la solución.

Modelo de datos lógico

Es un puente intermedio entre el modelo conceptual y el físico. Generalmente incluye todas las entidades y las relaciones entre ellas, atributos, tipos de datos y llaves, sin tener en cuenta cómo serán posteriormente implementados físicamente en la base de datos. Posee un alcance mayor que cubre las áreas y los procesos más importantes.

Modelo de datos físico

Describe las estructuras de almacenamiento y los métodos usados para tener un acceso efectivo de los datos.

1.10 Definición de la metodología

Uno de los retos más importantes en el mundo de la Informática es crear un producto con buena calidad y además con un costo reducido. Con el propósito de mejorar los indicadores en el proceso de desarrollo de un software, se crean las metodologías de desarrollo, estas son usadas para estructurar, planear y controlar el proceso de creación de un software. Con el transcurso de los años se han desarrollado nuevas metodologías y mejorado las existentes. Entre las más utilizadas se encuentran:

1.10.1 Metodología HEFESTO:

Es una metodología propia, cuya propuesta está fundamentada en una muy amplia investigación, comparación de metodologías existentes, experiencias propias en procesos de confección de un almacén de datos. Se caracteriza por basarse en los requerimientos de los usuarios, por lo cual su estructura es capaz de adaptarse con facilidad y rapidez ante los cambios en el negocio; ser independiente de las herramientas que se utilicen para su implementación y reducir la resistencia al cambio, ya que involucra a los usuarios finales en cada etapa para que tome decisiones respecto al comportamiento y funciones de un almacén de datos.

A pesar de las facilidades que no va a ser la seleccionada a utilizar ya que no está completa y se necesita seleccionar una metodología que este completa, que sea abarcadora, especifica y bien detallada, que tenga bien definido el ciclo de vida y las tareas a desarrollar. (17)

1.10.2 Metodología de Inmon

Se basa en conceptos bien conocidos del diseño de bases de datos relacionales. Está enfocada de forma descendente. Tiene un enfoque empresarial para la construcción de un almacén de datos. Sus principales características respecto a los almacenes de datos son: Orientado a temas, Integrado, No volátil y Variables en el tiempo.

No va a ser la seleccionada ya que va de lo macro a lo micro o sea se referencia normalmente como Top-Down. (18), (19)

1.10.3 Metodología de Kimball

Proporciona un enfoque de abajo hacia arriba (Bottom-up), muy versátil, y una serie de herramientas prácticas que ayudan a la implementación de un DW. Permite implementar pequeños mercado de datos en áreas específicas. El ciclo de vida está basado en cuatro principios básicos:

- **Centrarse en el negocio:** Hay que concentrarse en la identificación de los requerimientos del negocio y su valor asociado, y usar estos esfuerzos para desarrollar relaciones sólidas con el negocio, agudizando el análisis del mismo y la competencia consultiva de los implementadores. (20)
- **Construir una infraestructura de información adecuada:** Diseñar una base de información única, integrada, fácil de usar, de alto rendimiento donde se reflejará la amplia gama de los requisitos de negocio identificados en la empresa. (20)
- **Realizar entregas en incrementos significativos:** Crear el almacén de datos en incrementos entregables. Usar el valor de negocio de cada elemento identificado para determinar el orden de aplicación de los incrementos. (20)
- **Ofrecer la solución completa:** Proporcionar todos los elementos necesarios para entregar valor a los usuarios de negocios. Esto significa tener un almacén de datos sólido, bien diseñado, con calidad probada, y accesible. También se deberá entregar herramientas de consulta ad hoc, aplicaciones para informes y análisis avanzado, capacitación, soporte, sitio web y documentación. (20)

Propone tres caminos que se enfocan en diferentes áreas, los cuales son:

Camino superior o tecnológico: indica tareas relacionadas con un software específico.

Camino medio o de datos: se diseña e implementa el modelo dimensional y se desarrolla el proceso ETL.

Camino inferior o de aplicaciones de inteligencia de negocio: se aplica la inteligencia de negocio para los usuarios finales.

Figura 5: Fases del ciclo de vida.

El presente trabajo se enmarca en las etapas siguientes: Planificación, Análisis de requerimientos, Modelado Dimensional, Diseño Físico**,** Diseño e implementación de ETL y Especificación y desarrollo de aplicaciones de inteligencia de negocio.

Planificación

La planificación del proyecto del almacén de datos, es una de las etapas que involucra todo el ciclo de vida de la metodología; desde su inicio hasta su fin. En esta etapa se determinan los objetivos específicos y la meta a alcanzar con el desarrollo del almacén de datos, así como la necesidad de información de la empresa. (5)

Análisis de requerimientos

La definición de los requerimientos es el proceso de entrevistar al personal de negocio y técnico, pero siempre conviene tener un poco de preparación previa. Se debe aprender tanto como se pueda sobre el negocio, los competidores, la industria y los clientes del mismo. Hay que leer todos los informes posibles de la organización; rastrear los documentos de estrategia interna; entrevistar a los empleados, analizar lo que se dice en la prensa acerca de la organización, la competencia y la industria. Se deben conocer los términos y la terminología del negocio. (5) A partir de las entrevistas se identifican los temas analíticos y los procesos de negocios.

Modelado Dimensional

En este paso es donde se diseña el modelo lógico del almacén de datos, se establece el nivel de granularidad, se eligen las dimensiones, se identifican las tablas de hechos. (5)

Diseño Físico

El diseño físico se focaliza sobre la selección de las estructuras necesarias para soportar este diseño. En esta etapa se decide la cantidad de índices a tener en cuenta, el tipo de índice y donde se guardan los mismos. También se especifica el tipo de particionamiento y el rango de este, cuando se utilizan las llaves únicas. Finalmente, se definen los nombres que tendrán los índices, las particiones y las llaves. (5)

Diseño e implementación de ETL

El sistema de Extracción, Transformación y Carga es la base sobre la cual se alimenta el almacén de datos. Si el sistema ETL se diseña adecuadamente, da la posibilidad de extraer los datos de los sistemas de origen de datos, aplicar diferentes reglas para aumentar la calidad y consistencia de los mismos, consolidar la información proveniente de distintos sistemas, y finalmente cargar (grabar) la información en el almacén de datos en un formato acorde para ser utilizado con las herramientas de análisis. (5)

Especificación y desarrollo de aplicaciones de inteligencia de negocio

Son utilizadas para consultar, analizar y presentar información dirigida a apoyar las necesidades del negocio. Brindan al usuario un acceso rápido de la información almacenada mediante reportes previamente definidos.

Se selecciona la metodología de Kimball ya que es la más abarcadora, tiene bien detallado el ciclo de vida y las tareas. Proporciona una base empírica y metodológica adecuada para las implementaciones de un almacén de datos, dada su gran versatilidad y su enfoque ascendente, que permite construir el almacén de datos en forma escalonada.

1.11 Herramientas a utilizar

1.11.1 Selección del gestor de la base de datos

Un sistema gestor de bases de datos (SGBD) es un software informático especializado y disponible en el mercado que se utiliza para la creación, acceso, control, y gestión de las bases de datos. (21)

Los gestores de bases de datos más conocidos son: MySQL, PostgreSQL, Oracle y SQL Server.

Oracle: es básicamente una herramienta cliente/servidor para la gestión de bases de datos. Consume gran cantidad de recursos en el servidor, lo que trae consigo que se necesite de un buen sistema de hardware. Cuenta con herramientas necesarias para su

utilización así como con herramientas de programación básicas de Oracle. (22) Entre sus principales características se encuentra su seguridad, garantizando la autenticidad apropiada de los usuarios y la privacidad e integridad de la información. Posee lectura de multiversión, proporcionándoles a los usuarios respuestas consistentes.

Oracle es configurable en ambientes OLTP (Procesamiento de Transacciones en línea), paralelos, clúster, también es una buena solución a nivel de almacén de datos. (23) Se mantiene actualmente como una solución privativa, lo que conlleva a pagar altos costos para adquirir la licencia.

Microsoft SQL Server: es un sistema gestor de base de datos relacionales producido por Microsoft. Es un sistema cliente/servidor que funciona como una extensión natural del sistema operativo Windows. Entre otras características proporciona integridad de datos, optimización de consultas, recuperación, control de concurrencia y backup. Es relativamente fácil de administrar a través de la utilización de un entorno gráfico para casi todas las tareas de sistema y administración de bases de datos. Utiliza servicios del sistema operativo Windows para ofrecer nuevas capacidades o ampliar la base de datos.

SQL Server es fácil de usar y proporciona funciones de almacenamiento de datos que sólo estaban disponibles en Oracle y otros sistemas gestores de bases de datos. (24) Entre sus principales desventajas se encuentran que el costo de la licencia puede ser alto comparado con otros competidores (25) y está atado a la plataforma del sistema operativo sobre la cual se instala, por lo que lo convierte en un software que no puede ser ejecutado en otras plataformas excepto Windows.

MySQL: Es un sistema de gestión de bases de datos relacional, fue creada por la empresa sueca MySQL AB, la cual tiene el copyright del código fuente del servidor SQL, así como también de la marca. MySQL es un software de código abierto licenciado bajo la GPL de la GNU, aunque MySQL AB distribuye una versión comercial, en lo único que se diferencia de la versión libre, es en el soporte técnico que se ofrece y la posibilidad de integrar este gestor en un software propietario, ya que de otra manera se vulneraría la licencia GPL.

El principal objetivo de MySQL es la velocidad y robustez. Se destacan en la herramienta sus principales características: soporta gran cantidad de tipos de datos para las columnas, gran portabilidad entre sistemas, puede trabajar en distintas plataformas y sistemas operativos, aprovecha la potencia de sistemas multiproceso gracias a su implementación multihilo. Como principales desventajas presenta que no es intuitivo

respecto a otros sistemas gestores existentes, y que un gran porcentaje de sus utilidades no están documentadas. (26)

PostgreSQL: es un sistema de gestión de bases de datos objeto-relacionales (ORDBMS) que ha sido desarrollado de varias formas desde 1977. Comenzó como un proyecto denominado Ingres en la Universidad Berkeley de California. En 1986 otro equipo continuó el desarrollo del código de Ingres para crear un sistema de bases de datos objeto -relacionales llamado Postgre. En 1996, debido a un nuevo esfuerzo de código abierto y a la incrementada funcionalidad del software, Postgre fue renombrado a PostgreSQL.

El proyecto PostgreSQL sigue actualmente un activo proceso de desarrollo a nivel mundial gracias a un equipo de desarrolladores y contribuidores de código abierto. Está considerado como el sistema de bases de datos de código abierto más avanzado del mundo. (27)

Cuenta con una documentación muy bien organizada, pública y libre y es altamente adaptable a las necesidades del cliente. Soporta todas las características de una base de datos profesional entiéndase triggers, funciones, secuencias, relaciones, reglas, tipos de datos definidos por usuarios, vistas materializadas. Sus características técnicas lo convierten en un sistema gestor de base de datos muy potente y robusto en el mercado. Estabilidad, potencia, robustez, facilidad de administración e implementación de estándares han sido las características que más se han tenido en cuenta durante su desarrollo a lo largo de los años. (28)

Luego del análisis realizado a los principales gestores de bases de datos, se decide utilizar el PostgreSQL, ya que es fiable, tiene la capacidad de soportar múltiples plataformas (multiplataforma), y su licencia está libre por los fabricantes. Tiene gran velocidad en sus transacciones, rendimiento, facilidad de administración y conexión con otros productos, que sean software bien documentados y con perspectivas de evolución. Además posee varias herramientas gráficas para la administración de las bases de datos.

1.11.2 Suite Pentaho

Es una plataforma de BI (Business Intelligence) “orientada a la solución” y “centrada en procesos”. Es una herramienta desarrollada bajo la filosofía del software libre para la gestión y toma de decisiones empresariales. Está compuesta de diferentes programas que satisfacen los requisitos de BI y ha sido desarrollada bajo el lenguaje de

programación Java. Ofreciendo soluciones para la gestión y análisis de la información, incluyendo el análisis multidimensional OLAP, presentación de informes, minería de datos y creación de cuadros de mando para el usuario, por lo que constituye una solución factible para cualquier empresa que quiera hacer BI en su organización. (29)

Pentaho Schema Workbench (PSW)

Es una herramienta de desarrollo que permite crear, modificar y publicar un esquema de Mondrian. Está muy orientada al desarrollador conocedor de la estructura de un esquema de Mondrian. Permite crear todos los objetos que soporta Mondrian: esquema, cubo, dimensiones, métricas. El motor de Mondrian procesa las consultas MDX (Expresiones Multidimensionales) utilizando ROLAP. Trabaja con esquemas que utilizan simplemente ficheros XML, que contienen toda la información acerca de los datos que serán utilizados por Mondrian para formar la estructura del cubo. (30)

Pentaho Data Integration (Kettle)

Incluye un conjunto de herramientas para implementar el proceso ETL. Uno de sus objetivos es que el proceso ETL sea fácil de general, mantener y desplegar. Se compone de 4 herramientas:

- **Spoon:** es la herramienta gráfica que nos permite el diseño de las transformaciones y trabajos. Incluye opciones para pre-visualizar y testear los elementos desarrollados. Es la principal herramienta de trabajo de PDI y con la que construiremos y validaremos nuestros procesos ETL.
- **Pan:** es la herramienta que nos permite la ejecución de las transformaciones diseñadas en Spoon (bien desde un fichero o desde el repositorio). Nos permite desde la línea de comandos preparar la ejecución mediante scripts.
- **Kitchen:** similar a Pan, pero para ejecutar los trabajos o jobs.
- **Carte:** es un pequeño servidor web que permite la ejecución remota de transformaciones y jobs.

Entre sus características principales se destaca que está basado en metadatos y que es una aplicación escrita en Java con algunas características avanzadas en escritas en Java Script. (31)

Pentaho Metadata Edition

Es una herramienta que pertenece a la Suite Pentaho, la cual es utilizada para crear los metadatos del mercado de datos, permitiendo realizar consultas Ad - Hoc.

Pentaho Report Designer

Es una herramienta de reporte que permite crear informes, los cuales pueden ser ejecutados o publicados en la plataforma BI y a su vez puedan ser utilizados por los usuarios. Permite trabajar con múltiples orígenes de datos (JDBC, Olap4J, Pentaho Analysis, Pentaho Data Integration, XML) incluido el metadatos que se tenga definido en el sistema. (32)

Servidor OLAP Mondrian

Mondrian es una de las aplicaciones más importantes de la plataforma Pentaho BI. Mondrian es un servidor OLAP open source que gestiona comunicación entre una aplicación OLAP (escrita en Java) y la base de datos con los datos fuente. Es decir, Mondrian actúa como "JDBC (Java Database Connectivity) para OLAP". (33)

Hay cinco tipos de acciones de servidor.

Dirección URL: se utiliza normalmente para mostrar una dirección URL (*Uniform Resource Locator*, localizador uniforme de recursos) en un explorador. Por ejemplo, se puede hacer clic en el nombre de un cliente y, a continuación, mostrar el perfil de ese cliente en una página Web.

Informe: genera un informe basado en direcciones URL con la siguiente sintaxis:

http://<nombre del servidor de informes>/<directorio del servidor>

Conjunto de filas: devuelve un conjunto de filas basado en una condición especificada en una instrucción MDX. Esta acción resulta útil para vincular un conjunto de filas a datos en un cubo distinto, pero relacionado, en la misma base de datos OLAP, como las ventas de almacén de hace años. Los datos se muestran en una hoja de cálculo como una tabla de Excel.

Nivel de detalle: envía una consulta basada en una condición especificada en una instrucción MDX y devuelve datos de la tabla de administración empresarial, que es el origen de los valores totales de la medida. Los datos se muestran en una hoja de cálculo como una tabla de Excel

Expandir al detalle: esta acción envía una consulta al cubo basada en una condición especificada en una instrucción MDX y devuelve datos de la tabla de administración empresarial, que es el origen de los valores totales de la medida. (34)

1.11.3 SQL-Power-Architect

Es una herramienta de modelado de datos que fue creada por los diseñadores de almacenamiento de datos y tiene muchas características únicas dirigidas específicamente para el arquitecto de almacenamiento de datos. Permite a los usuarios de la herramienta realizar ingeniería inversa de bases de datos existentes, realizar perfiles de datos en bases de datos de origen y generar automáticamente los metadatos de ETL. Es una herramienta ideal para grupos de desarrollo donde se pude realizar el modelado de datos y poder así tener documentado el modelo de datos de todas las aplicaciones que se desarrollan. (35)

1.12 Conclusiones

Se realizó un estudio del funcionamiento del trabajo con datos bibliográficos con el objetivo de conocer cómo se gestiona la información para el proceso de toma de decisiones. Se definió la necesidad de desarrollar un almacén de datos que apoye la toma de decisiones para el análisis de datos bibliográficos. La tecnología adecuada para resolver la problemática expuesta es el mercado de datos. Para su desarrollo se decidió utilizar la metodología de Kimball.

Entre las herramientas seleccionadas está el sistema gestor de base de datos PostgreSQL. La suite Pentaho seleccionando las siguientes herramientas: Pentaho Schema Workbench para la gestión de los esquemas Mondrian, Pentaho Report Designer para los reportes, Pentaho Metadata-edition para generar los metadatos, y para el modelado de datos el software SQL-Power-Architect. Seleccionando para el diseño físico el esquema estrella y para el procesamiento analítico en línea se escogió el modelo ROLAP.

CAPÍTULO 2: Análisis y diseño del mercado de datos.

2.1 Introducción

En este capítulo se definirán los requisitos informacionales con los cuales el mercado de datos debe cumplir, realizándose una descripción de cada uno, estos se identificarán usando las técnicas definidas para ello. Se definirán y describirán las tablas de dimensiones, las tablas de hechos y las medidas necesarias para realizar el diagrama modelo de datos, se explicará además la tipología a utilizar para su confección.

2.2 Identificación de los temas analíticos

Los temas analíticos agrupan características comunes en un tema común (20). Permiten que los datos se organicen por temas para facilitar su acceso y entendimiento por parte de los usuarios finales.

2.2.1 Caracterización de los temas analíticos

El análisis de datos bibliográficos se encarga de medir la productividad científica de las publicaciones realizadas. A partir de un análisis realizado se identificaron los siguientes temas analíticos que agrupan requisitos en temas comunes.

- **Productividad científica:** está enfocado en medir cuantitativamente la productividad científica.
- **Predicción de tendencias:** está enfocado en predecir las tendencias de publicaciones científicas.

2.3 Procesos de negocio

Un proceso de negocio es un conjunto de tareas relacionadas lógicamente llevadas a cabo para lograr un resultado de negocio definido. Cada proceso de negocio tiene sus entradas, funciones y salidas. Las entradas son requisitos que deben tenerse antes de que una función pueda ser aplicada. Cuando una función es aplicada a las entradas de un método, tendremos ciertas salidas resultantes.

Se definieron para el presente trabajos los procesos de negocio productividad y predicción.

- **Productividad:** comprende la cantidad de publicaciones de un autor, de una editorial o de un país, además de la cantidad de publicaciones sobre una disciplina o para determinar el mes o año en que más se ha publicado, etc.

- **Predicción:** comprende el posible descarte de publicaciones o determinar las tendencias de publicación de una disciplina.

2.4 Definición de los requisitos del negocio

Actualmente representan un punto clave en el desarrollo de cualquier software. Para que un software tenga éxito, se deben comprender detalladamente los requisitos del negocio, en caso contrario el producto puede llegar a fracasar.

2.4.1 Técnicas de identificación de requisitos

Para realizar una correcta identificación de los requisitos y un adecuado análisis del problema a resolver se pueden implementar una serie de técnicas de acuerdo al cliente con que se esté tratando. Esto está relacionado con la comprensión de los procesos de toma de decisiones y los objetivos que el cliente pretende alcanzar con la información suministrada. Apoyando a esta etapa la metodología de Kimball propone algunas técnicas para poder identificar las necesidades de manera tal que se satisfagan las expectativas del cliente. Para la captura de requisitos del mercado de datos fueron usadas las siguientes técnicas:

Análisis de documentación: Se realizó un estudio de las leyes y metodologías que rigen a la Cienciometría y que están relacionados con la Bibliometría, así como de sus conceptos y aplicaciones, con el objetivo de lograr comprender el negocio.

Entrevista: Se realizó un conjunto de entrevistas a los especialistas en Ciencia de la Información del Instituto de Cibernética, Matemática y Física (ICIMAF), permitiendo recopilar información, criterios y opiniones que aportaron una serie de elementos esenciales para la definición de los requisitos.

Para realizar las entrevistas se utilizaron las preguntas que propone la metodología de Kimball. Una muestra de estas preguntas se encuentra en la **figura 1.1**. (Ver anexos)

2.4.2 Requisitos informacionales

Los requisitos de información describen los datos que se deben almacenar en el sistema para entonces poder satisfacer las necesidades del cliente. Atendiendo a las necesidades para la construcción del mercado de datos se definieron los siguientes requisitos informacionales:

Productividad:

RF 1. Determinar la productividad dado un país.

RF 2. Determinar productividad dada una institución.

RF 3. Determinar productividad dado un autor.

RF 4. Determinar la productividad de los autores por trimestre.

RF 5. Identificar los autores que más publican dada una disciplina.

RF 6. Determinar número de trabajos publicados sobre una disciplina.

RF 7. Identificar disciplina más publicada según mes o año.

RF 8. Determinar autores que han publicado dado un tipo de publicación.

Predicción:

RF 9. Predecir las tendencias de disciplina de las publicaciones.

RF 10. Adaptar políticas de descarte de publicaciones.

2.5 Especificación de los requisitos informacionales

La metodología de Kimball establece la siguiente tabla para la especificación de los requisitos, la cual está compuesta por los siguientes campos: número del requisito, nombre del requisito, descripción, campos asociados, tipo de dato de los campos, reglas o restricciones y la prioridad.

<table>
<tr><td rowspan="5">No.1</td><td colspan="2">Nombre</td><td colspan="2">Descripción</td></tr>
<tr><td colspan="2">Determinar la productividad dado un país.</td><td colspan="2">Muestra la cantidad de publicaciones realizadas por autores dado un país.</td></tr>
<tr><td>Campos</td><td>Tipo dato</td><td>Reglas</td><td>Prioridad</td></tr>
<tr><td>nombre_autor
pais_autor
cant_public</td><td>cadena de caracteres
cadena de caracteres
numérico</td><td>No procede</td><td>Alta</td></tr>
<tr><td></td></tr>
<tr><td rowspan="2">No.2</td><td colspan="2">Nombre</td><td colspan="2">Descripción</td></tr>
<tr><td colspan="2">Determinar productividad dada una institución.</td><td colspan="2">Muestra la cantidad de publicaciones realizadas por una institución editora determinada.</td></tr>
</table>

	Campos	**Tipo dato**	**Reglas**	**Prioridad**
	cant_public nombre_inst	numérico cadena de caracteres	No procede	Alta
No.3	**Nombre**		**Descripción**	
	Determinar productividad dado un autor.		Muestra la cantidad de publicaciones realizadas por un autor.	
	Campos	**Tipo dato**	**Reglas**	**Prioridad**
	nombre_autor cant_public	cadena de caracteres numérico	No procede	Alta
No.4	**Nombre**		**Descripción**	
	Determinar la productividad de los autores por trimestre.		Muestra la cantidad de publicaciones realizada por un autor en el trimestre seleccionado.	
	Campos	**Tipo dato**	**Reglas**	**Prioridad**
	nombre_autor no_trimestre cant_public	cadena de caracteres cadena de caracteres numérico	No procede	Alta
No.5	**Nombre**		**Descripción**	
	Identificar los autores que más publican dada una disciplina.		Al seleccionar una disciplina muestra los autores que más publican en esta.	
	Campos	**Tipo dato**	**Reglas**	**Prioridad**
	disciplina nombre_autor cant_public	cadena de caracteres cadena de caracteres numérico	No procede	Alta

<table>
<tr><td rowspan="4">No.6</td><td colspan="2">Nombre</td><td colspan="2">Descripción</td></tr>
<tr><td colspan="2">Determinar número de trabajos publicados sobre una disciplina.</td><td colspan="2">Al seleccionar una disciplina muestra la cantidad de trabajos que han sido publicados.</td></tr>
<tr><td>Campos</td><td>Tipo dato</td><td>Reglas</td><td>Prioridad</td></tr>
<tr><td>disciplina
cant_public</td><td>cadena de caracteres
numérico</td><td>No procede</td><td>Alta</td></tr>
<tr><td rowspan="4">No.7</td><td colspan="2">Nombre</td><td colspan="2">Descripción</td></tr>
<tr><td colspan="2">Identificar disciplina más publicada según mes o año.</td><td colspan="2">Muestra cuales son las disciplinas en la que más publican los autores de acuerdo a un mes o año seleccionado.</td></tr>
<tr><td>Campos</td><td>Tipo dato</td><td>Reglas</td><td>Prioridad</td></tr>
<tr><td>disciplina
cant_public
nombre_autor
mes
anno</td><td>cadena de caracteres
numérico
cadena de caracteres
cadena de caracteres
cadena de caracteres</td><td>No procede</td><td>Alta</td></tr>
<tr><td rowspan="4">No.8</td><td colspan="2">Nombre</td><td colspan="2">Descripción</td></tr>
<tr><td colspan="2">Determinar autores que han publicado dado un tipo de publicación.</td><td colspan="2">Al seleccionar un tipo de publicación muestra el nombre de los autores que han realizado publicaciones relacionadas con el tipo de publicación deseado.</td></tr>
<tr><td>Campos</td><td>Tipo dato</td><td>Reglas</td><td>Prioridad</td></tr>
<tr><td>tipo_public
nombre_autor</td><td>cadena de caracteres
cadena de caracteres</td><td>No procede</td><td>Alta</td></tr>
</table>

No.9	**Nombre**		**Descripción**	
	Predecir las tendencias de disciplina de las publicaciones.		Muestra cuales son las disciplinas en las que más han publicado los autores.	
	Campos	**Tipo dato**	**Reglas**	**Campos**
	disciplina cant_public	cadena de caracteres numérico	No procede	Alta
No.10	**Nombre**		**Descripción**	
	Adoptar políticas de descarte de publicaciones.		Muestra cuales son las disciplinas en las que menos se ha publicado.	
	Campos	**Tipo dato**	**Reglas**	**Prioridad**
	disciplina cant_public	cadena de caracteres numérico	No procede	Alta

Tabla 1: Descripción de los requisitos informacionales.

Requerimientos no Funcionales:

Los requisitos no funcionales son propiedades o cualidades que el producto debe tener. Se definieron los siguientes requisitos no funcionales para la construcción del mercado de datos.

1. Rendimiento:

- Los tiempos de respuesta a las consultas que generen reportes, recorran la base de datos, o interactúen con el tablero de mando no deben exceder de 30 segundos.

2. Seguridad:

- El sistema está definido para operar conjuntamente con el sistema del control del acceso (*Administration Console*) y garantizar de esta forma el acceso a los datos solamente a los administradores y usuarios autorizados.

3. Accesibilidad:

- El sistema debe estar disponible para todos los usuarios en cualquier momento.

4. Usabilidad:

- El sistema deberá ser sencillo, flexible y de fácil uso.

- El sistema debe poder ser usado por cualquier persona que posea conocimientos básicos de computación.

5. Software:

En los servidores:

- Sistema operativo Debian 6 o superior, Ubuntu Server 12.04 o superior, o CentOS 5 o superior.
- PostgreSQL 9.0 o superior.
- Maquina virtual de Java 1.6 o superior.

En los clientes:

- Sistema operativo Microsoft Windows XP o superior, o cualquier distribución GNU/Linux.
- Navegador web Firefox versión 15 o superior, o cualquier otro navegador compatible.

6. Hardware:

En los servidores (requerimientos mínimos):

- Procesador Dual-Core a 1 GHz.
- 2 GB de memoria RAM.
- 10 GB de Disco Duro.

En los clientes:

- Procesador P-IV a 1 GHz.
- 512 MB de memoria RAM.

2.6 Matriz dimensional

Se considera necesario utilizar una matriz para poder representar la relación entre las dimensiones y los procesos del negocio.

	Procesos	
Dimensiones	**Productividad**	**Predicción**
Editorial	X	X
Institución rectora	X	X
Tiempo	X	X
Autor	X	
Palabras claves	X	X
Revistas	X	
Eventos	X	
Tipo _ Publicaciones	X	
Publicacion	X	X

Tabla 2: Matriz dimensional.

2.7 Validación de requisitos

Luego de haber realizado la definición de los requisitos, se hace necesario validarlos ya que de esta forma se eliminan ambigüedades y se confirma que estos cumplan sus verdaderos propósitos, logrando un producto con mayor calidad. La metodología seleccionada propone como técnica de validación de requisitos la confección de un prototipo funcional, el cual consiste en seleccionar un proceso del negocio y realizar con el mismo una pequeña versión de un mercado de datos, con el objetivo de mostrar las facilidades que brinda la herramienta y determinar si cumple con los requisitos definidos. Para realizar el prototipo se utilizó como herramienta de apoyo un ejemplo que trae implementado la plataforma Pentaho.

2.8 Modelado dimensional

El modelado dimensional es el modo de acercar los datos a la manera en que estos serán convertidos en información útil para los usuarios del negocio. El objetivo es que estos

puedan encontrar de manera intuitiva y rápida la información que necesitan. El proceso de diseño comienza con el modelado dimensional de alto nivel que se obtiene a partir de los procesos que se encuentran en la matriz dimensional y cuenta con varios pasos para su construcción, elegir el proceso de negocio, establecer el nivel de granularidad, identificar las dimensiones y luego la tabla de hechos y sus medidas.

2.8.1 Identificación de los niveles de granularidad del mercado de datos

La granularidad representa el nivel de detalle con el que se desea almacenar la información sobre el negocio analizado. Mientras mayor sea el nivel de detalle de los datos, se tendrán mayores posibilidades analíticas; ya que los mismos podrán ser resumidos. Es decir, los datos que posean granularidad fina (nivel de detalle) podrán ser resumidos hasta obtener una granularidad media o gruesa. No sucede lo mismo en sentido contrario, ya que los datos almacenados con granularidad media podrán resumirse, pero no tendrán la facultad de ser analizados a nivel de detalle. En este paso se determina la variable que integrará cada perspectiva, actualizándose el diccionario de datos. Esta acción determinará la granularidad de la información encontrada en el mercado de datos.

Para todos los procesos el nivel de granularidad definida para el tiempo es el grano mensual. La granularidad específica para publicaciones está definida por la disciplinas. La granularidad de autores, eventos, instituciones y editorial está definida por el país.

2.8.2 Tablas de dimensiones

Las tablas de dimensiones son elementos que contienen atributos utilizados para restringir y agrupar los datos que participan en el análisis. Las dimensiones poseen entre sus características principales la definición de jerarquías entre sus atributos, que tienen como objetivo plasmar explícitamente la forma en que se puede consolidar. Son las compañeras integrales de las tablas de hechos, ellas contienen la descripción textual del negocio. (36)

Las dimensiones identificadas fueron las siguientes:

Dim_Tiempo: Contiene los datos fundamentales referentes a la fecha denominada por el mes y año, además del número del trimestre en el que se encuentre. Integrado por los siguientes atributos: mes, año y número de trimestre.

Dim_Editorial: Contiene la información referente a la editorial. Esta información hace referencia a los siguientes atributos: nombre_edit y pais_edit.

Dim_Institucion_rectora: Guarda la información referente a la institución principal, permitiendo conocer los siguientes atributos: nombre_inst y pais_inst.

Dim_Revistas: Contiene los datos fundamentales referentes a la revista, permitiendo conocer los atributos: nombre_rev, editorial_rev, nivel_rev y acceso_rev.

Dim_Eventos: Contiene la información referente a los eventos que se realizan, permitiendo conocer los atributos: nombre_event, tipo_event y pais_sede.

Dim_Publicacion: Permite conocer a información referente a la publicación, está compuesta por los atributos: título, disciplina, resumen.

Dim_Palabras_claves: Permite conocer la información referente a las palabras claves mediante el atributo: palabra_clave.

Dim_Autor: Contiene los datos fundamentales referentes a el autor. Integrado por los siguientes atributos: nombre_ autor, país_ autor, institución.

Dim_Tipo_Publicacion: Hace referencia a los tipos de publicaciones mediante el atributo tipo_public.

2.8.3 Tablas de hechos

Las tablas de hechos contienen los hechos que serán utilizados por los analistas de negocio para apoyar el proceso de toma de decisiones y entre su contenido están los datos cuantitativos. Esta se encuentra rodeada de las tablas de dimensiones.

Fact_Publicacion: Está compuesta por las llaves primarias de las dimensiones Dim_Tiempo, Dim_Editorial, Dim_Institucion_rectora, Dim_Revistas, Dim_Eventos, Dim_Tipo_Publicacion, Dim_Publicacion, Dim_Autor, Dim_Palabras_claves y la medida cant_public.

2.8.4 Modelo dimensional detallado

Luego de identificar los procesos del negocio y las dimensiones, se confecciona el modelo dimensional del mercado de datos.

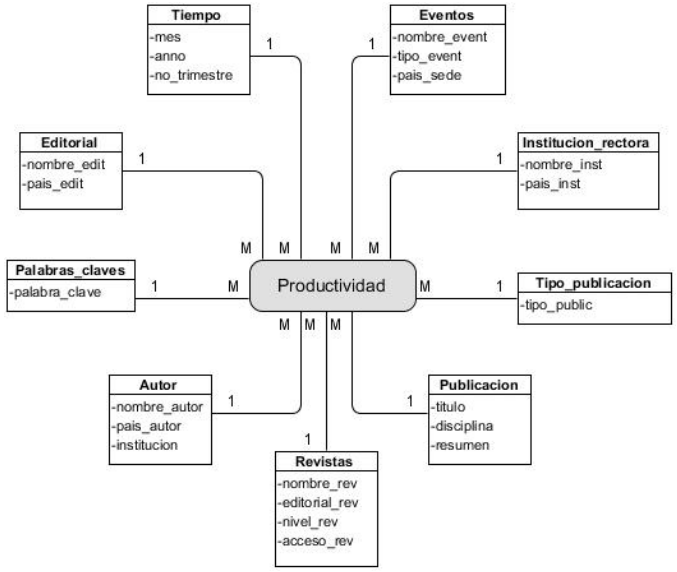


Figura 6: Modelo dimensional detallado del proceso Productividad.

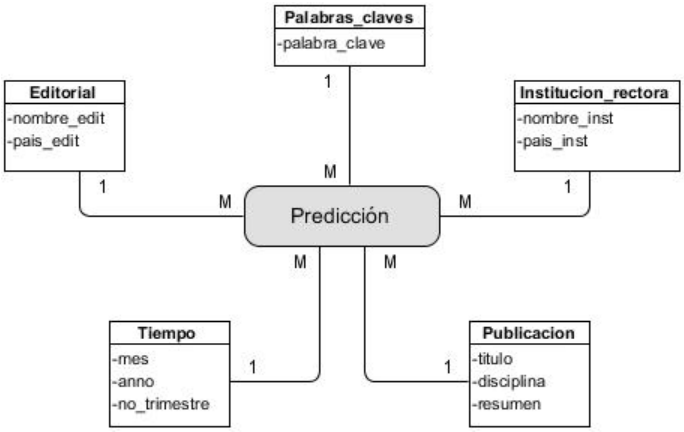


Figura 7: Modelo dimensional detallado del proceso Predicción.

2.8.5 Realizar uniones

Luego de haber diseñado las tablas de dimensiones y las de hechos a partir de los indicadores se realiza un modelo lógico previo, el cual expresa las correspondencias que existen entre dichas tablas. Una vez que se obtiene este modelo se analizan las tablas de dimensiones para verificar que no existan datos redundantes. De ser necesario se pueden unir tablas de dimensiones con el objetivo de eliminar la redundancia de datos y obtener el modelo lógico final. Al realizar las uniones se utilizaron los patrones de diseño llaves subrogadas y convenciones de nombre y tipo.

2.8.6 Unión de las tablas de dimensiones con la tabla de hechos.

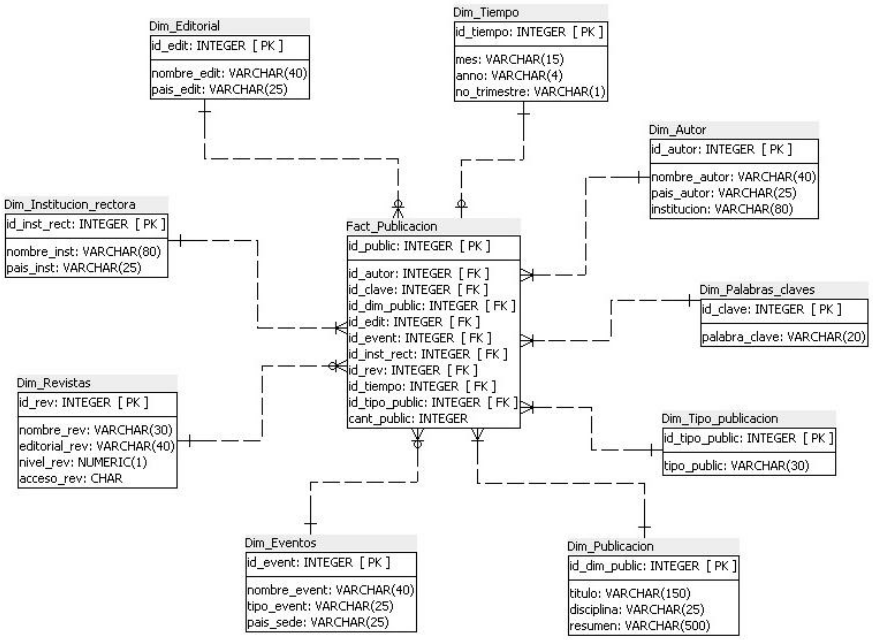


Figura 8: Modelo de datos.

2.9 Patrones de diseño

El uso de patrones para diseñar y construir el almacén de datos permite disminuir la complejidad de dicha acción y solucionar problemas que se presentan de modo recurrente. A la hora de realizar el diseño se identificaron los siguientes patrones:

Llaves subrogadas: este patrón se aplica para solucionar el problema de que los identificadores de las tablas de dimensiones no se repitan, dándole solución mediante la creación de secuencias, generándose una llave primaria para cada entidad, garantizando que el identificador de la entidad sea único.

Este patrón se evidencia en el modelo lógico del negocio en el identificador que presenta cada dimensión.

Naming and type conventions (convenciones de nombre y tipo): se utiliza para fijar prefijos seguidos de un guión bajo indicando su rol y función en el mercado de datos. Todas las dimensiones y la tabla de hechos reciben un prefijo de acuerdo a su función y las columnas de clave de dimensión llevan el nombre de la dimensión a la que pertenecen.

Este patrón se evidencia en el modelo lógico del negocio cuando se utiliza Dim seguido de un guión bajo y el nombre de la dimensión o de la tabla de hecho, ejemplo Dim_Autor y Fact_Publicacion.

2.10 Diseño del proceso ETL

Luego de haber construido el modelo dimensional se inicia el proceso de extracción, carga y transformación de los datos. Este proceso incluye la corrección de errores, borrado de campos que no es de interés para el usuario e incluso agregación de información. Este proceso es clave para lograr que los datos extraídos de diversos orígenes se integren finalmente en un mismo entorno. Para realizar este proceso se identifican las tablas en la base de datos de donde se extrae la información que permiten llenar los campos de las tablas dimensiones del mercado de datos.

Proceso ETL para la dimensión Dim_Institucion_rectora:

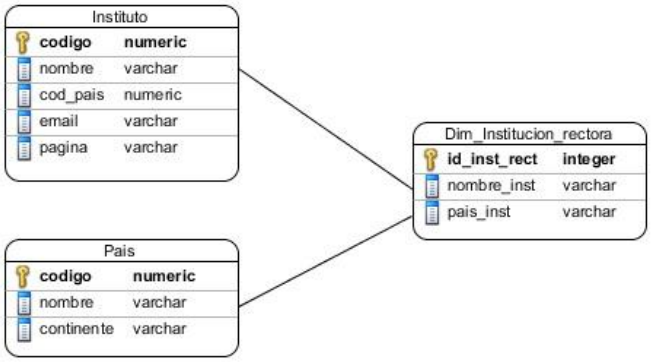


Figura 9: Diseño de la ETL la dimensión Dim_Institucion_rectora.

Para el llenado de la dimensión Dim_Institucion_rectora se identifican el atributo nombre perteneciente a la tabla Institución, y el atributo nombre perteneciente a la tabla Pais.

Proceso ETL para la dimensión Dim_Editorial:

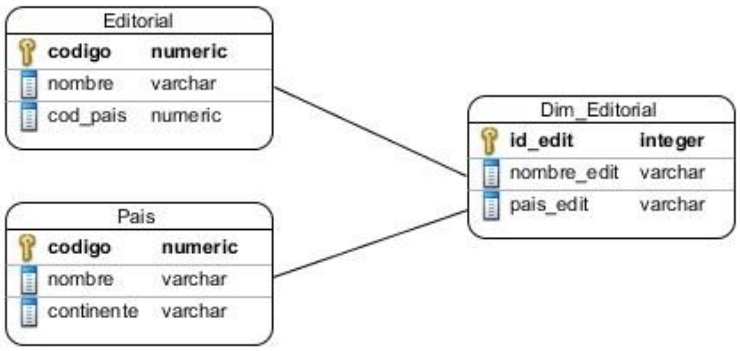


Figura 10: Diseño de la ETL la dimensión Dim_Editorial.

Para el llenado de la dimensión Dim_Editorial se identifican el atributo nombre de la tabla Editorial y el atributo nombre de la tabla País.

Proceso ETL para la dimensión Dim_Revistas:

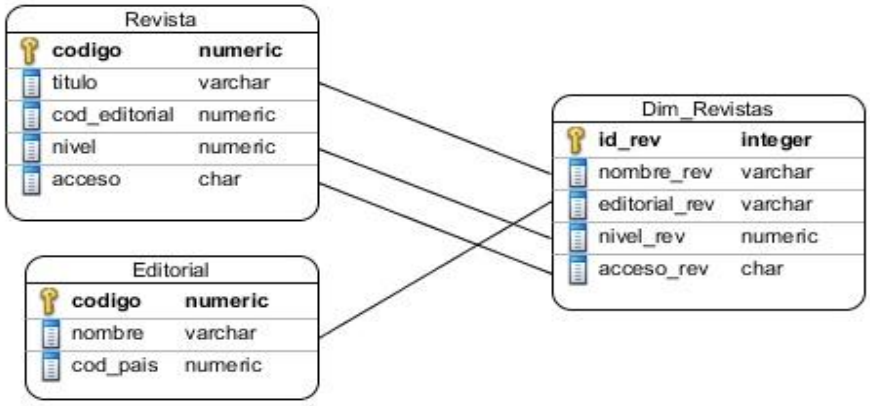


Figura 11: Diseño de la ETL la dimensión Dim_Revista.

Para el llenado de la dimensión Dim_Revistas se identifican los atributos título, nivel y acceso de la tabla Revista y el atributo nombre de la tabla Editorial.

Proceso ETL para la dimensión Dim_Eventos:

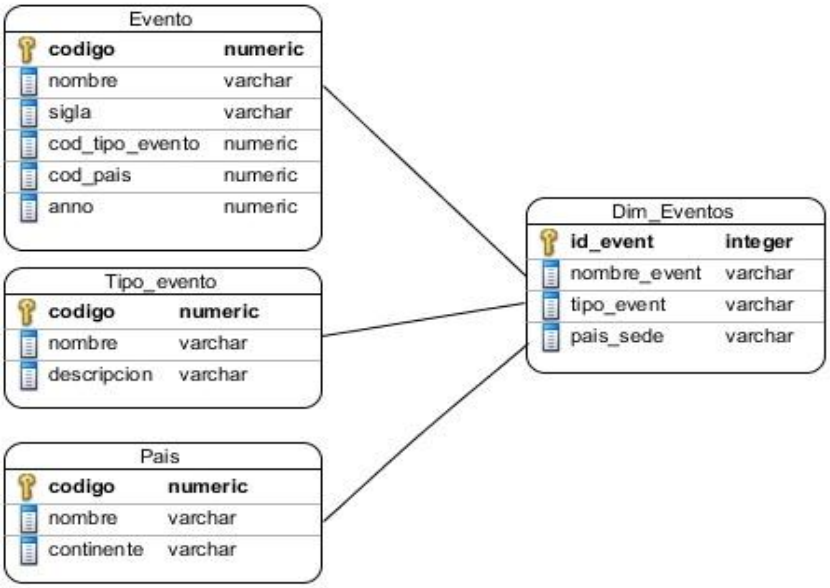


Figura 12: Diseño de la ETL la dimensión Dim_Eventos.

Para el llenado de la dimensión Dim_Eventos se identifican los atributos nombre de la tabla Evento, nombre de la tabla Tipo_evento y nombre de la tabla País.

Proceso ETL para la dimensión Dim_Autor:

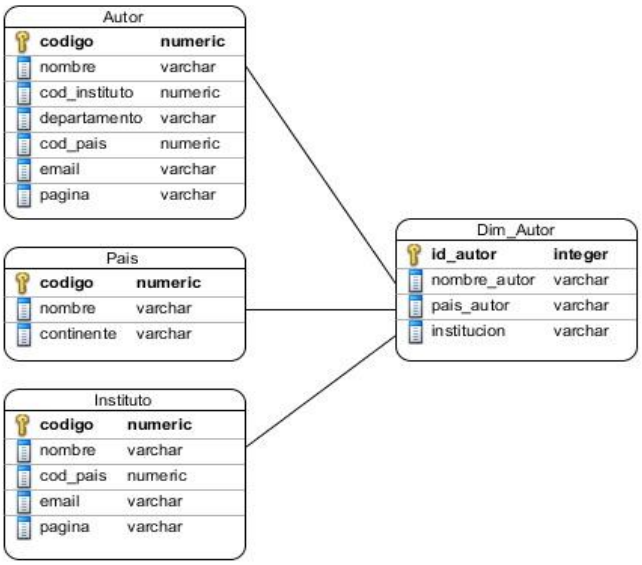


Figura 13: Diseño de la ETL la dimensión Dim_Autor.

Para el llenado de la dimensión Dim_Autor se identifican los atributos nombre de la tabla Autor, el atributo nombre de la tabla País y el atributo nombre de la tabla Institucion.

Proceso ETL para la dimensión Dim_Tiempo:

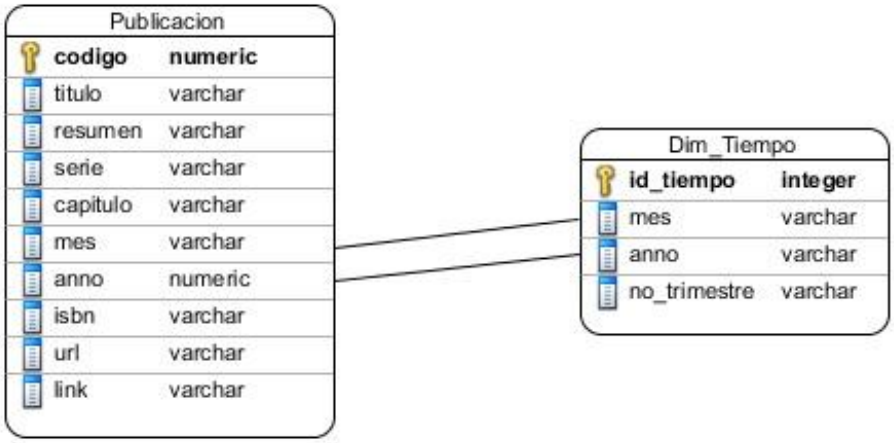


Figura 14: Diseño de la ETL la dimensión Dim_Tiempo.

Para el llenado de la dimensión Dim_Tiempo se identifican los atributos mes y anno de la tabla Publicacion.

Proceso ETL para la dimensión Dim_Palabras_claves:

Figura 15: Diseño de la ETL la dimensión Dim_Palabras_claves.

Para el llenado de la Dim_Palabras_claves se identifica el atributo nombre de la tabla Palabra_clave.

Proceso ETL para la dimensión Dim_Tipo_Publicacion:

Figura 16: Diseño de la ETL la dimensión Dim_Tipo_Publicacion.

Para el llenado de la dimensión Dim_Tipo_Publicacion se identifica el atributo nombre de la tabla Tipo_publicacion.

Proceso ETL para la dimensión Dim_Publicacion:

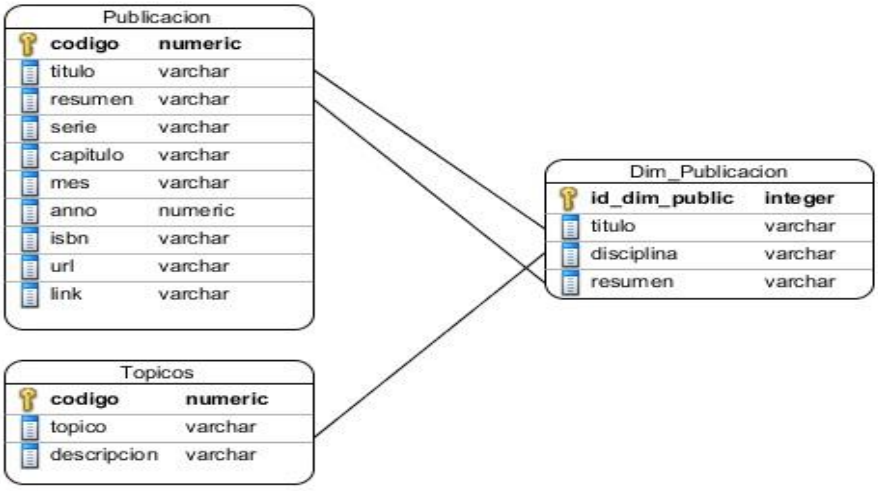


Figura 17: Diseño de la ETL la dimensión Dim_Publicacion.

Para el llenado de la dimensión Dim_Publicacion se identifica el atributo título y resumen de la tabla Publicacion y tópico de la tabla Topicos.

2.11 Conclusiones

Se definieron temas analíticos y dos procesos de negocio, los cuales conllevaron a la identificación y descripción de diez reglas del negocio para la creación del mercado de datos para el análisis de datos bibliográficos. Se modelaron nueve tablas de dimensiones y una tabla de hechos a partir de las perspectivas identificadas, llegando a un diseño del modelo lógico del negocio luego de realizar las uniones correspondientes entre las tablas de dimensiones y la tabla de hecho. Se diseñaron las ETL que conforman el mercado de datos.

CAPÍTULO 3: Evaluación del proceso de implementación del mercado de datos.

3.1 Introducción

En este capítulo se realiza la implementación del proceso ETL utilizando la herramienta Pentaho Data Integration y del cubo OLAP utilizando Pentaho Schema Workbench. Se implementan los reportes del mercado de datos, además, se definen y aplican las pruebas al mercado de datos con el objetivo de validar la solución propuesta, mediante la utilización de una lista de chequeo para obtener las *No Conformidades* que presente.

3.2 Implementación

Para implementar el mercado de datos se realiza el proceso ETL; luego de cargado el mercado de datos se procede a configurar el repositorio de metadatos, al concluir se generan los reportes correspondientes y se elabora el tablero de mando para visualizar la información almacenada.

3.2.1 Implementación de las ETL

Para la implementación de las ETL se realiza el proceso de extracción de los datos de las fuentes correspondientes, los cuales se copian en un área temporal para realizar su transformación. Luego se realiza la limpieza de los datos para darle tratamiento a los valores nulos y codificarlos. Finalmente se cargan los datos transformados en el mercado de datos.

Se implementaron diez ETL para la confección del mercado de datos para el análisis de datos bibliográficos. A continuación se muestra como se realizó la implementación.

- **Implementación de la ETL para la dimensión Autor:**

En el proceso de implementación de la ETL para la dimensión Autor, luego de haber creado el repositorio damos doble click en la carpeta INPUT, seleccionamos el componente Table input además de los componentes mostrados en la siguiente figura y se arrastran hasta la zona de diseño.

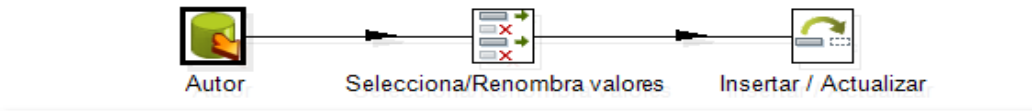


Figura 18: Componentes seleccionados en la zona de diseño.

En la **Table input** se configura la conexión a la base de datos origen. Luego se selecciona la tabla a la que se le va a realizar el proceso ETL, para este ejemplo la tabla seleccionada es la tabla Autor. Como la tabla Autor se relaciona con otras tablas y es necesario escoger otros campos se utiliza la cláusula inner join para realizar la unión con otras tablas.

Después se elige el componente **Selecciona/renombra valores**, donde se definen los campos de entrada de la tabla escogida y los campos de salida, siempre teniendo en cuenta que el nombre de los atributos sea igual a como se encuentra en sus respectivas tablas.

Luego el componente **Insertar/actualizar** es utilizado para insertar o actualizar los datos seleccionados, se configura la conexión a la base de datos destino. Se escoge la tabla que guardará los datos, pero antes se compara con su id en esta base de datos para comprobar que no se inserte la misma información.

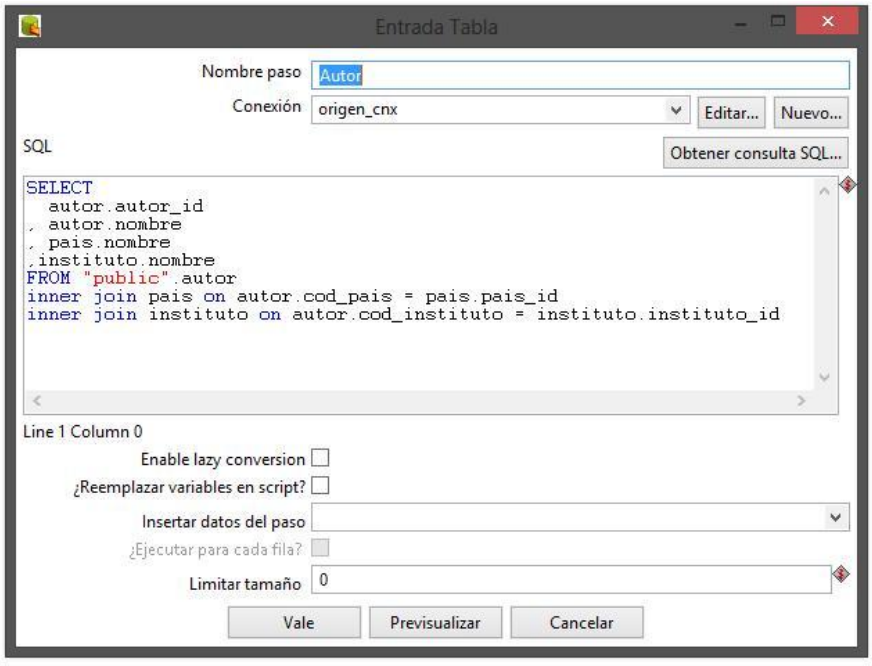


Figura 19: Configuración de la Table input.

- **Implementación de la ETL para la dimensión Tiempo:**

Para la implementación de la ETL Tiempo la cual es la encargada de cargar la fecha de la fuente de datos, y representa el mes y año en que fue publicada una publicación. Se coloca en la zona de diseño los siguientes componentes: Table input, Valor Java Script Modificado, Selecciona/Renombra valores e Insertar/actualizar.

Figura 20: Componentes seleccionados en la zona de diseño.

El componente **Valor Java Script Modificado** se utiliza para obtener el trimestre. Se realiza la implementación Java Script para determinar el trimestre, el cual puede ser de 4 tipos:

- Trimestre = 1 si el mes es Enero, Febrero o Marzo.
- Trimestre = 2 si el mes es Abril, Mayo o Junio.
- Trimestre = 3 si el mes es Julio, Agosto o Septiembre.
- Trimestre = 4 si el mes es Octubre, Noviembre o Diciembre.

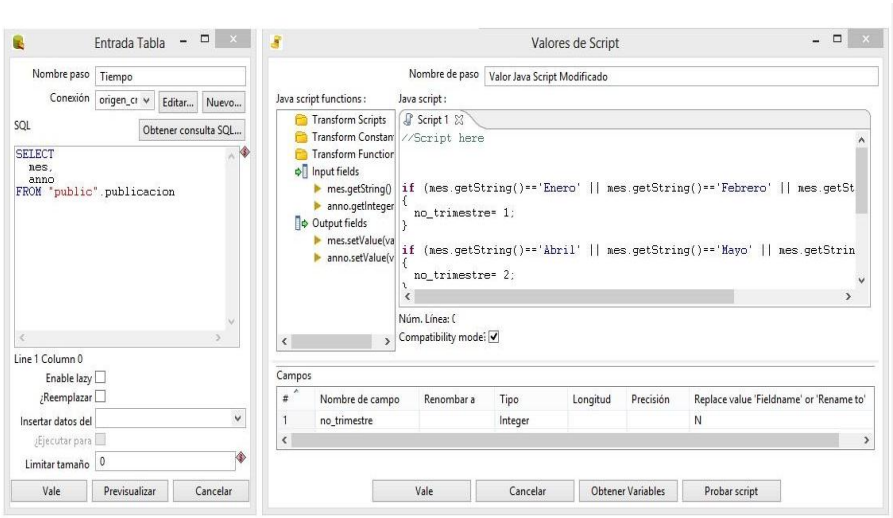


Figura 21: Configuración de la Table input y Valor Java Script Modificado.

- **Implementación de la ETL para la Tabla Fact_Publicacion:**

La implementación de la ETL para la Tabla Fact_Publicacion se seleccionó un conjunto de componentes, para proceder al llenado de la misma. Los componentes utilizados son: Table input, Búsqueda, Valor Java Script Modificado, Filtrar filas, Selecciona/Renombra valores, Ordenar filas, Agrupar por e Insertar/actualizar.

El componente **Búsqueda** es utilizado para chequear en cada tabla si el identificador (id) que se va a insertar se encuentra en la tabla, de tal manera que si el identificador no se encuentra en la tabla se agrega y si está no se permite agregarlo.

El componente **Filtrar filas** es utilizado para chequear que ninguno de los campos que provienen de la búsqueda sean nulos.

En el componente **Ordenar filas** se definen los identificadores que van a entrar a la Tabla Fact_Publicacion.

En el componente **Agrupar por** se define la medida cant_publicacion, la cual va ser determinada contando el identificador (id) de las publicaciones.

Figura 22: Componentes seleccionados en la zona de diseño.

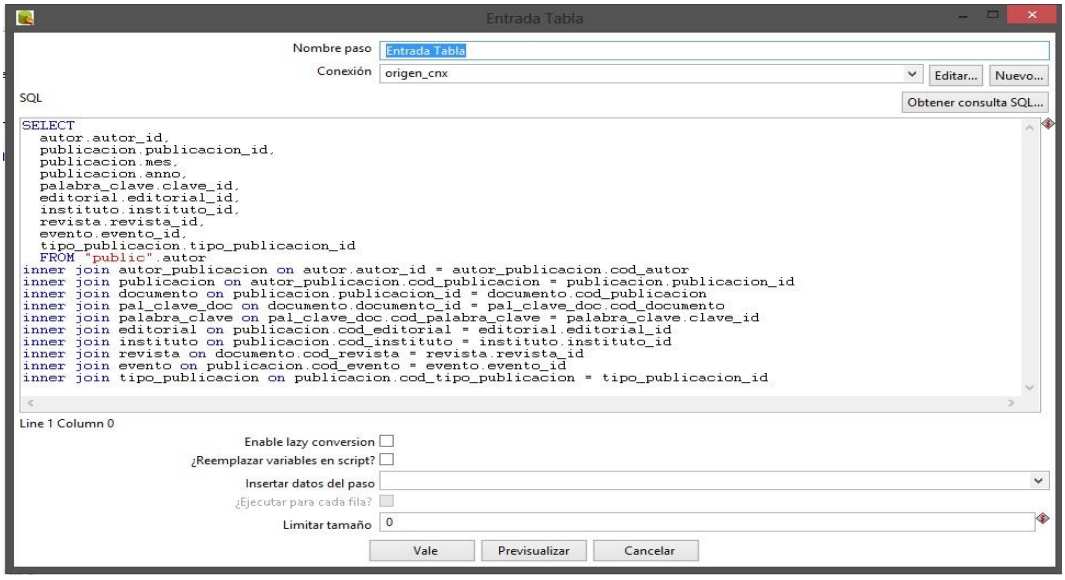


Figura 23: Configuración de la Table input.

Diseño del Job:

Una vez terminada la implementación de todas las ETL es necesario utilizar un Job (trabajo). Este es implementado y utilizado con el objetivo de que ejecute todas las ETL al ser iniciada esta funcionalidad.

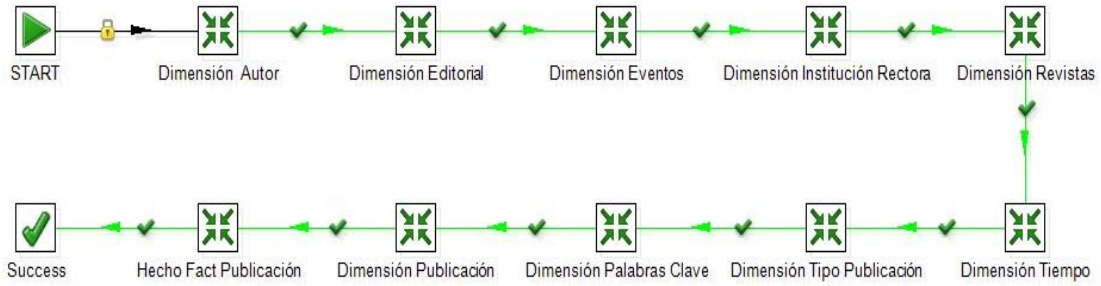


Figura 24: Diseño del Job.

3.2.2 Implementación del Cubo OLAP

Primeramente se crea el esquema (Schema) y el cubo. Luego se crea una tabla (Table) y se nombra como fpublicacion. Seguidamente se adicionan las dimensiones y se le establecen las jerarquías (Hierarchy). Al ser adicionadas todas las dimensiones y establecer su jerarquía se añade el indicador dando clic derecho encima del cubo y seleccionando la opción adicionar mesura (Add Measure), la cual tiene por nombre cant_public y se realiza su cálculo.

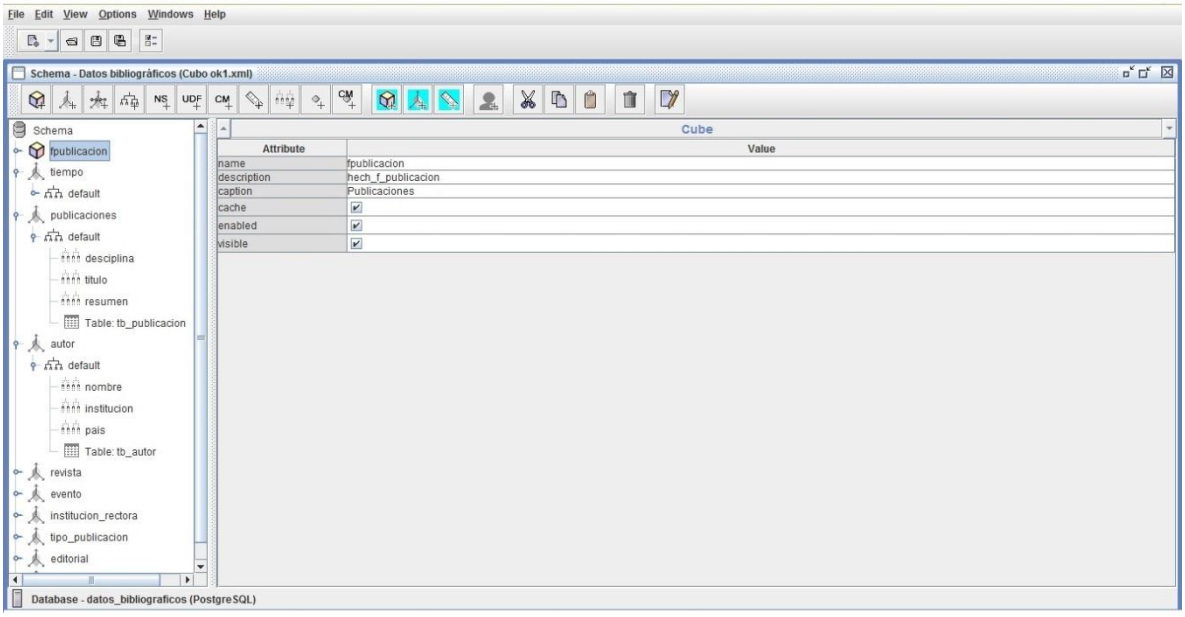


Figura 25: Implementación del cubo.

3.3 Especificación y desarrollo de las aplicaciones de la Inteligencia de Negocio

La aplicación de la Inteligencia de negocio es un elemento fundamental en la explotación del mercado de datos. Son utilizadas para consultar, analizar y presentar información dirigida a apoyar las necesidades del negocio. Proporcionan al usuario el acceso rápido de la información almacenada en el mercado de datos. Permiten además generar escenarios, pronósticos y reportes que apoyen a la toma de decisiones. Estas aplicaciones se desglosan en las herramientas de construcción de reportes y tablero de mando.

3.4 Herramienta de administración del mercado de datos

La consola de administración de Pentaho proporciona al usuario una ubicación central desde la que se pueden administrar las implementaciones de Pentaho. Simplifica muchas tareas administrativas comunes, como la gestión de usuarios y roles, trabajos de planificación y gestión de servicios. Además permite automatizar algunas de las tareas que antes se realizaban manualmente. (37)

Para la administración del mercado de datos se utilizó la herramienta Administration Console, en la cual se crearon dos usuarios administradores los cuales tienen todos los

permisos para trabajar con el mercado de datos. Además se crearon roles de usuarios los cuales solo podrán realizar tareas de visualización de los reportes predefinidos y otras tareas predefinidas.

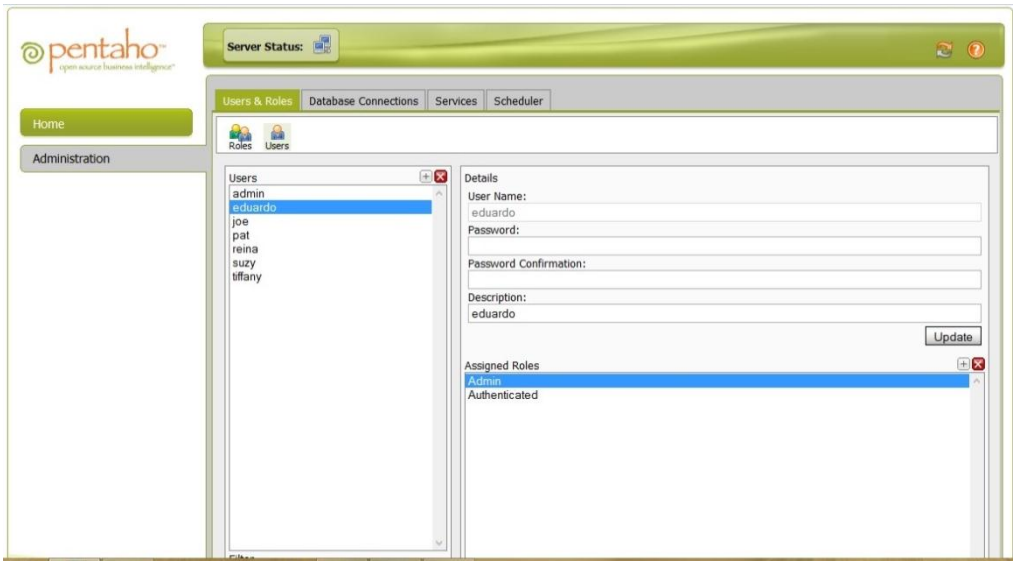


Figura 26: Herramienta de administración.

3.5 Creación del repositorio de metadatos

Luego de haber terminado de cargar y publicar el cubo del mercado de datos, se pasa a configurar el repositorio de metadatos, con el objetivo de brindar al usuario la posibilidad de interactuar con los datos almacenados. Para su desarrollado se utilizó la herramienta Pentaho Metadata-edition.

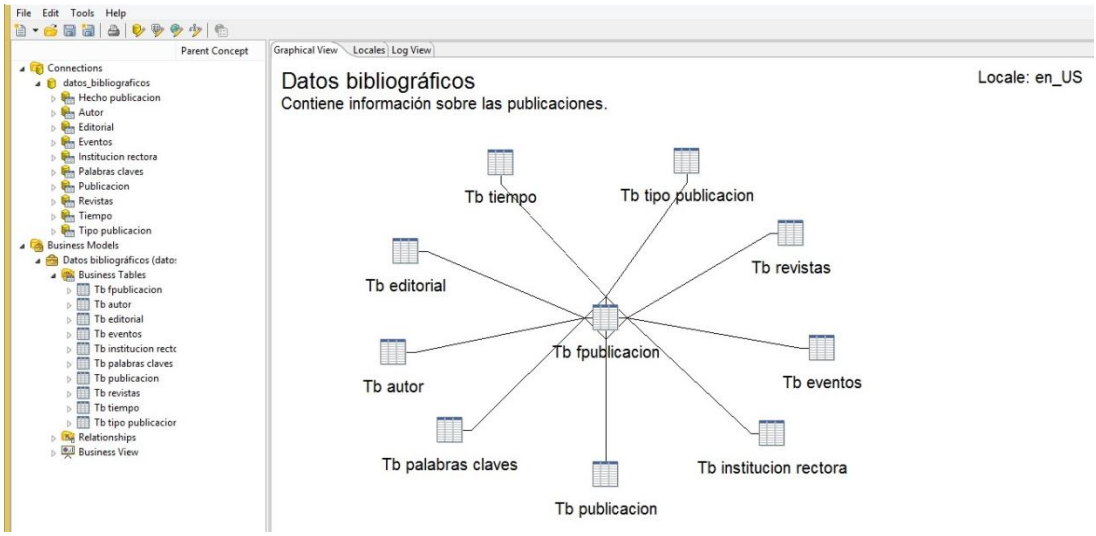


Figura 27: Reporte de metadatos.

A continución se muestra un ejemplo de una consulta Ad-hoc realizada para determinar la productividad de los autores por instituciones.

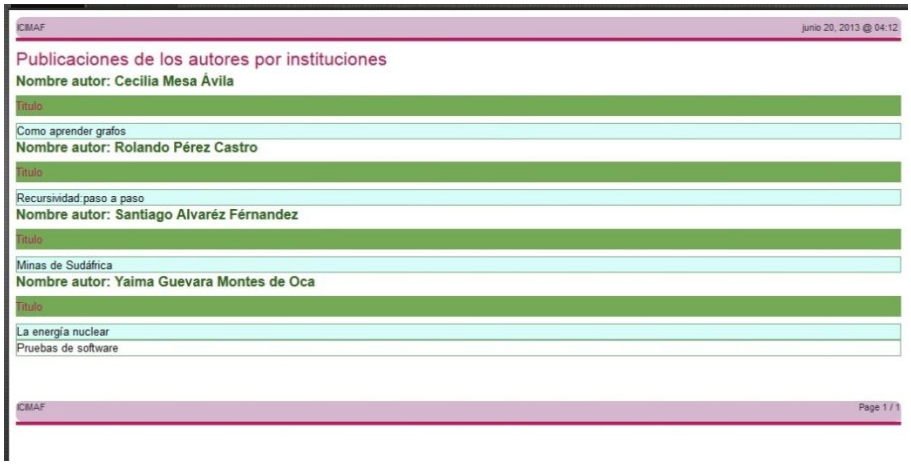


Figura 28: Ejemplo de consulta Ad-hoc.

3.5.1 Herramienta de construcción de reportes

Para realizar los reportes se emplea la herramienta Pentaho Report Designer. Se realiza un reporte para cada requisito informacional definido. A continuación se muestran algunos de los reportes diseñados.

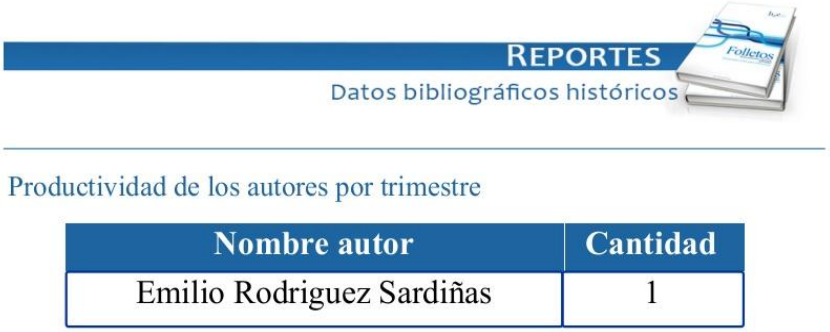


Figura 29: Reporte de la productividad de los autores por trimestre.

Productividad del país seleccionado.

Institución	Autores	Publicaciones
Eliseo Reyes	Yaima Guevara Montes de Oca	1
Instituto de Medicina ..	Cecilia Mesa Ávila	1
Eliseo Reyes	Emilio Rodriguez Sardiñas	1

Figura 30: Reporte de la productividad por países.

Productividad por instituciones editoras

Autores	Cantidad
Rolando Pérez Castro	1
Yaima Guevara Montes de Oca	1
Emilio Rodriguez Sardiñas	1

Figura 31: Reportes de la productividad por instituciones rectoras.

Tendencias de publicación

Disciplina	Cantidad
Informática	2

Figura 32: Reporte de las tendencias de publicaciones.

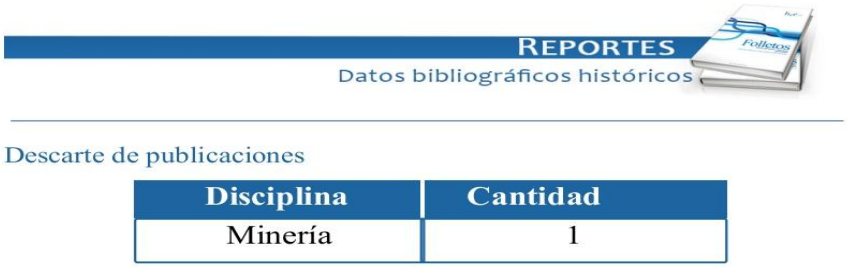


Figura 33: Reporte de descarte de publicaciones.

3.6 Tablero de Mando

El tablero de mando (*Dashboard*) les permite a los usuarios un acceso interactivo a la información y ofrece personalización dinámica, brindándoles la posibilidad a los usuarios de navegar por la aplicación, modificar los reportes e interactuar con los resultados.

La solución está integrada por los tableros de mando: Productividad y Predicción. El tablero de mando Productividad contribuye al análisis de toda la información recopilada de las publicaciones realizadas por los autores. El tablero de mando Predicción contribuye al análisis de la información recopilada para descartar o predecir temas de publicaciones.

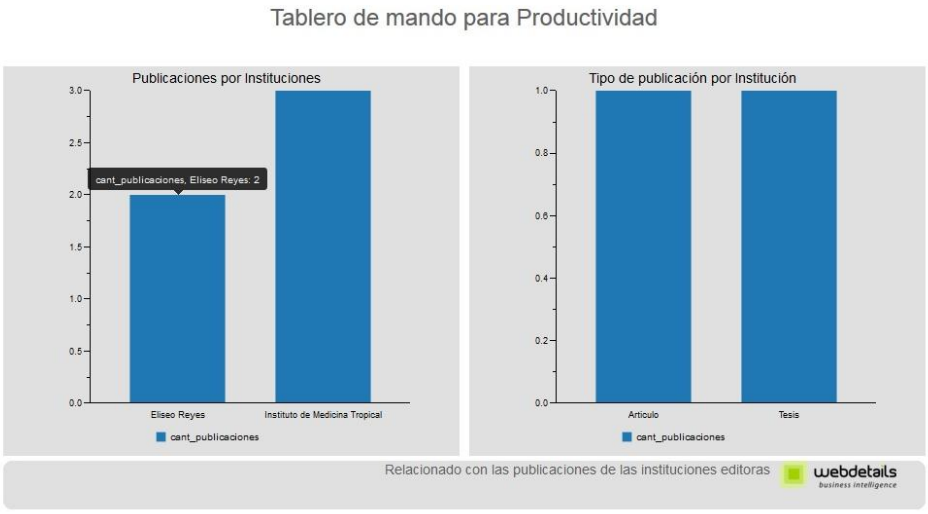


Figura 34: Tablero de mando para el proceso Productividad.

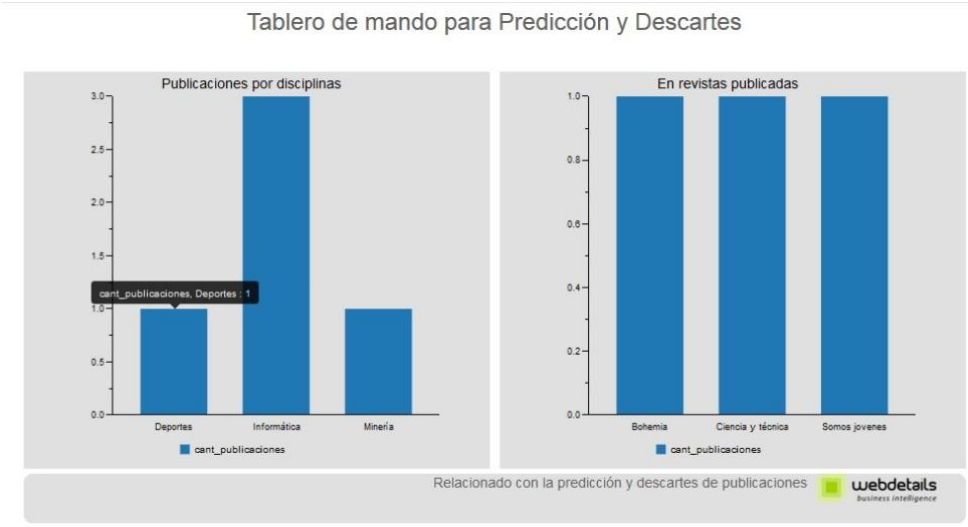


Figura 35: Tablero de mando para el proceso Predicción.

3.7 Pruebas

Para evaluar el funcionamiento del mercado de datos es necesario realizar pruebas en diferentes aspectos. Con el resultado obtenido de estas pruebas se pueden observar las fortalezas y debilidades que presenta el producto y de esta forma mejorar los aspectos débiles que sean identificados. Las pruebas tienen como objetivo verificar que el sistema cumpla con los requerimientos que se definieron y describieron para su desarrollo, además, ayudan a encontrar errores en la implementación, que no se hayan identificado en el proceso de desarrollo del producto.

El mayor problema que presenta la metodología Kimball es la fase de pruebas ya que solamente propone utilizar las listas de chequeo para evaluar la aplicación desarrollada.

Para establecer las pruebas se definieron dos iteraciones por parte del grupo de desarrollo, la primera para encontrar las no conformidades y una segunda para comprobar que no se introdujeron nuevos errores cuando se arreglaron las no conformidades detectadas.

3.7.1 Listas de Chequeo

¿Qué es una lista de chequeo?

La lista de chequeo es un listado de preguntas, en forma de cuestionario que se utiliza para verificar el grado de cumplimiento de determinadas reglas establecidas a priori, con un fin determinado. Describen de manera organizada criterios en relación al conocimiento

o los procedimientos de determinadas acciones. Aunque no exista un modelo genérico para la confección de una lista de chequeo; existen diversos formatos debido a que cada lista debe ser particular de cada negocio o proceso, a tono con las características y necesidades del mismo. Además se está al tanto del hecho que su estructura consta de indicadores con respuestas que califican la ocurrencia o no de un fenómeno relacionado con el tema a evaluar, así como la intensidad del mismo. (38)

Son herramientas importantes para concentrar gran cantidad de información y conocimiento, de manera concisa, evitando en su aplicación errores de omisión, creando con ello un mecanismo fiable y reproducible, mediante evaluaciones que permiten mejorar normas de calidad. También ayudan a los evaluadores a identificar posibles errores en procedimientos a evaluar y poseen cualidades de reproducibilidad y contextualización.

Elaboración de la lista de chequeo

Para elaborar la lista de chequeo, se tuvieron en cuenta los elementos de evaluación que no deben faltar una vez que se realice el proceso de diseño e implantación del mercado de datos. Contiene diferentes indicadores a evaluar los cuales se encuentran distribuidos en tres secciones fundamentales (38):

Estructura del documento: abarca todos los aspectos definidos por el expediente de proyecto o el formato establecido por el proyecto.

Indicadores definidos por la etapa: abarca todos los indicadores a evaluar durante la etapa de diseño e implantación del mercado de datos.

Semántica del documento: contempla todos los indicadores a evaluar respecto a la ortografía, redacción y demás.

Elementos que forman parte de la estructura de la lista de chequeo:

Peso: define si el indicador a evaluar es crítico o no.

Indicadores a evaluar: son los indicadores a evaluar en las secciones Estructura del documento, Semántica del documento e Indicadores definidos por la etapa; estos últimos dependen de los elementos de evaluación definidos para la etapa del proceso de diseño e implantación del mercado de datos. A un elemento de validación puede responder uno o varios indicadores.

Eval. (Evaluación): es la forma de evaluar el indicador en cuestión. El mismo se evalúa de 1 en caso de que exista alguna dificultad sobre el indicador y 0 en caso de que el indicador revisado no presente problemas.

NP (No Procede): se usa para especificar que el indicador no es necesario evaluarlo en ese caso.

Cantidad de elementos afectados: especifica la cantidad de errores encontrados sobre el mismo indicador.

Comentario: especifica los señalamientos o sugerencias que quiera incluir la persona que aplica la lista de chequeo. Pueden o no existir señalamientos o sugerencias.

Estructura del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿El documento a entregar contiene las secciones obligatorias de la plantilla estándar definida? (Portada, Resumen, Tabla de Contenidos, Introducción, Contenido, Conclusión, etc.)	0		0	
Indicadores definidos por la etapa					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
	1. ¿La metodología satisface las necesidades del problema?	0		0	
	2. ¿La granularidad definida respeta el nivel de detalle con el que desea almacenar la información sobre el negocio analizado?	0		0	
	3. ¿Las dimensiones están bien definidas?	0		0	
	4. ¿Se utilizó el menor número de transformaciones posibles al cargar los datos hacia el Staging area?	0		0	
crítico	5. ¿Se creó el Modelo Físico a partir del Modelo Lógico?	0		0	
crítico	6. ¿Cumple la implementación del proceso de ETL con la metodología definida?	0		0	

	7. ¿Se tuvo en cuenta los formatos fuentes y tipos de datos de las perspectivas de análisis?	0		0	
	8. ¿La extracción de los datos se realiza a partir de las fuentes de datos?	0		0	
	9. ¿Se realiza una limpieza de los datos antes de realizar la carga de los mismos?	0		0	
crítico	10. ¿El sistema garantiza la confidencialidad y seguridad de acceso a los datos por el rol de los usuarios?	0		0	
crítico	11. ¿Se cargan primero los datos de las tablas de dimensiones y luego los de las tablas de hechos?	0		0	
crítico	12. ¿La tipología de esquema seleccionada es la que mejor se adapta a los requerimientos del negocio?	0		0	
	13. ¿Los cubos OLAP se cargan con rapidez y con los datos establecidos?	0		0	
	14. ¿El sistema responde de forma rápida y veraz a la información que le sea solicitada por parte del usuario?	0		0	
Semántica del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿Se han identificado errores ortográficos en el documento a entregar o en los modelos diseñados?	0		0	
crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?	0		0	
	3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?	1		4	

Tabla 3: Lista de Chequeo.

Luego de realizar la prueba basada en lista de chequeos, se continúa verificando el funcionamiento del mercado de datos al que se le realizaron otras pruebas, tales como:

Prueba de Integración: esta prueba consistió en integrar el cubo realizado con el servidor de Inteligencia de Negocio (bi-server) mediante su publicación. Además se integraron los reportes creados y metadatos, para luego ser integrados con el tablero de mando permitiendo una visualización de la información almacenada en el mercado de datos.

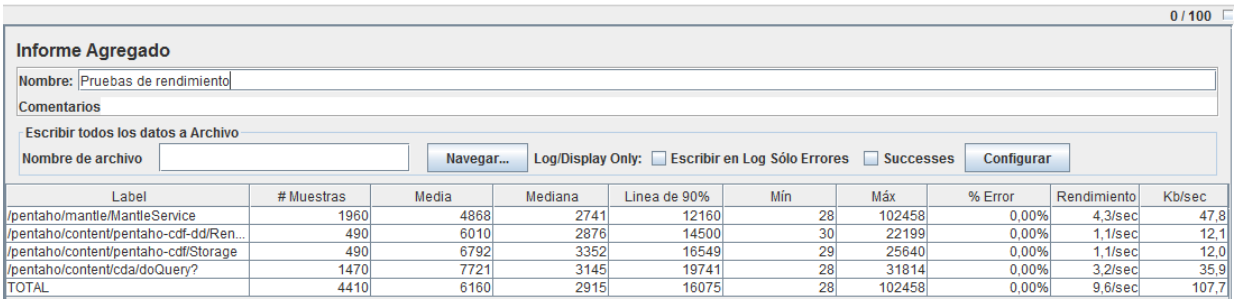
Prueba de Seguridad: esta prueba consiste en chequear que cada usuario creado solo pueda acceder a los permisos que le fueron asignados; verificando que ningún usuario pueda realizar tareas concedidas a los administradores. Los usuarios y contraseñas fueron definidos en la herramienta *Pentaho Administration Console*.

Prueba de Carga y Stress

Con el objetivo de garantizar la fiabilidad y rendimiento del mercado de datos se realizaron las pruebas de carga y stress a partir de la utilización de la herramienta *JMeter*, la cual ofrece una serie de facilidades para realizar este tipo de pruebas y está validada y probada su capacidad a nivel internacional. Los resultados arrojados luego de realizadas las pruebas mostraron que el rendimiento se mantuvo estable sin existir una variación exponencial en el tiempo de respuesta en las acciones, ya sea, para un número reducido o considerable de usuarios que manipulen simultáneamente el mercado de datos.

	Cantidad de usuarios	
Rendimiento	**50**	**100**
Tablero de mando	6,5 seg	9,6 seg
Consultas Ad Hoc	1,5 seg	1,6 seg

Tabla 4: Resultado de pruebas de rendimiento.



Label	# Muestras	Media	Mediana	Linea de 90%	Mín	Máx	% Error	Rendimiento	Kb/sec
/pentaho/mantle/MantleService	1960	4868	2741	12160	28	102458	0,00%	4,3/sec	47,8
/pentaho/content/pentaho-cdf-dd/Ren...	490	6010	2876	14500	30	22199	0,00%	1,1/sec	12,1
/pentaho/content/pentaho-cdf/Storage	490	6792	3352	16549	29	25640	0,00%	1,1/sec	12,0
/pentaho/content/cda/doQuery?	1470	7721	3145	19741	28	31814	0,00%	3,2/sec	35,9
TOTAL	4410	6160	2915	16075	28	102458	0,00%	9,6/sec	107,7

Figura 36: Informe agregado de pruebas con JMeter.

3.8 Resultados de las pruebas

Una vez aplicada la lista de chequeo se detectan los indicadores evaluados de mal y con el objetivo de darles solución se especifican en una tabla de no conformidades, la cual presenta la siguiente estructura (38):

No.: es un número consecutivo e indica la cantidad de no conformidades identificadas.

Elemento de evaluación: se refiere a un número que identifica al elemento de evaluación para el cual se corresponden los indicadores identificados. La evaluación estará definida en un intervalo de 1 a 10; donde de 1 a 5 representa mala calidad, de 6 a 7 baja calidad, de 8 a 9 aceptable y 10 calidad superior.

NC (No Conformidad): especifica la NC a la que se refiere.

Fase correspondiente: especifica la fase del procedimiento a la que corresponde la NC encontrada.

Significación: especifica si la NC es o no significativa, dependiendo si el indicador es o no crítico.

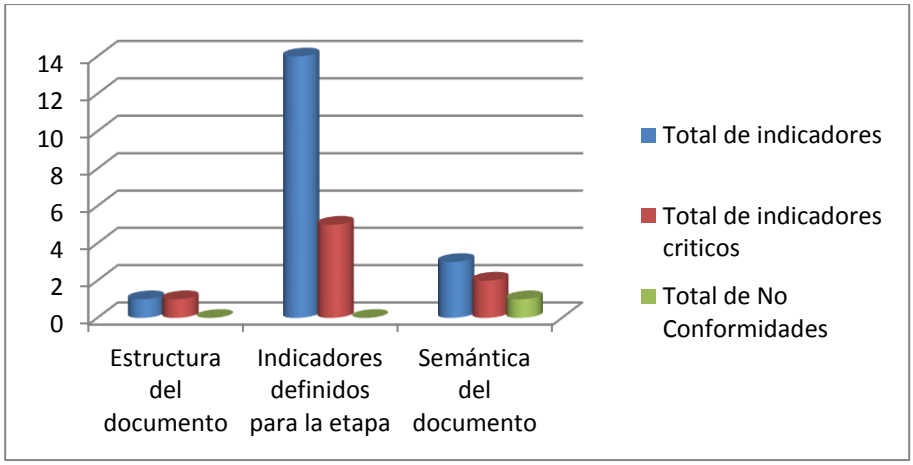
Recomendación: especifica si la NC es una recomendación, es decir que no es de obligatorio cumplimiento que se solucione por parte de los especialistas técnicos.

Estado NC: especifica el estado de solución en que se encuentra la NC, puede ser Pendiente o Solucionada.

Respuesta del equipo de desarrollo: si es necesario se especifica la respuesta que le da el equipo de desarrollo a la NC.

No	Elemento de Evaluación	NC	Fase correspondiente	Significación	Recomendación	Estado NC	Respuesta del equipo de desarrollo
1	7	Se encontraron algunas páginas en las cuales el número de página con que aparece en el índice no coincide con el contenido que se refleja realmente en dicha página.	Confección del documento.	No significativa.		Resuelta.	Este error fue provocado por no realizar la actualización de la tabla de contenido luego de a ver agregados nuevos subtítulos.

Tabla 5: Lista de No Conformidades.



Gráfica 1: Comportamiento de los indicadores por secciones.

3.9 Conclusiones

Con la implementación del proceso ETL y del cubo OLAP se realizó el mercado de datos desarrollado para el análisis de datos bibliográficos. Se confeccionaron los reportes predefinidos dados los requisitos informacionales. Se definieron los indicadores para la lista de chequeo, se realizaron pruebas de integración, fiabilidad, rendimiento y seguridad. Luego de realizar las pruebas definidas fueron corregidos los errores identificados en las mismas.

CONCLUSIONES

Con el término de la investigación se logró una herramienta que facilita el trabajo de análisis de datos bibliográficos, mediante la obtención de varios indicadores, para apoyar el proceso de toma de decisiones por parte de los especialistas del ICIMAF. Se obtuvo una solución realizada con la suite Pentaho, que incluye una serie de reportes y tableros de mando con gráficos interactivos, que facilitan el análisis y ayudan en la comprensión de los resultados de los diferentes indicadores definidos. Durante el desarrollo de la investigación se determinaron los requisitos informacionales que dan respuesta a las capacidades del mercado de datos para el análisis de datos bibliográficos. Se modeló el mercado de datos teniendo en cuenta los pasos que define la metodología de Kimball para la construcción de los mismos. Se implementó el proceso de extracción, transformación y carga de los datos con el objetivo de seleccionar específicamente la información relevante. Se diseñó e implementó el cubo OLAP, reportes, metadatos y tableros de mando. Se realizaron las pruebas para validar la calidad del mercado de datos a partir de las listas de chequeo. La solución propuesta es un aporte más al desarrollo y uso del software libre en nuestro país, así como un paso más a la independencia tecnológica.

RECOMENDACIONES

Para el presenta trabajo de diploma se recomienda:

- Trabajar más el tema del almacén de datos y mercado de datos en los proyectos productivos de nuestra facultad.
- Incorporar paulatinamente más información al mercado de datos.
- Agregar nuevos indicadores de información bibliográfica a la solución propuesta.
- Llevar el mercado de datos realizado hasta la fase de despliegue.
- Proponer la construcción de mercado de datos en otras áreas relacionadas con el análisis de datos bibliográficos.

GLOSARIO DE TÉRMINOS.

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

MDX: es el acrónimo de Multidimensional eXpression o expresiones multidimensionales, es un lenguaje de consulta para bases de datos multidimensionales sobre cubos OLAP, se utiliza en Business Intelligence para generar reportes para la toma de decisiones basados en datos históricos, con la posibilidad de cambiar la estructura o rotación del cubo.

Ad-Hoc: son consultas en bases de datos, el sistema le permite al usuario personalizar una consulta en tiempo real, en vez de estar atado a las consultas prediseñadas para informes.

Cubos: Representación o visualización de una tabla de hechos y sus dimensiones correspondientes.

ICIMAF: Instituto de Cibernética, Matemática y Física.

SGBD: Sistema Gestor de Base de Datos.

Kettle: Sistema de ETL de Pentaho Data Integration.

REFERENCIAS BIBLIOGRÁFICAS.

1. Real Academia de la Lengua Española. [En línea] http://lema.rae.es/drae/?val=inform%C3%A1tica.

2. **Spinak, Ernesto.** *Indicadores cienciométricos.* Sao Paulo : s.n., 1998.

3. **ICIMAF.** - *Instituto de Cibernética, Matemática y Física .* Habana - Cuba : http://www.icimaf.cu/., 2013.

4. **Pritchard, Alan.** *Statiscal bibliography or Bibliometrics.* 1969.

5. **Kimball, Ralph y Ross, Margy.** *The Data Warehouse Toolkit. .* 2003.

6. **Raquel Abella, Lucía Cóppola, Diego Olave.** *Sistema Datawarehousing. Carga y control de la calidad.* Uruguay : s.n., 1999.

7. **Fernández, Carlos.** Dataprix Knowledge Is The Goal. [En línea] 2009. http://www.dataprix.com/arquitectura-data-warehouse-areas-datos-nuestro-almacen-corporativo.

8. **Farias, Ing. Jonathan Merino.** [En línea] 2011. https://sites.google.com/site/jmerinocorporation/.

9. **Callejo, Miguel Ángel Manso.** Universidad Politécnica de Madrid. [En línea] 2010. http://www2.topografia.upm.es/pdi/m.manso/docencia/IDE_plan92_ITT/IDE-2010/Tema_5/Que%20son%20los%20metadatos.pdf.

10. **González Ferrer, Yailin y Carmenate Acevedo, Yudit.** *Diseño e Implementación de un mercado de datos para el área de Registro Legal.* Habana : s.n., 2012.

11. **Ponniah, Paulraj.** *Data Warehousing Fundamentals:A Comprehensive Guide for IT Professionals.* New York, Chichester, Weinheim, Brisbane, Singapore y Toronto : s.n., 2001. s.n..

12. **Federico Piedrabuena, Gustavo Vázquez.** *Relevamiento: Diseño Físico de Sistemas OLAP.* Montevideo : s.n., 2005.

13. **Palominos, Fredi.** Sistemas de Soporte a la toma de Decisiones. [En línea] 2005. http://palomo.usach.cl/bdnc/2005-02/Presentaciones/U3-1-OLAP.pdf.

14. **Ramón Cueto, Ariagna y Valido Pérez, Yenisel.** *Implementación de un Data Warehouse para el control del Recurso Humano de la Salud.* Habana - Cuba : s.n., 2009.

15. **Hudomalj, E. Vidmar G.** *OLAP and Bibliographic Databases, Scientometrics, Vol. 58.* 2003.

16. **Baid, A. y Balmin A. y Hwang H. y Nijkamp E. y Rao J. y Reinwald B. y Simitsis A. y Sismanis Y. y Ham F.** *DBPubs: Multidimensional Exploration of Database Publications. Proceedings of the 34th International Conference on Very Large Data Bases, Vol. 1.* 2008.

17. **Bernabeu, Ricardo Dario.** *DATA WAREHOUSING: Investigación y Sistematización de Conceptos - HEFESTO Metodología propia para la Construcción de un Data Warehouse. V2.0.* 2010.

18. **Tufiño López, Jorge Luis.** Desarrollo e Implantación del Datamart para el Sistema Nacional de Vigilancia Tecnológica de Software Libre. Tesis de Grado. [En línea] 2011. http://bibdigital.epn.edu.ec/handle/15000/4101.

19. **Inmon, W. H.** *Building the Data Warehouse, Fourth Edition. .* 2005.

20. **Rivadera, Gustavo.** *La Metodología de Kimball para el Diseño de almacenes de datos.* Salta : s.n., 2010.

21. **Alarcón, Vicenç Fernández.** *Desarrollo de sistemas de información. Una metodología basada en el modelado.* s.l. : Edicions UPC, 2006.

22. **Cruz-Chávez, Dr. Marco Antonio.** Universidad Autónoma del Estado de Morelos. [En línea] 2009. http://www.uaem.mx/posgrado/mcruz/cursos/miic/oracle.pdf .

23. **Daniel Fernández Arencibia, Yunieski Fábregas Santos.** *Administración, configuración y optimización de un Sistema de Bases de Datos Descentralizado en Oracle Database 10g.* 2007.

24. **Petković, Dušan.** *Microsoft SQL Server 2005: a beginner's guide.* s.l. : McGraw-Hill Professional, 2005.

25. **Cruz-Chávez, Dr. Marco Antonio.** Universidad Autónoma del Estado de Morelos. [En línea] 2012. http://www3.uaem.mx/posgrado/mcruz/cursos/miic/sql5.pdf.

26. —. Universidad Autónoma del Estado de Morelos. [En línea] 2009. http://www.uaem.mx/posgrado/mcruz/cursos/miic/MySQL.pdf.

27. **D.González, Carlos.** Curso de base de datos PostgreSQL, SQL avanzado y PHP. [En línea] 2010. http://www.usabilidadweb.com.ar/postgre.php.

28. **A Quiñones, Ernesto.** Introducción a PostgreSQL. [En línea] http://www.postgresql.org.pe/articles/introduccion_a_postgresql.pdf.

29. **Bouman, Roland y Van Dongen, Jos.** *Pentaho Solutions. Business Intelligence and DataWarehousing with Pentaho and MySQL.* 2010.

30. **Díaz, Josep Curto.** *Introducción al Business Intelligence.* s.l. : Editorial UOC, 2010.

31. **Hidalgo López, Leydis y Caballero Cartaya, Alier .** *Implementación de un DataMart para la Unidad.* Habana : UCI, 2010.

32. **Designer., Reporting en Pentaho con Pentaho Report.** El Rincon del BI. [En línea] 2011. http://churriwifi.wordpress.com/2010/07/15/17-4-reporting-en-pentaho-con-pentaho-re.

33. Portada sobre la plataforma Pentaho Open Source Business Intelligence. [En línea] http://pentaho.almacen-datos.com/mondrian.html.

34. Microsoft. [En línea] http://office.microsoft.com/ca-es/excel-help/ejecutar-una-accion-de-servidor-olap-en-un-informe-de-tabla-dinamica-HA010177379.aspx.

35. SQL Power Group. [En línea] 2009. http://power-architect.googlecode.com/files/PowerArchitectUserGuide-0.9.13.pdf.

36. **Ralph Kimball, Margy Ross, Warren Thornthwaite, Joy Mundy, Bob Becker.** *The Data Warehouse Lifecycle Toolkit.* 2002.

37. **Doug, Moran.** Pentaho . *Pentaho .* [En línea] 17 de Febrero de 2010. http://wiki.pentaho.com/display/ServerDoc2x/.01+Introduction.

38. **Rodríguez, Javier Sotolongo y Peralta, Yohan Orlando Góngora.** *Implementación del proceso de extracción, transformación y carga de un Datawarehouse para los Ensayos Clínicos del Centro de Inmunología Molecular.* Habana - Cuba : s.n., 2010.

39. **Imhoff Claudia, Galemmo Nicholas y Geiger Jonathan G.** *Mastering Data Warehouse Desing, Relational and Dimentional Techniques.* EUA : Wiley Publishing Inc, 2003.

40. **Hobbs, Lilian.** *Oracle Database 10g Data Warehousing.* EUA : ELSEVIER Digital Press, 2005.

41. **Rouse, Margaret.** SearchSqlserver. [En línea] 2005. http://searchsqlserver.techtarget.com/definition/hybrid-online-analytical-processing.

42. **Lee Gilbert, Jan Hewitt, Margaret Rouse.** SearchDataCenter. [En línea] 2005. http://searchdatacenter.techtarget.com/definition/OLTP.

43. **Esteban Zimányi, Elzbieta Malinowski.** *Advanced Data Warehouse Design.* 2008.

44. **Rivadera, Gustavo R.** Ucasal- Universidad Católica de Salta. Argentina. [En línea] 2010. http://www.ucasal.net/templates/unid-academicas/ingenieria/apps/5-p56-rivadera-formateado.pdf.

45. **Darmawikarta, Djoni.** *Dimensional Data Warehousing with MySQL.* s.l. : Pub. BrainySoftware. ISBN 0-9752128-2-6, 2007.

46. Software científico. [En línea] http://www.softwarecientifico.com/paginas/refviz.htm.

47. BibWord Microsoft Word Citation and Bibliography Styles . [En línea] 2009. http://bibword.codeplex.com.

48. **Miguel Antonio Alonso Escribano, Raúl Blanco Ramos, Manuel Óscar Sánchez Yagüe.** Universida Complutense de Madrid. [En línea] 2009. http://eprints.ucm.es/9532/1/Sistema_gestor_de_referencias_bibliogr%C3%A1ficas_y_elaborador_de_bibliograf%C3%ADas_para_MS_Word.pdf.

ANEXO

TEMA: RESPONSABILIDADES

1. Describe tu organización y sus relaciones internas con el resto de la compañía
2. ¿Cuáles son tus responsabilidades primarias?

TEMA: ASUNTOS Y OBJETIVOS DE NEGOCIO

1. ¿Cuáles son los objetivos de tu organización? ¿Qué usted está tratando de lograr?
2. ¿Cuáles son tus metas prioritarias sobresalientes de negocio?
3. ¿Cuál es tu métrica de éxito? ¿Cómo sabe usted que está bien?
4. ¿Cuál es el impacto en la organización?
5. ¿Cómo identifica usted problemas /excepciones o sabe que va encaminado al problema?

TEMA: ANALISIS DE LOS REQUERIMIENTOS

1. ¿Qué tipo de análisis de rutina usted hace actualmente?
2. ¿Cómo obtiene usted actualmente los datos?
3. ¿Qué datos son usados?
4. ¿Qué hace con la información una vez que la obtiene?
5. ¿Qué análisis le gustaría realizar?
6. ¿Hay mejoras potenciales para su método /proceso actual?
7. ¿Quién pide análisis ad hoc?
8. ¿Qué hacen con el análisis?
9. ¿Cuáles reportes usa usted actualmente?
10. ¿Qué dato en el reporte es importante?
11. ¿Cómo utiliza usted la información?
12. ¿Qué capacidades analíticas le gustaría a usted tener?
13. ¿Cuánta información histórica es requerida?
14. ¿Cuál es el impacto financiero?

Figura 1.1: Ejemplo de preguntas de la entrevista para los especialistas del centro ICIMAF.