Universidad de Cuenca
Facultad de Ingeniería
Escuela de Informática

Data Warehouse para el Centro de Documentación Regional "Juan Bautista Vázquez"

Autores:
Valeria Alexandra Haro Valle
Wilson Rodrigo Pérez Rocano

Director:
Ing. Víctor Hugo Saquicela Galarza

Tesis de grado previa a la obtención del Título de
Ingeniero de Sistemas

Cuenca, Junio de 2014

Resumen

El proceso de toma de decisiones en las bibliotecas universitarias es de suma importancia, sin embargo, se encuentra complicaciones como la gran cantidad de fuentes de datos y los grandes volúmenes de datos a analizar. Las bibliotecas universitarias están acostumbrados a producir y recopilar una gran cantidad de información sobre sus datos y servicios. Las fuentes de datos comunes son el resultado de sistemas internos, portales y catálogos en línea, evaluaciones de calidad y encuestas. Desafortunadamente estas fuentes de datos sólo se utilizan parcialmente para la toma de decisiones debido a la amplia variedad de formatos y estándares, así como la falta de métodos eficientes y herramientas de integración. Este proyecto de tesis presenta el análisis, diseño e implementación del Data Warehouse, que es un sistema integrado de toma de decisiones para el "Centro de Documentación Juan Bautista Vázquez". En primer lugar se presenta los requerimientos y el análisis de los datos en base a una metodología, esta metodología incorpora elementos claves incluyendo el análisis de procesos, la calidad estimada, la información relevante y la interacción con el usuario que influyen en una decisión bibliotecaria. A continuación, se propone la arquitectura y el diseño del Data Warehouse y su respectiva implementación la misma que soporta la integración, procesamiento y el almacenamiento de datos. Finalmente los datos almacenados se analizan a través de herramientas de procesamiento analítico y la aplicación de técnicas de Bibliomining ayudando a los administradores del centro de documentación a tomar decisiones óptimas sobre sus recursos y servicios.

Palabras Claves:
Sistema de soporte a la toma de decisión, centro de documentación, biblioteca, Data Warehouse, Bibliomining, Pentaho, Weka.

Abstract

The decision-making process in academic libraries is paramount; however highly complicated due to the large number of data sources, processes and high volumes of data to be analyzed. Academic libraries are accustomed to producing and gathering a vast amount of statistics about their collection and services. Typical data sources include integrated library systems, library portals and online catalogues, and systems of consortiums and quality surveys. Unfortunately, these heterogeneous data sources are only partially used for decision-making processes due to the wide variety of formats, standards and technologies, as well as the lack of efficient methods of integration. This thesis presents the analysis, design and implement of Data Warehouse for an academic library "Juan Bautista Vázquez". Firstly, an appropriate methodology documented in a previous study is used for data collection. This methodology incorporates key elements including process analysis, quality estimation, information relevance and user interaction that may influence a library decision. Based on the above mentioned approach, this study defines a set of queries of interest to be issued against the integrated system proposed. Then, relevant data sources, formats and connectivity requirements for a particular example are identified. Next, Data Warehouse architecture is proposed to integrate, process, and store the collected data transparently. Eventually, the stored data are analyzed through reporting techniques of analytical processing and prototype of Bibliomining. By doing so, the thesis provides the design of an integrated solution that assists library managers to make tactical decisions about the optimal use and leverage of their resources and of services.

Índice general

Índice de figuras

Índice de tablas

Universidad de Cuenca
Clausula de derechos de autor

Yo, Valeria Alexandra Haro Valle, autora de la tesis "Data Warehouse para el Centro de Documentación Regional Juan Bautista Vázquez", reconozco y acepto el derecho de la Universidad de Cuenca, en base al Art. 5 literal c) de su Reglamento de Propiedad Intelectual, de publicar este trabajo por cualquier medio conocido o por conocer, al ser este requisito para la obtención de mi título de Ingeniero en Sistemas. El uso que la Universidad de Cuenca hiciere de este trabajo, no implicará afección alguna de mis derechos morales o patrimoniales como autora.

Cuenca, 20 de junio de 2014

Valeria Alexandra Haro Valle

C.I: 0604043679

Universidad de Cuenca
Clausula de derechos de autor

Yo, Wilson Rodrigo Pérez Rocano, autor de la tesis "Data Warehouse para el Centro de Documentación Regional Juan Bautista Vázquez", reconozco y acepto el derecho de la Universidad de Cuenca, en base al Art. 5 literal c) de su Reglamento de Propiedad Intelectual, de publicar este trabajo por cualquier medio conocido o por conocer, al ser este requisito para la obtención de mi título de Ingeniero en Sistemas. El uso que la Universidad de Cuenca hiciere de este trabajo, no implicará afección alguna de mis derechos morales o patrimoniales como autor.

Cuenca, 20 de junio de 2014

Wilson Rodrigo Pérez Rocano

C.I: 0104784491

Universidad de Cuenca
Clausula de derechos de autor

Yo, Valeria Alexandra Haro Valle, autora de la tesis "Data Warehouse para el Centro de Documentación Regional Juan Bautista Vázquez", certifico que todas las ideas, opiniones y contenidos expuestos en la presente investigación son de exclusiva responsabilidad de su autora.

Cuenca, 20 de junio de 2014

Valeria Alexandra Haro Valle

C.I: 0604043679

Universidad de Cuenca
Clausula de derechos de autor

Yo, Wilson Rodrigo Pérez Rocano, autor de la tesis "Data Warehouse para el Centro de Documentación Regional Juan Bautista Vázquez", certifico que todas las ideas, opiniones y contenidos expuestos en la presente investigación son de exclusiva responsabilidad de su autor.

Cuenca, 20 de junio de 2014

Wilson Rodrigo Pérez Rocano

C.I: 0104784491

Agradecimientos

La presente Tesis de grado es un logro obtenido por el esfuerzo y la constancia, en la que participaron varias personas las cuales nos brindaron su ayuda, nos asesoraron, nos dieron ánimos y nos guiaron para una exitosa culminación del mismo.

Agradecemos de manera especial al Ing. Víctor Saquicela por el apoyo desinteresado e incondicional, por haber confiado en nosotros y guiarnos durante el transcurso del proyecto como director de tesis, a la Ing. Lorena Sigüenza por su asesoramiento y comentarios desde el inicio hasta la culminación del mismo. Así como también, le agradecemos a la Lic. Rocío Campoverde, al Ing. Mauricio Brito, al Ing. Andrés de los Reyes, a la Lic. Margarita Gutierrez y a todos los bibliotecarios del Centro de Documentación "Juan Bautista Vázquez" que siempre estuvieron prestos en apoyarnos y asesorarnos en las dudas que se nos presentó.

Gracias también a nuestros queridos compañeros y amigos de clase que nos supieron brindar su amistad incondicional, una meta más cumplida pero siempre quedan los buenos recuerdos de los momentos que supimos aprovechar y disfrutar en la universidad. Gracias en general a todas las personas que ayudaron directa e indirectamente en la realización de este proyecto.

Valeria, Wilson

Dedicatoria

Este trabajo que es la cumbre de la lucha constante que significó este sueño profesional lo dedico a mi madre por ser el pilar más importante en mi vida, quien ha sabido formarme con buenos sentimientos, hábitos y valores. A mis hermanos por ser ese apoyo incondicional, por su cariño, comprensión y por toda la felicidad que me han brindado en esta vida. Finalmente, este trabajo va dedicado a ese ángel que es mi motivación de ser mejor cada día, quien me enseño el significado del compromiso y la responsabilidad, a no desfallecer ni rendirme ante nada a través de sus sabios consejos.

Valeria

Dedicatoria

Con todo cariño y amor a mis padres, por todo el apoyo brindado y esos ánimos que siempre sabían mantenerlos pese a los problemas que se presentaban. A mis hermanos y cuñados que los quiero mucho gracias por toda esa paciencia que han sabido tener conmigo, gracias por los consejos y la confianza depositada en mí. A mis sobrinos, a ustedes que siempre los llevo en mi mente que son lo más hermoso que Dios me ha regalado ya que son la alegría y la vida misma en la familia. A toda mi familia por sabernos comprender, reconociendo que lo más importante es vivir en comunión, compartiendo los mejores momentos que una persona puede experimentar. A mis grandes amigos que siempre están presentes en todo momento, por tantas desveladas que sufrimos y al final sirvieron de mucho, por esos momentos agradables y tristes que compartimos permitiendo valorar a las personas que nos rodean. Y a Dios por seguirme dando la oportunidad de vivir y regalarme una familia maravillosa.

"Un paso más en mi vida, un nuevo reto por vivirlo."

Wilson

Capítulo 1

Introducción

En este capítulo se aborda la descripción del problema que se pretende resolver al implementar la presente tesis con el título denominado **"Data Warehouse para el Centro de Documentación Juan Bautista Vázquez"** además, se incluye el alcance del mismo y se definen los objetivos a cumplir incluyendo la justificación respectiva.

1.1. Identificación del problema

Uno de los problemas principales hoy en día en toda empresa con o sin fines de lucro es la gestión de grandes volúmenes de información y la forma de explotar dicha información para lograr soporte de las decisiones financieras, administrativas y económicas. Toda empresa actualmente debe mantener un control de la información generada día a día para poder tomar decisiones de una forma óptima.

Al estar inmersos en la era del conocimiento, cuyo valor primordial es la información, la cual tiene un peso inmenso sobre cualquier otro recurso de una empresa; si un gerente está bien informado de lo que sucede dentro de la empresa con datos concretos y reales puede tomar decisiones mucho más acertadas para el bien de la empresa, permitiendo a los administradores manejar sus recursos de manera eficiente.

En las bibliotecas que son centros de información surge la necesidad de tener un control sobre los datos generados años atrás y de integrarlos para tener una visión general de su situación y así poder tomar decisiones sobre el manejo de recursos,

para beneficiar a los profesionales, estudiantes y profesores que acuden diariamente en busca de conocimiento a estos centros de información.

El Centro de Documentación Regional "Juan Bautista Vázquez" de la Universidad de Cuenca está conformado por las Bibliotecas de los Campus Central, El Paraíso (Áreas de la Salud) y Yanuncay (Áreas de Agropecuarias y Artes); donde, actualmente no existe una integración de la información que permita una adecuada toma de decisiones. Su conjunto documental está constituido por publicaciones convencionales en todas sus formas, así como por diversos soportes digitalizados, audiovisuales y bases de datos en línea. Razón por lo cual, se ve la necesidad de integrar los datos en un repositorio común para analizar la información y la toma de decisiones.

1.2. Justificación

Gracias al avance informático en la actualidad los procesos se han sistematizado, sean estos de empresas o de instituciones sin fines de lucro, como resultado de esta informatización se obtiene una gran cantidad de información que se almacenan en diferentes archivos y bases de datos.

Las aplicaciones de estos procesos que se manejan en la biblioteca han generado de una u otra manera la necesidad de contar con un sistema que ayude a una correcta toma de decisiones administrativas. Sin embargo, aunque estas aplicaciones han permitido la automatización de procesos también ha complicado la selección de información para la toma de decisiones, especialmente porque es necesario consultar uno por uno cada sistema.

Proveer servicios tanto a estudiantes como al personal docente en los préstamos de una variedad de libros, sean estos digitales o físicos, consultas de documentos, tesis o trabajos de investigación son procesos que se viven día a día en el Centro de Documentación Regional "Juan Bautista Vázquez" de la Universidad de Cuenca.

El poder saber qué temática de libros son más consultados, el intentar conocer a qué estudiantes de que facultad realizan préstamos o consultas y con qué frecuencia, el decidir en qué tipo de libros se debe actualizar o conseguir un número mayor de ejemplares, son inquietudes que se generan después de un proceso.

Por esta razón, la información obtenida diariamente por estos procesos sería de gran utilidad para la toma de decisiones institucionales, por lo cual, el siguiente paso a realizarse es analizar los datos que se tiene e interpretarlos. Por consiguiente, una herramienta que permita el análisis, el enlace de diferentes fuentes de datos, la interpretación y proporcionar resultados óptimos y eficientes, es un Data Warehouse, que junto con un proceso de minería de datos permita obtener resultados decisivos en base a las necesidades de la institución, ayudando así a una mejor administración.

A partir del análisis holístico realizado en la tesis doctoral en ejecución por la Ing. Lorena Sigüenza denominada **"OPTIMIZACIÓN DE PRESUPUESTOS Y RECURSOS EN BIBLIOTECAS UNIVERSITARIAS"**se pretende implementar esta herramienta de ayuda en la toma de decisiones que realice la integración de datos para proporcionar información consistente a los administradores.

1.3. Alcance

El proyecto pretende realizar el análisis, el diseño e implementación de un sistema de ayuda a la toma de decisiones según los requerimientos que plantean las personas involucradas tales como la directora de la biblioteca conjuntamente con su grupo de trabajo, y de los miembros del proyecto "**Cambio institucional para fortalecer la investigación y la educación**" del programa de cooperación institucional entre la Universidad de Cuenca y el Consejo de Universidades Flamencas (VLIR). Para ello se realizará un análisis de la mejor alternativa de software y demás herramientas necesarias.

El sistema que se va a implementar apoyará en la toma de decisiones para mejorar la administración de todos los recursos que posee el Centro de Documentación "Juan Bautista Vázquez" para lo cual, se realizará un análisis de las fuentes de información que pueden ser bases de datos relacionales y base de datos documentales tanto de la biblioteca como de los distintos sistemas de información que se manejan en la Universidad de Cuenca que se relacionan entre sí para satisfacer los requerimientos de la organización.

Para realizar la integración de éstas se tendrá que analizar el acceso a la información de los diferentes repositorios. En el caso de que no existiera dicha herramienta

para el acceso a alguna fuente de información, la presente tesis también contemplara el análisis y desarrollo de la herramienta de acceso a la información.

Además, se plantea realizar como parte de este proyecto un prototipo básico de minería de datos para el centro de documentación que permita extraer conocimiento del Data Warehouse para guiar la estrategia y planificación en el centro de documentación analizado.

1.4. Objetivo general

Analizar, diseñar e implementar un sistema de Data Warehouse de soporte de decisiones para el Centro de Documentación "Juan Bautista Vázquez" de la Universidad de Cuenca.

1.4.1. Objetivos Específicos

- Establecer un modelo lógico multidimensional que permita acceder de manera eficiente a la información.
- Analizar y comparar herramientas para la implementación de un Data Warehouse y seleccionar la más adecuada.
- Analizar las fuentes de información asegurando el acceso desde la herramienta tecnológica.
- Integrar las diferentes fuentes de información del centro de documentación y de la universidad en un repositorio común.
- Aplicar algoritmos de Data Mining para extraer conocimiento del Data Warehouse.

1.5. Métodos y Procedimientos

Se utilizará una combinación de métodos teóricos y empíricos para la construcción de un sistema de ayuda a la toma de decisiones en el Centro de Documentación

"Juan Bautista Vázquez". Se seguirá el método analítico para lograr la descomposición del problema en partes y el método sintético para lograr la integración de los componentes.

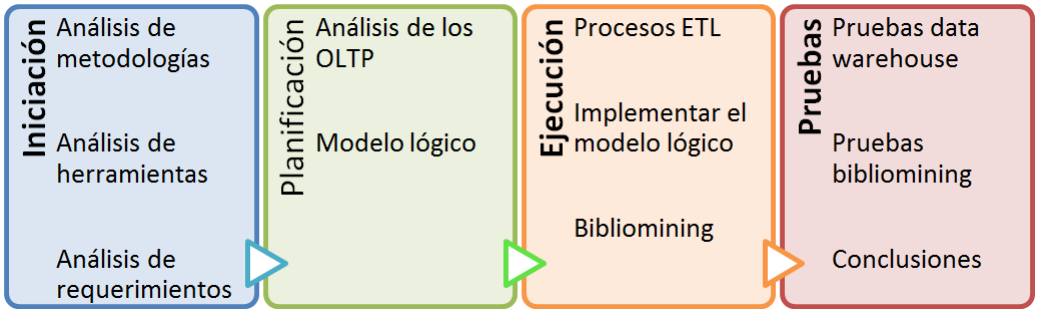


Figura 1.1: Fases de Implementación del proyecto

Como se observa en la figura 1.1 en la fase inicial se realizará una comparación de metodologías utilizadas para el desarrollo de un Data Warehouse, luego se seleccionará la más adecuada, así como también, se va ha comparar las principales herramientas de software que existen en el mercado para la implementación de un Data Warehouse y se elegirá una para dar soluciones a las necesidades planteadas por el personal del centro de documentación. Además, los métodos empíricos son necesarios para la recolección de requerimientos en ésta fase por lo cual, se realizará entrevistas con los expertos en los procesos que se realicen en el centro de documentación.

Posteriormente en la fase de planificación se creará un modelo multidimensional que tenga la capacidad de proveer respuestas para ayudar en la toma de decisiones en el centro de documentación y así ayudar a tener un manejo eficiente de los recursos.

En la siguiente fase, la fase de ejecución se realizará la implementación del Data Warehouse/Data Mart donde se llevarán a cabo procesos de extracción, transformación y carga de datos para posteriormente crear el modelo de datos. Además, en esta fase se creará un prototipo básico de Bibliomining que permita extraer la información del Data Warehouse/Data Mart creado con anticipación.

Finalmente, en la fase de pruebas se examinará el correcto funcionamiento del sistema de apoyo en la toma de decisiones para el centro de documentación y se documentarán las conclusiones.

Capítulo 2

Marco Teórico

2.1. Introducción

En este capítulo se aborda los fundamentos teóricos necesarios para la implementación de un datawareshouse incluyendo conceptos básicos sobre las metodologías y herramientas, las fuentes de información que se podrían llegar a analizar y formatos o sistemas de codificación de datos propios de una biblioteca o centro documental.

2.2. Data Warehouse y Data Mart

Data Warehouse

Según Ralph Kimball el Data Warehouse es "una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis" (Kimball y Ross, 2002), es decir el Data Warehouse es una base de datos que integra toda la información proveniente de diferentes fuentes de información obtenidos de ambientes operacionales necesarios de una empresa, para generar reportes y poder analizar los datos de forma estratégica.

Bill Inmon uno de los primeros autores del Data Warehouse lo define en términos de las características de los repositorios afirmando que es: "Una colección de datos que sirve de apoyo a la toma de decisiones, organizados por temas, integrados, no

volátiles y en los que el concepto de tiempo varía respecto a los sistemas tradicionales" (Inmon, 2005).

Una definición más completa es dada por Bernabeu: "El Data Warehouse posibilita la extracción de datos de sistemas operacionales y fuentes externas, permite la integración y homogeneización de los datos de toda la empresa, provee información que ha sido transformada y sumarizada, para que ayude en el proceso de toma de decisiones estratégicas y táctica" (Bernabeu, 2010), por ende el Data Warehouse tiene la finalidad de hacer que la información esté integrada y accesible ayudando en la toma de decisiones dentro de una organización.

Entre las características de un Data Warehouse según Bernabeu están: (Bernabeu, 2010)

- Orientado a un tema
- Administra grandes cantidades de información
- Guarda información en distintos repositorios
- Condensa y agrega información
- Integra y asocia información
- Ayuda en la decisión estratégica
- Permite explotar la información histórica existente

Las ventajas del Data Warehouse según Bernabeu son: (Bernabeu, 2010)

- Posibilita la extracción de datos de sistemas operacionales y fuentes externas.
- Permite la integración y homogeneización de los datos de toda la empresa.
- Provee información que ha sido transformada y sumarizada, para que ayude en el proceso de toma de decisiones estratégicas y tácticas

Data Mart

Es un subconjunto del Data Warehouse que se enfoca en solucionar una área específica de la organización, la cual consta con las mismas características que un Data Warehouse. Según Inmon "un Data Mart dependiente es un subconjunto lógico (vista) o un subconjunto físico (extracto) de un almacén de datos más grande" (Inmon, 2005)

2.2.1. Modelo Multidimensional de un Data Warehouse

Un modelo multidimensional es una base de datos en donde su información se almacena en forma multidimensional, es decir, a través de tablas de hechos y tablas de dimensiones.

- **Tabla de Dimensiones:** Definen como están los datos organizados lógicamente y proveen el medio para analizar el contexto del negocio, contienen datos cualitativos.
- **Tabla de Hecho:** Las tablas de hechos contienen hechos que serán utilizados por los analistas de negocio para apoyar el proceso de toma de decisiones, conformado generalmente por datos cuantitativos. Los hechos son datos instantáneos en el tiempo, que son filtrados, agrupados y explorados a través de condiciones definidas en las tablas de dimensiones.

Al momento de realizar el modelamiento de un Data Warehouse se puede optar por distintas formas de relacionar la tabla de hecho y las tablas de dimensiones, entre las cuales están(Bernabeu, 2010):

- **Esquema Estrella:** Consiste en una tabla de hechos central y las tablas de dimensión están relacionadas mediante claves. En este modelo los datos deben de estar totalmente normalizados. (Ver figura 2.1)
- **Esquema Copo de Nieve:** Es una extensión del esquema en Estrella, esta posee una tabla de hechos central y las tablas de dimensión están relacionadas a este mediante claves, pero a su vez las tablas de dimensión pueden relacionarse con otras tablas de dimensión. (Ver figura 2.2)

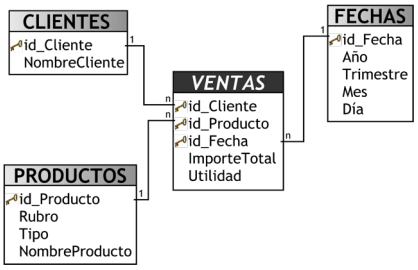


Figura 2.1: Esquema Estrella

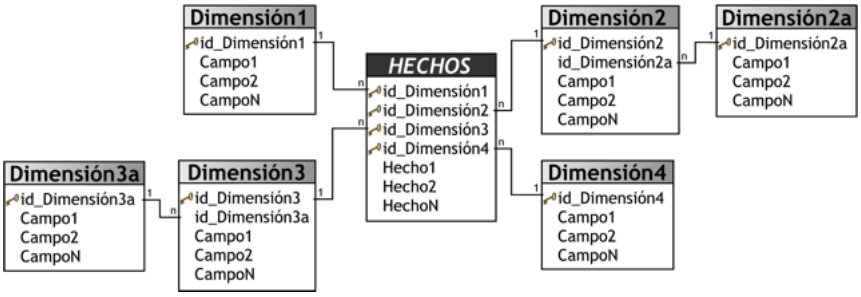


Figura 2.2: Esquema Copo de Nieve

- **Esquema Constelación:** Compuesta por una tabla de hechos central relacionado con otras tablas de hechos, donde cada tabla de hechos posee sus propias dimensiones. (Ver figura 2.3)

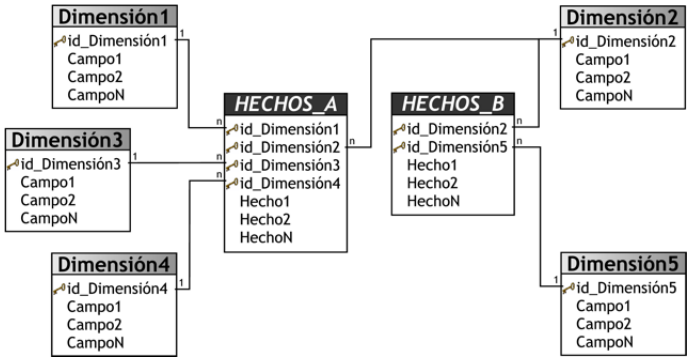


Figura 2.3: Esquema Constelación

2.3. Fuentes de información

Todas las empresas disponen de varias fuentes de información sean estas bases de datos, archivos de texto, entre otros. El centro de documentación en particular disponde de una base de datos documental diferente a las bases de datos relacionales conocidas por los informáticos.

2.3.1. Bases de datos documentales

Las bases del tipo IsisDB son bases de tipo documental, donde la información se almacena en archivos planos. Los índices de los mismos están almacenados en forma de un árbol balanceado. La ventaja de manejar bases de datos documentales frente a una relacional, es la rapidez al momento de la entrega de resultados cuando se realiza una consulta (Bigio, 1998).

Organización del Archivo

Los registros almacenados en las bases de datos IsisDB siguen el formato MARC21 (Machine - Readable Cataloging) que permite almacenar la información en forma de etiquetas, este formato se detalla en la sección 2.3.4.1.

Arquitectura ISIS

La figura 2.4 describe la forma de acceso por parte del usuario final a los repositorios de la información, donde:

- .fdt: Contiene la definición de los campos a mostrar en la interfaz.
- .fmt: Contiene la estructura a mostrar en la interfaz.
- .pft: Formato de impresión de la información (similar al .css de una página web).
- .mst: Archivo maestro
- .fst: Contiene la definición de los campos a ser indizados

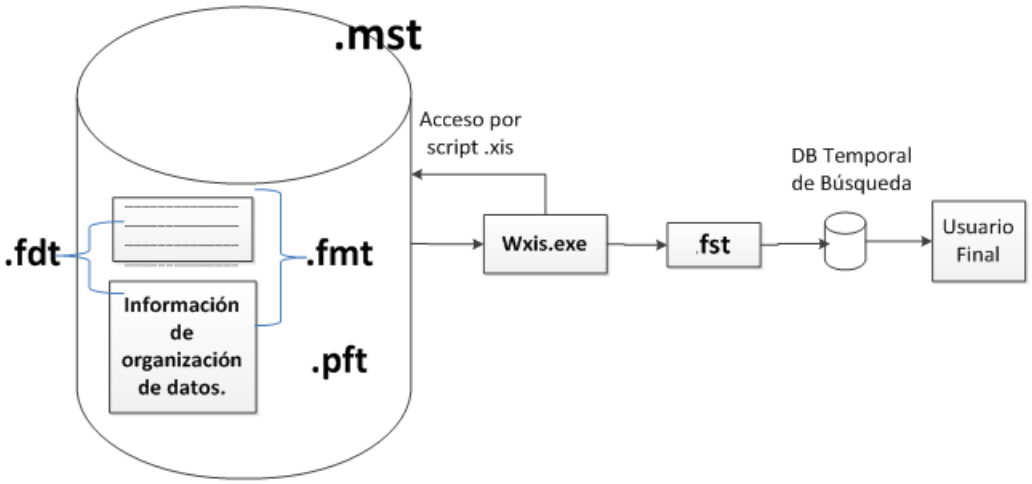


Figura 2.4: Arquitectura Base de Datos ISIS

Wxis.exe construido en C++, es el encargado de acceder al achivo maestro sobre el cual se realiza la carga de información. Por cual, wxis.exe contiene las herramientas de conexión a los diferentes formatos y bases de datos documentales.

2.3.2. Bases de datos relacionales

Es un modelo propuesto a fines de los años 60 por Edgar Frank Codd de IBM, basado en tablas y relaciones interconectando los datos que permite representar la información del mundo real de una forma intuitiva. Además, es el modelo más empleado en la actualidad debido a su versatilidad, potencia y formalismos matemáticos en los que se basa. (Sánchez, 2004)

La terminología básica involucrada al modelo relacional es la siguiente (Sánchez, 2004):

- **Tupla:** Cada fila de la tabla
- **Atributo:** Cada columna de la tabla
- **Grado:** Número de atributos de la tabla
- **Cardinalidad:** Número de tuplas de una tabla
- **Dominio:** Conjunto válido de valores representables por un atributo

2.3.2.1. Gestor de Base de Datos

Un sistema de gestión de base de datos es un programa que permite introducir, almacenar, ordenar y manipular los datos. Actúa como interfaz entre los programas de aplicación y el sistema operativo (Cobo, 2008). Algunos de los gestores de bases de datos relacionales son: MySQL, PostgreSQL, Oracle, entre otros.

2.3.2.2. MySQL

MySQL es un gestor de bases de datos relacional, multiplataforma, multihilo y multiusuario de la empresa Oracle desde el 2009. Se oferta bajo licencia GPL o Uso comercial.(Okamura, 2012)

2.3.2.3. PostgreSQL

PostgreSQL es un potente sistema gestor de bases de datos relacional, orientado a objetos, de código abierto y multiplataforma. Creado por una comunidad de desarrolladores que trabajan de forma desinteresada. Es muy fiable y sólido permitiendo incluso el almacenamiento de grandes objetos como imágenes, sonidos y videos. (PostgreSQL, 2013)

2.3.2.4. Oracle

Es un ORDBMS(Object Relational Database Management System) sistema de gestión de base de datos objeto-relacional multiplataforma de licencia privativa desarrollado por Oracle Corporation (Juan PAblo, 2007). Es considerado uno de los sistemas de bases de datos más completos (Oracle, 2013).

2.3.3. Archivos

Los archivos informáticos son un conjunto de bits equivalentes de los archivos escritos en libros, papel o fichas usados en las oficinas de cualquier institución. Facilitan la organización de los recursos usados para almacenar permanentemente datos en un sistema informático, con gran capacidad de almacenamiento. Son identificados por un nombre y una dirección de la carpeta que lo contiene (Tenzer, 2007).

Todos los archivos en los distintos formatos existentes se convierten en una fuente de información válida para la creación de un Data Warehouse, sean estos log, archivos de texto, xml, entre otros. Éstos deben disponer información ordenada para poder identificar los metadatos sin ningún tipo de problema.

2.3.3.1. Log's

En los archivos log's se registran datos de eventos ocurridos durante un período de tiempo determinado. Estos archivos se generan con la finalidad de almacenar evidencia sobre algún acontecimiento en particular (Tenzer, 2007). Posteriormente estos archivos pueden ser interpretados para generar cuadros estadísticos que proporcionen información útil sobre algún proceso en particular.

2.3.3.2. Archivos de Texto

Son archivos que no usan ningún tipo de codificación, el texto se almacena directamente permitiendo ser legible por personas (Tenzer, 2007). La información almacenada es simple, no tiene ningún formato ni tipo de letra elegida por el usuario.

2.3.4. Formato de Datos

Las bibliotecas hacen uso de ciertos formatos de datos que deben ser explicados teóricamente para entender la estructura o el significado de los datos.

2.3.4.1. MARC21

Un registro MARC(MAchine- Readable Cataloging) es un registro catalográfico legible por máquina, es la información que se presenta en una ficha de catálogo de una biblioteca incluyendo una descripción de un ítem, el asiento principal, los asientos secundarios y la signatura topográfica.

La signatura topográfica tiene como propósito colocar el material bibliográfico de un mismo tema en un mismo lugar, para lo cual se utiliza los esquemas de clasificación del Sistema Decimal de Dewey o de la Biblioteca del Congreso (LC).

El formato MARC21 está compuesto por:

- **Campo:** Unidad lógica de un registro bibliográfico que se divide en uno o varios subcampos.
- **Etiqueta:** Número de tres dígitos que identifica al campo.
- **Indicador:** Son los dos caracteres siguientes a la etiqueta, cada indicador puede contener un valor numérico del 0 al 9 cuyo significado se detalla en la documentación MARC21
- **Subcampo:**Los subcampos estan identificados por un código de subcampo y los delimitadores. Los códigos están formados por una letra minúscula, y los delimitadores son símbolos utilizados para separar los diferentes subcampos. Los campos del 001 al 009 no tienen subcampos y se denominan campos de control(ControlFields).
- **Designador de contenido:** Se usan para hacer referencia a un conjunto de etiquetas, los indicadores y los códigos de subcampo.

La figura 2.5 muestra las partes de un registro con la etiqueta 300, los indicadores vacíos "##" y la información en los subcampos "a,b y c".

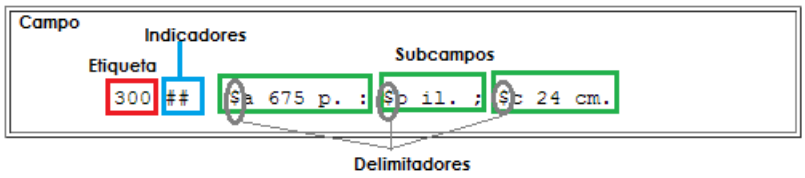


Figura 2.5: Ejemplo de un registro MARC21

Las etiquetas se dividen en centenas, como se presenta en la tabla 2.1 . Los 9XX se reservan para incluir información local de acuerdo a las necesidades.

La cabecera es una cadena de 24 caracteres que se encuentra al inicio del registro antes de la definición de campos y subcampos. (Ver tabla: 2.2)

El estado del registro de la cabecera con formato MARC21 del ejemplo de la figura 2.3 es un material textual corregido, en concreto es una monografía cuyo nivel de codificación es abreviado y su forma de catalogación no es ISBD como datos más relevantes.

Etiqueta	Descripción
0XX	Números, información y códigos de control
1XX	Asiento principal
2XX	Títulos, edición, pie de imprenta
3XX	Descripción física, etc.
4XX	Mención de serie(tal como se presenta en el libro)
5XX	Notas
6XX	Asientos secundarios temáticos
7XX	Asientos secundarios(autores y títulos)
8XX	Asientos secundarios de serie (formas normalizadas)

Tabla 2.1: Etiquetas MARC21

2.3.4.2. Dublin Core

El formato Dublin Core respaldado por la organización Dublin Core Metadata Initiative (DCMI), es usado para describir un recurso Web de manera estándar sin importar el formato, tipo o especialización (Weibel, 2005).

Este sistema de metadatos está definido por un total de 15 etiquetas, las cuales se describen a continuación (Weibel, 2005):

- DC.Title.- Título del recurso a almacenar.
- DC.Subject.- Palabras claves que describe el recurso.
- DC.Description Descripción del contenido del recurso.
- DC.Source.- Recurso al que pertenece el documento.
- DC.Languaje.- Idioma del recurso.
- DC.Relation.- Referencia a un recurso relacionado con el contenido.
- DC.Coverage.- Ámbito del contenido del recurso.

Caracter	Descripción	Ejemplo de posibles valores	Generada por el sistema
00-04	Longitud del registro lógico		Si
05	Estado del registro	c-corregido revisado d - suprimido n-nuevo	Si
06	Tipo de registro	a - material textual c - músic notada e - material cartográfico j - Grabación sonora musical	
07	Nivel bibliográfico	a - parte monográfica b - parte seriada c - colección m - monografía	
08	Tipo de control	# - no especifica a - archivístico	
09	Esquema de codificación de carateres	# - MARC-8 a - UCS/UNICODE	Si
10	Longitud de los indicadores		Si
11	Longitud del código de subcampo		Si
12-16	Posición de inicio de los datos		Si
17	Nivel de codificación	# - nivel completo z - nivel incompleto 3 - nivel abreviado u - desconocido	
18	Forma de codificación	# - no es ISBD a - AACR 2 c - ISBD sin puntuación u - desconocido	
19	Nivel de registro de un recurso	# - no especifica a - conjunto b - parte con título independiente	Si
20-23	Estructura del directorio	Campos fijos 4,5,0,0 respectivamente	Si

Tabla 2.2: Cabecera MARC21

Posición	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Contenido	0	0	8	5	3	c	a	m			2	2	0	0	2	6	3		a		4	5	0	0

Tabla 2.3: Ejemplo de la cabecera MARC21

- DC.Creator.- Autor de la creación del contenido.
- DC.Publisher.- Entidad responsable de la creación del recurso.
- DC.Contributor.- Colaborador en la creación del contenido del recurso.
- DC.Rights.- Derechos de la propiedad intelectual del recuso.
- DC.Date.- Fecha de creación o modificación del recurso (aaaa-mm-dd).

- DC.Type.- El tipo del recurso a almacenar.
- DC.Format.- Detalles físicos del recurso.
- DC.Identifier.-Referencia del recurso.

Las etiquetas se usan en formato html ya que se utilizan en un entorno Web, un ejemplo de la utilización de las etiquetas se muestra a continuación:

```
<meta name="DC.Format" content="video/mpeg; 10 minutes">

<meta name="DC.Language" content="en" >

<meta name="DC.Publisher" content="publisher-name" >
```

2.3.4.3. Sistema de Clasificación Decimal Dewey

El joven bibliotecario Melvil Dewey formuló en 1876 la clasificación decimal Dewey también denominada CDD, que permite clasificar los recursos bibliográficos de una biblioteca.(Segundo, 1996)

Dewey es un esquema numérico dividido en disciplinas, para asignar una disciplina se debe conocer el contenido del material bibliográfico. Según Morimer los sistemas de clasificación presentan una estructura con los siguientes elementos(Carreón y Figueroa, 2009)

- Esquema: Representa a detalle las categorías del universo del conocimiento.
- Notación: Son números, letras y/o otros símbolos usados para representar las divisiones principales y subordinadas de un esquema.
- Índice: Se encarga de enlistar alfabéticamente los términos que han sido incluidos en los esquemas junto con su notación correspondiente.

Las disciplinas en las que puede ser catalogado el material bibliográfico en el nivel más alto son:

- 000: Generalidades

- 100: Filosofía y Psicología
- 200: Religión
- 300: Ciencias Sociales
- 400: Lenguaje
- 500: Ciencia
- 600: Tecnología
- 700: Arte y Recreación
- 800: Literatura
- 900: Historia y Geografía

Se puede tener más niveles de especificación, básicamente se usan tres niveles de detalle pero se puede usar cuatro o cinco separando las tres primeras cifras con un punto. Si es necesario utilizar más de seis cifras, se deja un espacio libre por cada tres cifras. La figura 2.6 muestra un ejemplo con 8 niveles de detalle.

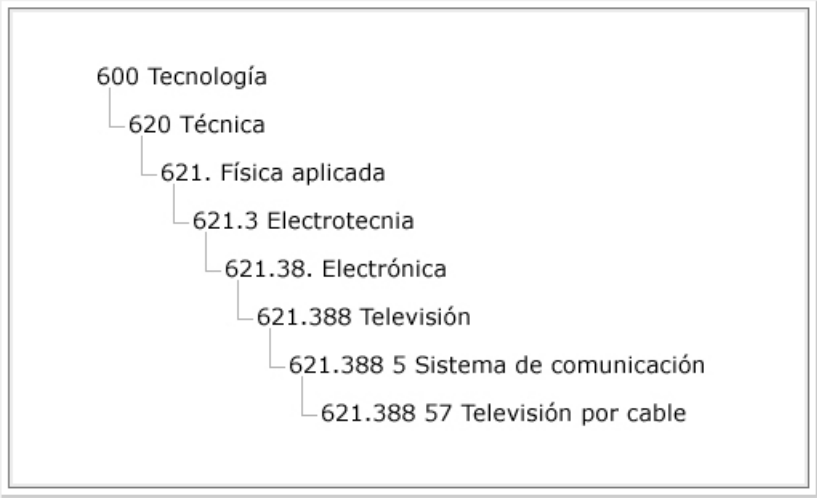


Figura 2.6: Ejemplo Dewey

2.4. Metodologías de desarrollo del Data Warehouse

Cada fabricante de software de inteligencia de negocios busca imponer una nueva metodología con sus productos. Cada metodologia de desarrollo del Data Warehouse posee sus ventajas y limitaciones. Sin embargo, se imponen entre la mayoría las metodologías de Kimball, la de Inmon y Hefesto.

2.4.1. Metodologia Bill Inmon

Inmon propone una metodología Top-down que transfiera la información de los diferentes OLTP (Procesamiento de Transacciones en Línea) a un repositorio centralizado(Inmon, 2005), considerando que éste tendrá algunas características:

- Orientado a temas
- Variante en el tiempo
- No volátil
- Integrado

Los datos extraídos se almacenan en una estructura de datos en tercera forma normal después de haber sido depurados, de la cual los Data Marts de cada departamento de la empresa obtienen su información como lo muestra la figura 2.7 (Fuente: Wordpress, Roberto Espinosa, Consultor SAP)

2.4.2. Metodologia Ralph Kimball

Este enfoque también se referencia como Bottom-up, permite construir un Data Warehouse de forma escalonada basándose en los procesos de negocio, considerando los Data Marts para construir un Data Warehouse, es decir Kimball propone: partir de los datos y procesos existentes para modelar un Data Warehouse que se adapte a ellos con el objetivo de lograr eficiencia en tiempo (Rivadera, 2010).

La arquitectura de la metodología se muestra en la fig 2.8 (Fuente: Wordpress, Roberto Espinosa, Consultor SAP)

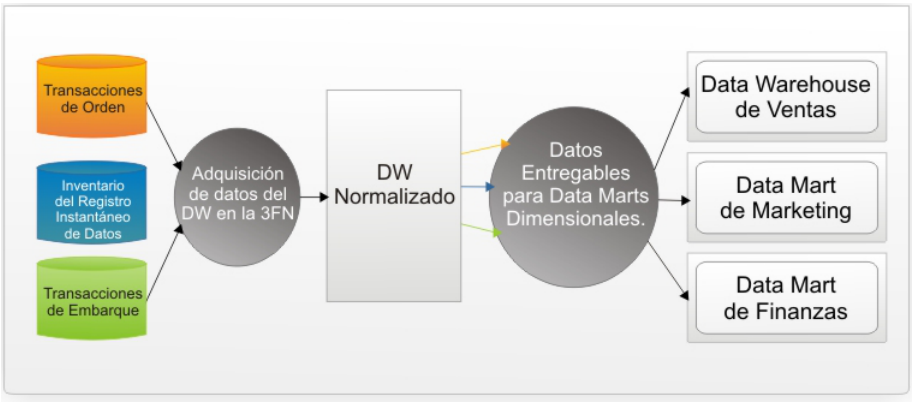


Figura 2.7: Metodología Inmon

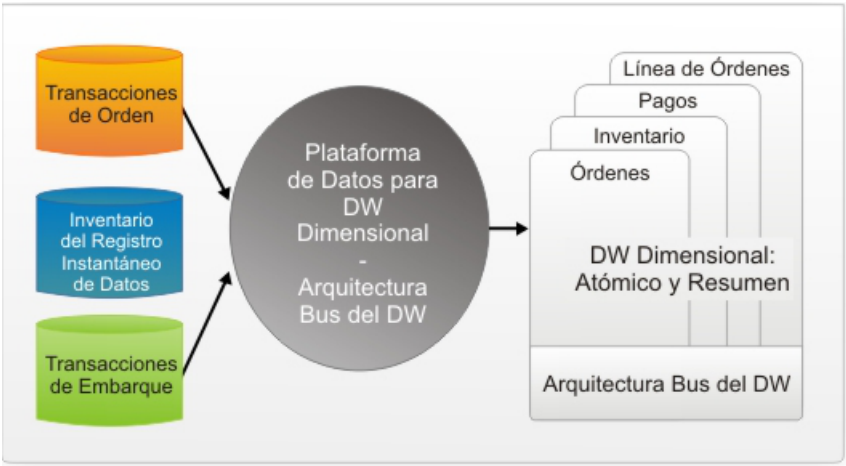


Figura 2.8: Metodología Kimball

Los distintos Data Marts están conectados entre sí por un bus que permite que los usuarios puedan realizar consultas a los diferentes Data Marts ya que este bus contiene los elementos en común que los comunican.

2.4.3. Metodología Hefesto

Las dos metodologías mencionadas anteriormente dan un enfoque diferente para implementar un Data Warehouse. Una metodología adicional que combina los objetivos de las metodologías anteriores se denomina Hefesto creada por Darío Bernabeu. La metodología Hefesto parte de la recolección de requerimientos de información

del usuario, seguido de los procesos de extracción, transformación y carga de datos (ETL) hasta definir un esquema lógico para la organización ya sean estos Data Marts o Data Warehouse.
Bernabeu indica que esta metodología cuenta con las siguientes características (Bernabeu, 2010):

- Los objetivos y resultados esperados en cada fase se distinguen fácilmente y son sencillos de comprender.
- Se basa en los requerimientos del usuario, por lo cual su estructura es capaz de adaptarse con facilidad y rapidez ante los cambios en el negocio.
- Reduce la resistencia al cambio, ya que involucra a los usuarios finales en cada etapa para que tome decisiones respecto al comportamiento y funciones del Data Warehouse.
- Utiliza modelos conceptuales y lógicos, los cuales son sencillos de interpretar y analizar.
- Es independiente del tipo de ciclo de vida que se emplee para contener la metodología.
- Es independiente de las herramientas que se utilicen para su implementación.
- Es independiente de las estructuras físicas que contengan el Data Warehouse y de su respectiva distribución.
- Cuando se culmina con una fase, los resultados obtenidos se convierten en el punto de partida para llevar a cabo el paso siguiente.
- Se aplica tanto para Data Warehouse como para Data Mart.

Hefesto es una metodología que contempla cuatro pasos para la construcción de un Data Warehouse, como se indica en la figura2.9 (Bernabeu, 2010)

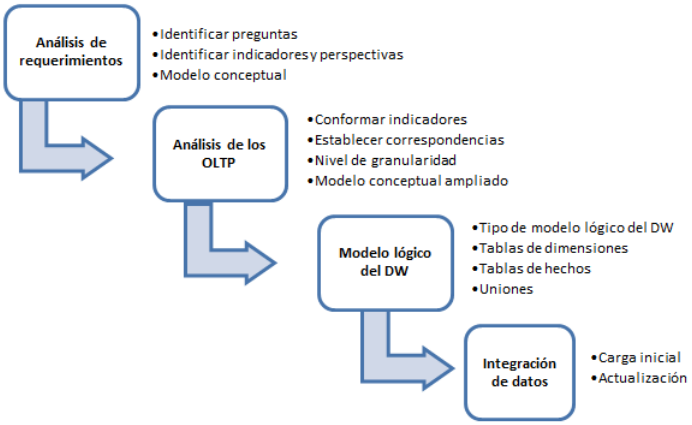


Figura 2.9: Pasos de la metodología Hefesto

A continuación se describen los pasos de la metodología Hefesto(Bernabeu, 2010):

1. ***Análisis de requerimientos***

   Identificar preguntas: Punto de partida en el que los usuarios guían el desarrollo ya que se recolecta requerimientos utilizando varias técnicas: entrevistas, cuestionarios, observaciones, etc. Considera la identificación de las necesidades de información clave de alto nivel soportada por alguna fuente de información.

   Identificar indicadores y perspectivas: Se debe descomponer las preguntas de negocio para descubrir indicadores y perspectivas. Los indicadores son valores numéricos y representan lo que se desea analizar como por ejemplo: saldos, promedios, cantidades, sumatorias, fórmulas, etc. Las perspectivas son objetos mediante los cuales se quiere examinar los indicadores como por ejemplo: clientes, proveedores, países, productos, etc.

   Modelo conceptual: A partir de los indicadores y perspectivas identificadas construir un modelo conceptual que permite comprender los resultados a obtener sin tener conocimientos previos.

2. ***Análisis de los OLTP***

   Conformar indicadores: En este paso se indica cómo se calcularán los indicadores definidos anteriormente indicando los hechos que lo componen y la

función de sumarización que se utilizará para la agregación por ejemplo: SUM, AVG, COUNT, etc.

Establecer correspondencias: Examinar las fuentes de información disponibles que contengan la información requerida para identificar las correspondencias con el modelo conceptual.

Nivel de granularidad: Seleccionar los campos que contendrá cada perspeciva, no es necesario considerar todos los datos solo se debe seleccionar los más relevantes para las consultas.

Modelo conceptual ampliado: Agregar al modelo conceptual creado en el paso anterior los atributos de cada perspectiva.

3. ***Modelo lógico del Data Warehouse***

   Tipo de modelo lógico del Data Warehouse: Seleccionar el esquema que se adapte mejor a los requerimientos y necesidades de los usuarios. Definir objetivamente si se empleará un esquema en estrella, constelación o copo de nieve ya que afectará considerablemente la elaboración del modelo lógico.

   Tablas de dimensiones: Diseñar las tablas de dimensión que formarán parte del Data Warehouse, cada perspectiva definida en el modelo conceptual constituirá una tabla de dimensión.

   Tablas de hechos: Definir la tabla de hechos que contendrá los hechos a través de los cuales se constiuirán los indicadores.

   Uniones: Definir la relación entre las tablas de dimensiones y las tablas de hechos.

4. ***Integración de datos***

   Carga inicial: Poblar el modelo lógico con datos utilizando técnicas de limpieza y calidad de datos, procesos ETL, etc. Primero cargar los datos de las dimensiones y luego los datos de las tablas de hechos.

   Actualización: Establecer políticas y estrategias de actualización.

2.5. Herramientas para implementar un Data Warehouse

En el mercado existe una gran gama de herramientas para implementar el Data Warehouse conocidas como herramientas de Inteligencia de Negocio (Business Intelligence BI). Cada una de las cuales están formadas por un paquete de aplicaciones integradas. Entre ellas están las del tipo comercial como: Oracle Business Intelligence Server One, IBM Cognos, Microsoft SQL Server - Suite de Herramienta de BI , etc. Y las de tipo Open Source como: Eclipse BIRT Project, JasperReports, Pentaho entre otras.

2.5.1. Pentaho BI

Pentaho BI es una suite de software orientada a la solución con componentes de BI que permiten a la empresa desarrollar soluciones integradas al problema Pentaho Open Source Business Intelligence Platform Technical White PaperCopyright,2006. Su plataforma se basa en flujos de trabajos, procesos y definición de procesos las cuales pueden ser integradas fácilmente.

Debido a que es una completa gama de programas integrados, la arquitectura de Pentaho se basa en servidores, motores y componentes muchos de ellos estándares; ofreciendo una plataforma de BI escalable y sofisticada que combina componentes de código abierto y código fuente escrita por desarrolladores de Pentaho.

Adicionalmente, es posible integrar software de terceros (Ver figura 2.10, Fuente: Pentaho Open Source Business Intelligence), entre los principales componentes que conforman la suite de pentaho están:

- Mondrian (Open Source OLAP Server).- Servidor OLAP.
- JFreeReport (Open Source Reporting).- Herramientas de Reportes.
- Kettle (Open Source Data Integration).-Integración de Datos.
- Weka (Open Source Data Mining).- Minería de Datos.

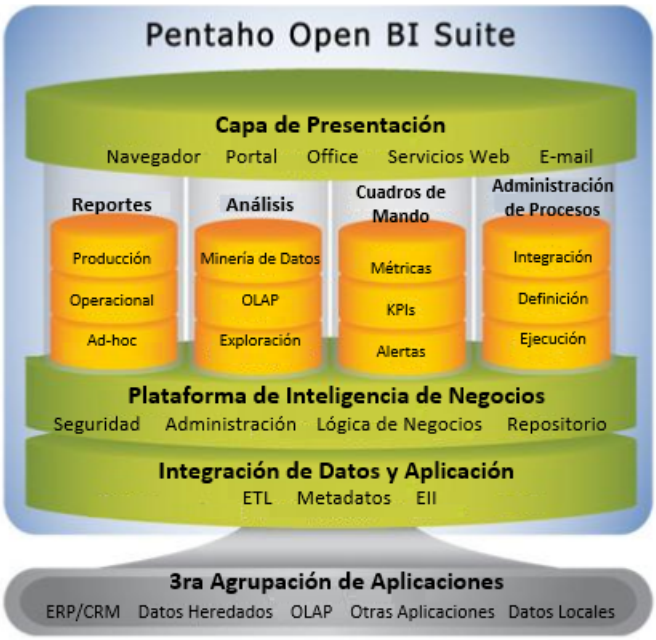


Figura 2.10: Suite de Pentaho

2.5.2. Jasper

Maneja un modelo de negocio del tipo comercial de código abierto ofreciendo informes, cuadros de mando, análisis, y servicios de integración de datos para los requisitos de BI tanto autónomos y embebidos con una arquitectura flexible y moderna, construida en un modelo escalable para que sea integrable con otras aplicaciones.

Jasper ofrece productos como se presenta en la figura 2.11 [1]:

2.6. Data Mining y Bibliomining

En esta sección se presentan los conceptos de Data Mining y Bibliomining que sustentan teóricamente los procesos realizados posteriormente.

[1] Jaspersoft, Jaspersoft Embedding Guide(2013)

Figura 2.11: Suite de Jasper

2.6.1. Data Mining

Data mining es un término que ha atraído gran atención sobre la información de las empresas y de la sociedad en general en los últimos años. La minería de datos es aplicada sobre grandes volúmenes de datos (Theodoulidis, 2003), generados por las empresas a lo largo del tiempo para convertir esa información en conocimiento útil (Frank, 2005).

La abundancia de información de las empresas con la utilización de técnicas de limpieza de datos e integración de los mismos, hace posible generar conocimiento oculto que deteminan patrones de comportamiento de los procesos empresariales.

Objetivos de Data Mining Los objetivos para el cual se desarrolla el Data Mining propuestos por Virseda-Roman son (Benito y Carrillo, 2010):

- **Descripción:** Al analizar una gran cantidad de datos se pueden descubrir reglas, que aclaran los procesos dentro de una empresa con la finalidad de ayudar en la planificación y la mejora continua.

- **Predicción:** Genera nuevas oportunidades de negocio al predecir tendencias y comportamientos corporativos.

Fases Data Mining

Las fases para el descubrimiento de conocimiento en bases de datos se muestran en la figura 2.12

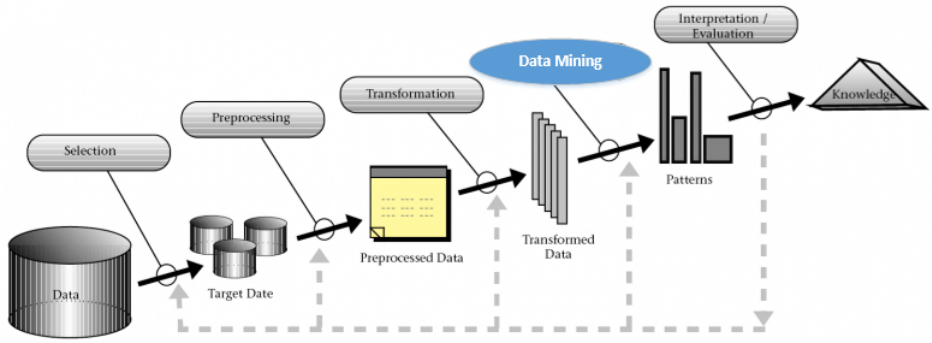


Figura 2.12: Fases del descubrimiento de conocimiento

Las fases son:

- **Selección:** Seleccionar los datos para dar solución al problema.
- **Prepocesamiento**: Limpieza de datos donde cierta información innecesaria es quitada.
- **Transformación**: Transformar los datos necesarios antes de ser transferidos.
- **Data Mining**: Extracción de patrones.
- **Interpretación y Evaluación**: Los patrones identificados son interpretados en conocimiento y usado para soportar la toma de decisiones.

Hay dos factores críticos para tener éxito con las técnicas de Data Mining:

- Disponer de un Data Warehouse amplio y bien integrado
- Comprender claramente el proceso comercial al cual las técnicas de Data Mining van a aplicarse.

2.6.2. Bibliomining

Bibliomining es una combinación de los términos biblioteca y Data Mining, utilizado por primera vez por Nicholson quien lo define como la aplicación de la minería de datos en el ámbito bibliotecario. Bibliomining es aplicar diversos algoritmos de Data Mining "para descubrir los patrones en grandes cantidades de datos en bruto que puedan proporcionar los patrones necesarios para crear un modelo y una ayuda automatizada de la colección del material bibliográfico" (Stanton, 2003).

Está relacionada con la biblioteconomía y documentación ya que pretende que los resultados obtenidos de la aplicación de las técnicas de minería de datos sean útiles para entender las comunidades de usuarios o bien para aplicarlos en cualquier entorno relacionado con el ámbito biliotecario.

Biblioming se describe como un conjunto de distintas técnicas basadas en la estadística utilizadas para el análisis de la información generada por las bibliotecas para extraer conocimiento que ayude a tomar decisiones y/o a solucionar problemas.

Para aplicar las técnicas de Biblioming es necesario seguir un conjunto de pasos: (Nicholson, 2003)

- Determinar el área de interés
- Identificar las fuentes de datos internas y externas
- Depositar los datos en un Data Warehouse
- Seleccionar herramientas de análisis apropiadas
- Descubrir patrones utilizando Data Mining o procesos tradicionales
- Analizar e implementar los resultados

Las bibliotecas han recopilado datos sobres sus colecciones y usuarios por años, pero no han sabido utilizar esos datos para la toma de decisiones. El Biblimining que integra en esfuerzos actuales de la investigación y de la evaluación permitirá a encargados e investigadores de las bibliotecas una idea más completa de los recursos contenidos en sus organizaciones y cómo están siendo alcanzados por los usuarios. Así tomando un acercamiento más activo basado en usos de minería de datos de la

visualización de los mismos, y de la estadística, estas organizaciones de información pueden conseguir una visión más clara de las necesidades de entrega y de la gestión de la información.

Una de las ventajas de Data Mining es que puede aprovecharse de cantidades grandes de datos y la información descubierta con el uso de técnicas de Bibliomining da a la biblioteca el potencial de ahorrar dinero, proporcionar programas más apropiados, resolver más necesidades de información de los usuarios, observar problemas de su colección y sirve como fuente de información más eficaz de sus usuarios.

Capítulo 3

Análisis y Diseño de un Data Warehouse para el Centro de Documentación "Juan Bautista Vázquez"

3.1. Introducción

El primer paso antes de empezar con el análisis para la implementación del Data Warehouse es la selección de la metodología, a partir de la cual se debe realizar la recolección de requerimientos en el centro documental para comprender las necesidades de información que se tengan y diseñar la solución más adecuada, considerando que los requerimientos esten acorde a la información disponible en las diferentes fuentes de datos que se dispongan. Para un mejor entendimiento de la aplicación de la metodología se realizará la implementación de la misma sobre el **Proceso de Préstamos** definiendo un modelo multidimensional, este proceso se repite para las demás actividades del centro documental llegando a definir un modelo multidimensional general. La aplicación de la metodología seleccionada en los demás procesos identificados en el desarrollo del Data Warehouse se presenta en los anexos.

3.2. Selección de la metodología para desarrollar un Data Warehouse

Los enfoques presentados por Inmon y Kimball tienen algunas diferencias, todo depende de los objetivos de negocio de una organización para elegir una u otra metodología, en la tabla 3.1 se presenta una comparativa de las metodologías mencionadas.

INMON	KIMBALL
Fácil mantenimiento	Orientado a los procesos de negocio
Modelo normalizado	Modelo dimensional
Integración de datos de toda la empresa	Integración de datos de las áreas de negocio
Complejo de diseñar	Tiempo de implementación corto

Tabla 3.1: Comparación de metodologías: Inmon y Kimball

Inmon presenta una alternativa para empresas estables que cuenten con el tiempo y los recursos necesarios para la implementación de un Data Warehouse, mientras que, el enfoque que Kimball propone una implementación rápida orientado a los procesos de negocio formando al Data Warehouse como un conglomerado de Data Marts.

La metodología seleccionada para la implementación del Data Warehouse de este proyecto de titulación es Hefesto, debido a que ofrece una metodología que integra los conceptos de Inmon como de Kimball, dando énfasis a los requerimientos de los usuarios permitiendo así que los resultados persigan los objetivos definidos al momento del planteamiento del Data Warehouse.

3.3. Aplicación de la Metodología Hefesto

En esta sección se aplicará el análisis de requerimientos de la metodología de Hefesto para los procesos internos como externos identificados en el desarrollo del Data Warehouse. Este análisis se desarrolla de manera global, sin profundizar en proceso alguno.

3.3.1. Definición de los requerimientos del negocio

El primer paso de la metodología Hefesto es el análisis de los requerimientos que en una biblioteca o centro de documentación se enfocará en proporcionar a los usuarios una amplia variedad de material bibliográfico y los servicios relacionados. Diariamente se genera información que puede ser de utilidad en la toma de decisiones relacionadas a la asignación de recursos tales como: económicos, personal, tiempo e infraestructura.

Al integrar los datos se tiene una mejor perspectiva de los procesos llevados a cabo en el centro documental de manera que éstos influyan en las decisiones con mayor certeza.

Los requerimientos que generan el proceso de integración están basados en cuatro cuadrantes, los cuales son analizadas por Siguenza-Guzman (Siguenza-Guzman et al., 2013) quien describe un estudio de implementación del proceso holístico, afirmando las ventajas y beneficios de este análisis en una biblioteca, los cuadrantes mencionados son:

1. **PRIMER CUADRANTE:** Es un análisis interno de la biblioteca que considera los recursos materiales y económicos utilizados, se orienta a la parte financiera analizando los costos y recursos consumidos por los procesos del centro documental. (Siguenza-Guzman, 2013b)

   Un estudio realizado en Estados Unidos (Ellis-Newman et al., 1996) acerca de los costos siguiendo los métodos tradicionales determina que en una biblioteca se realizan gastos directos tales como el consumo de recursos y horas laborales y gastos indirectos, tales como costos de mantenimiento, marketing, depreciación, capacitaciones e incluso gastos de electricidad. El problema de estos sistemas de costo tradicional es que deben ser adecuados cuando los gastos indirectos son bajos y los servicios son limitados (Ellis-Newman y Robinson, 1998).

   ABC (*Activity-Based-Costing*) es un sistema de costeo más avanzado propuesto por Cooper y Kaplan en 1998 que realiza un tratamiento eficiente de los costos indirectos, asigna un costo a cada una de las actividades realizadas las cuales pertenecen a un servicio que presta la biblioteca. La suma de los costos de las actividades determina el costo operacional de un servicio determinado, como:

catalogación, préstamos, adquisición, entre otros (Ellis-Newman, 2003). Sin embargo, el problema está en la subjetividad de los empleados de la biblioteca al estimar los tiempos dedicados a las actividades realizadas y el alto costo para la recolección de datos (Siguenza-Guzman et al., 2013).

Para superar estas limitaciones Kaplan y Anderson desarrollan el método de costeo TDABC, que es un ABC que para cada actividad posee ecuaciones de costos, las cuales son calculadas en base al tiempo requerido para llevarlas a cabo. El Centro de Documentación "Juan Bautista Vázquez" posee un sistema de costeo que sigue esté método TDABC para analizar los costos de los procesos y servicios evaluando los tiempos y recursos que se asignan a éstos.

2. **SEGUNDO CUADRANTE:** Considera la usabilidad y evaluación de los servicios bibliotecarios, enfocándose en la calidad de los mismos. Sigüenza-Guzmán recomienda el uso de al menos uno de los cinco métodos de evaluación: la recopilación de estadísticas, buzones de sugerencias, las pruebas de usabilidad web, facilidad de uso de la interfaz de usuario y la satisfacción mediante encuestas (Lorena et al., 2014).

   Matthews (Matthews, 2013) analiza la necesidad de combinar los datos sobre el uso de la biblioteca y sus servicios con otros datos disponibles en el campus universitario (fuentes externas a la biblioteca).

3. **TERCER CUADRANTE:** Evalúa la concurrencia, el acceso y las citaciones a documentos digitales, además, analiza las necesidades de los usuarios en relación del material bibliográfico existente en el centro de documentación, la información se puede recolectar de dos maneras (Nicholson, 2003):

   - A través del contacto directo con los usuarios con el fin de informarse que material bibliotecario es necesario para ellos.
   - Por el contacto indirecto a través del análisis bibliométrico.

4. **CUARTO CUADRANTE:** Se enfoca en descubrir patrones de comportamiento por parte de los usuarios. Los patrones a analizar se basan en acceso al material digital y las búsquedas realizadas sobre el material bibliográfico, las cuales se almacenan en registros Log's.

		Topic	
		Library System	Collection
Perspective	Internal (Library)	[1] What does the library system consist of?	[4] How is the library system manipulated?
	External (Users)	[2] How effective is the library system?	[3] How useful is the library system?

Figura 3.1: Matriz holística de Nicholson

A partir de los cuatro cuadrantes analizados por Nicholson (Nicholson, 2004) (Ver figura 3.1), en base a las necesidades del centro de documentación se tiene el modelo estudiado por Siguenza-Guzman (Siguenza-Guzman, 2013a), como se muestra en la figura 3.2. La cual adapta los requerimientos de un centro de documentación en una perspectiva holística considerando la matriz holística de Nicholson.

		Topic	
		Library System	Collection
Perspective	Internal (Library)	**Cost analysis** Processes, Time, Resources	**Log analysis** Implicit and explicit data
	External (Users)	**Quality** Statistics gathering, Suggestion boxes, Usability testing, Satisfaction surveys	**Bibliometrics** Citations patterns, Publishing patterns, Journals downloaded, and Journals' impact factor

Figura 3.2: Matriz holística de Siguenza-Guzmán and Leuven and Belgium

3.3.2. Análisis de las Fuentes de Datos

Una vez definidos y analizados los requerimientos, el segundo paso según la metodología Hefesto es el análisis de los OLTP (OnLine Transaction Processing). Los OLTP son fuentes de información orientadas al procesamiento de transaciones, que involucran ingresos, actualizaciones y eliminaciones de datos. Para el caso del centro de documentación analizado se manejan varios procesos que generan transacciones, entre las principales actividades que se desarrollan día a día y que producen una gran

cantidad de transacciones están: Préstamos, Catalogación, Evaluación de Servicios, Servicio al Usuario, Adquisiciones, entre otras como se presenta en la figura 3.3.

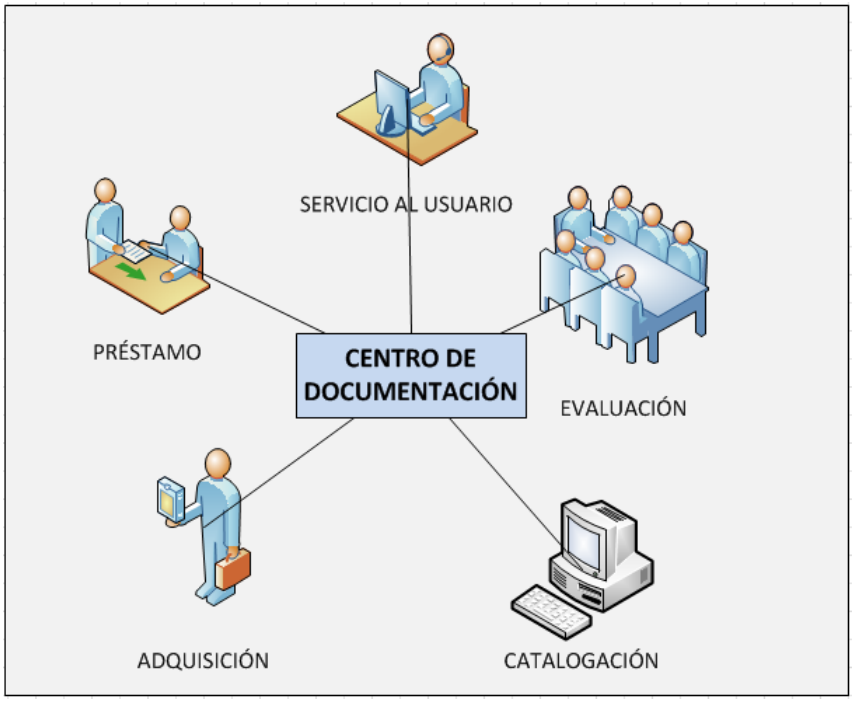


Figura 3.3: Procesos diarios

Los datos a los que se tiene acceso internamente son los relacionados a los procesos de: Catalogación, Préstamos y Evaluación de servicios, la figura 3.4 muestra las fuentes de información involucradas en dichos procesos.

El centro de documentación posee fuentes propias de información que puede obtener y explorar por sus propios medios y recursos sin necesidad de acudir a terceros, ya sean fuentes generadas por su propia gestión o aquellas fuentes que han sido elaboradas por alguien en un momento dado pero que están disponibles en la organización. Además, dispone de acceso a otras fuentes externas provenientes del DTIC (Departamento de Tecnologías de la Información y Comunicación) de la Universidad de Cuenca, relacionadas con ciertos procesos académicos y administrativos como lo muestra la figura 3.5:

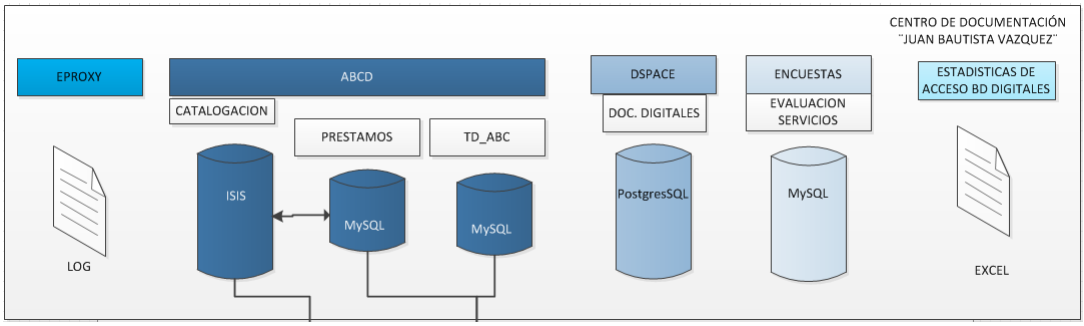


Figura 3.4: Fuentes de Información Internas

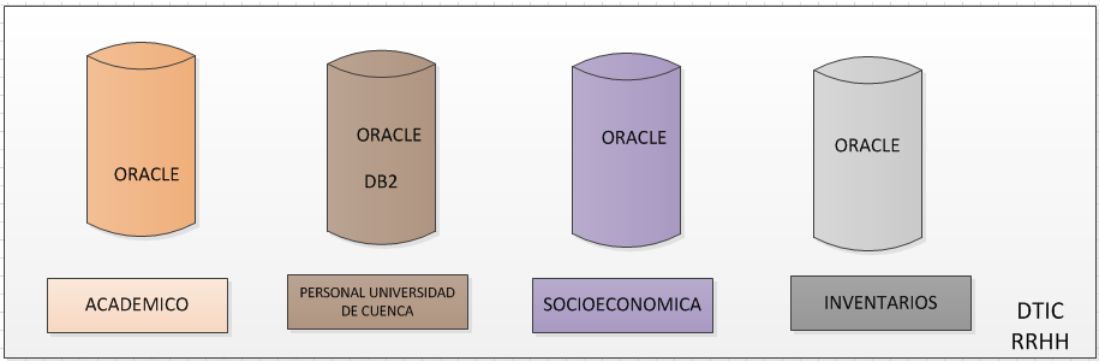


Figura 3.5: Fuentes de Información Externas

3.3.2.1. Fuente de información: Servidor EZproxy

Descripción

Software desarrollado por Chris Zagar en 1999 quien fundó Useful Utilities, posteriormente en el 2008 OCLC (Online Computer Library Center) adquirió el producto. [1] EZproxy es un servidor proxy que facilita el acceso remoto al contenido con licencia que ofrecen las bibliotecas, permite a los usuarios del centro de documentación conectarse a sistemas de autentificación local y acceder a la base de datos bibliográfica remota a la que se encuentre suscrito.

Características

- Provee un acceso intermedio transparente para los usuarios.

[1] http://jobs.code4lib.org/jobs/ezproxy/?page=3

- Acceso a un gran número de proveedores.
- Conexión a una gran variedad de servicios de autentificación.
- Reduce el número de autorizaciones y passwords.

Repositorio de Datos.- Archivos de extensión ".log".

Diagrama de Datos

Una muestra de datos del log del Centro de Documentación "Juan Bautista Vázquez" se muestra en la figura 3.6:

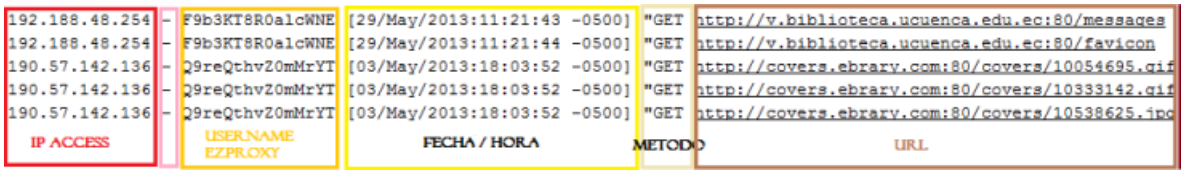


Figura 3.6: Muestra del Log EZproxy

Definición de Atributos

EZproxy mantiene un archivo log estándar de uso en el servidor web, cada vez que un usuario accede a una página web a través de EZproxy se registra una entrada en el archivo *ezproxy.log*, lo cual permite evaluar la cantidad de accesos a las distintas bases de datos para tomar decisiones en la renovación de licencias de estas bases. Los accesos se registran en el archivo log con un determinado formato por defecto, teniendo la posibilidad de agregar o quitar la información que será requerida. Por ejemplo, si se accede a la página http://www.somedb.com/index.html, la entrada del log común sería como:

```
132.174.1.1 - - [14/Mar/2008:09:39:18 -0700]
"GET http://www.somedb.com:80/index.html HTTP/1.0" 200 1234
```

El formato común definido en el archivo *config.txt o ezproxy.cfg* es:

```
LogFormat %h %l %u %t "%r" %s %b
```

Campo	Descripción
h	Dirección IP del host de acceso
l	Username remoto
u	Username usado para iniciar sesión en EZProxy
t	Fecha/Hora de conexión
r	URL de la solicitud
s	Código de estado de la solicitud
b	Número de bytes transferido

Tabla 3.2: Campos del Formato LOG EZProxy

Y los campos se describen en la tabla 3.2:

Otros campos que pueden agregarse para que sean almacenados en el archivo log son los que se presentan en la tabla 3.3:

Campo	Descripción
a	Dirección IP del host EZProxy
m	Método de solicitud (GET,POST)
v	Nombre del host del servidor web virtual
expressione	Evalúa la expresión EZProxy
ezproxy-sessioni	Identificador del usuario actual

Tabla 3.3: Campos adicionales del registro LOG EZProxy

Relación con otras Aplicaciones

Como se muestra en la figura 3.7, EZProxy está relacionado con todas las bases de datos virtuales con las que se enlace el centro de documentación, actualmente se tiene acceso a las siguientes:

- IEEE Xplore
- SpringerLink
- Taylor Francis

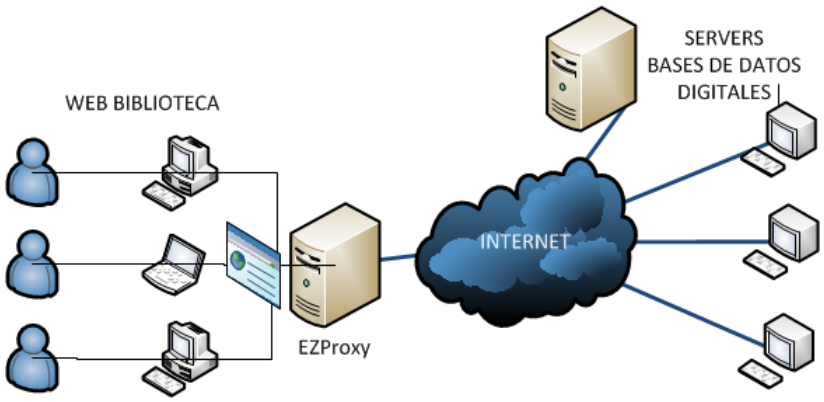


Figura 3.7: Diagrama EZProxy

- Ebrary
- Elibro
- IEEE Xplore
- EBSCO Research Databases
- ProQuest
- Prisma
- Gale Cengage Learning
- OAIster
- Directory of Open Access Journals
- INASP International Network for the Availability of Scientific Publications
- E-prints in Library and Information Science
- Flacso
- Scholar Google
- CiteSeerX

- Social Science Research Network
- EconPapers
- Dialnet
- EtnasSoft
- Biblioteca Digital Mundial
- El libro total
- World Turism Organization

Ubicación.- Los archivos log's son propios de los procesos internos del centro de documentación.

Responsable.- Ing. Andrés de los Reyes

Sistema Operativo.- Centos

Volumen.- Se genera un archivo log mensualmente de aproximadamente 1.7 Mb.

Acceso a Datos.- Intranet

3.3.2.2. Fuente de información: ABCD

Descripción

ABCD es un software open source desarrollado por BIREME y VLIR, es el acrónimo de Automatización de Bibliotecas y Centros de Documentación. ABCD es una aplicación Web que integra varios paquetes y herramientas para la automatización de bibliotecas que comprende las siguientes funciones:

- Catalogación
- Préstamos
- Adquisiciones
- Estadísticas
- OPAC (Catálogo en línea)

- Importar registros bibliográficos a través del protocolo Z39.50
- Servicios bibliotecarios como SDI (Selective Dissemination of Information), impresión de código de barras, control de calidad, etc

ABCD esta conformado por los módulos que se presentan en la figura 3.8 (de Sorporte Tecnico BVS)

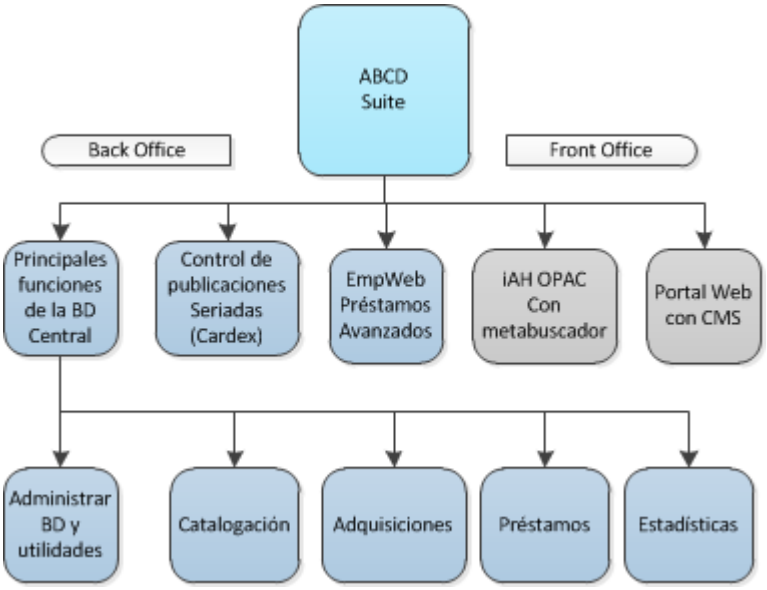


Figura 3.8: Suite ABCD

Características

ABCD es una herramienta integrada de gestión de bibliotecas que posee bases de datos Isis, tecnología Isis Script, Java Script, PHP, MySQL, Apache y YAZ. La aplicación ABCD puede manejar los registros de los siguientes tipos de materiales bibliográficos: (Dhamdhere, 2011)

- Material cartográfico
- Manuscritos
- Material lingüístico
- Archivos de computadora

- Gráficos
- Material audiovisual
- Registros musicales
- Registros sonoros no musicales
- Material seriado
- Entre otros

3.3.2.3. Fuente de información: Catalogación

Descripción

ABCD provee un módulo de catalogación para crear cualquier estructura bibliográfica como MARC21, UNIMARC y CEPAL, permitiendo personalizar los campos y subcampos que se ingresarán para cada registro bibliográfico.

Además, provee la posibilidad de importar registros bibliográficos de otras bibliotecas o centros documentales a través del protocolo Z39.50 facilitando el trabajo de los bibliotecarios y permitiendo usar un mismo formato universal de catalogación (Dhamdhere, 2011). Z39.50 es un protocolo que permite compartir registros bibliográficos identificándolos a través de una búsqueda de una serie de servidores Z39.50 como la Biblioteca del Congreso en Washington o la Biblioteca Británica en Londres y posteriormente, permite descargar los registros al sistema local. (de Smet y Spinak, 2009)

El proceso de creación de una nueva base de datos en ABCD sigue el proceso [2] mostrado en la figura (3.9)

Repositorio de Datos.- Bases de datos documental denominada ISISDB en base al formato MARC21.

Diagrama de Datos Como se muestra en la figura 3.10, ISISDB almacena registros los mismos que están compuestos por campos y estos a su vez constan de subcampos que siguen el formato MARC21.

[2]Listas de Sorporte Tecnico BVS, ABCD Automatizacion de Bibliotecas y Centros de Documentacion

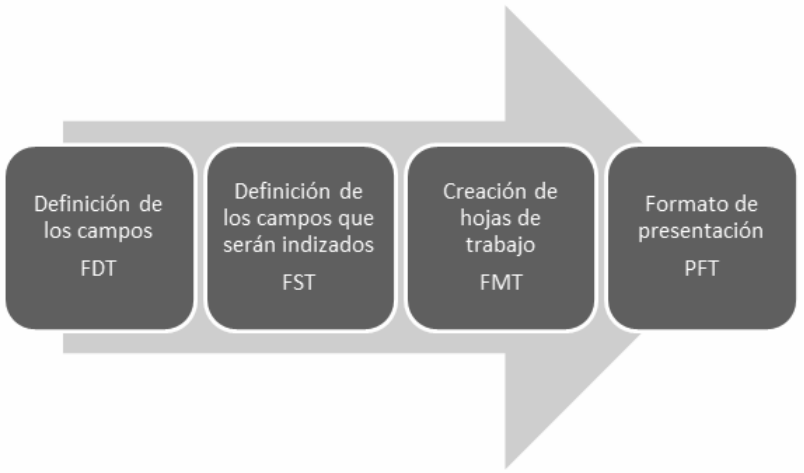


Figura 3.9: Proceso de creación de una base de datos en ABCD

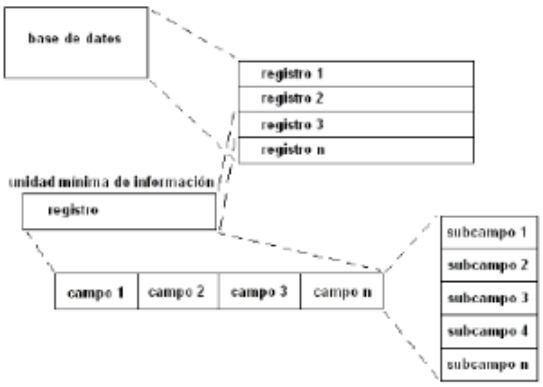


Figura 3.10: Diagrama de datos Isis

Definición de Atributos

En general en el Centro de Documentación "Juan Bautista Vázquez" se utilizan los campos del formato MARC21 que se presenta en la tabla 3.4 para almacenar los registros bibliográficos en la base de datos Isis.

Campo	Descripción	1er. Ind	2do. Ind	subcampo	Ejemplo
020(R)	ISBN			a	
022(R)	ISNN			a	
040(NR)	Centro catalogador			a b	UC-CDJBV Idioma de catalogación
041(R)	Código de idioma	0 no tiene trad 1 Incluye trad		a	
082(R)	CDD Dewey	0 ed. completa ed		a c	Nro. de clasificación Nro. de ingreso
100(NR)	Autor personal	0 nombre directo 1 apellido 2 Nombre de familia		a b d	Apellido, Nombre Numeración Fecha de nacim y muerte
110(NR)	Autor corporativo	0 nombre invertido 1 jurisdicción 2 Nombre directo		a e g	Autor institucional Entidad subordinada Sigla
245(NR)	Título	0 sin asiento secund. 1 con asiento secund	0 al 9	a b c h n p	Título Subtítulo Mensión de respons. Medio físico Número de la parte/Sección de la obra
246(R)	Variante del título	0 sin información 1 parte de título Título paralelo Título definitivo		a b f g	Título Subtítulo Fecha o designación secuenc Información miscelanea
250(NR)	Edición			a b	Edición Resto de la mención de edición
260(NR)	Pie de imprenta			a b c	Lugar de public. Editorial Fecha de public.
300(NR)	Descripción física			a b c e	Nro. páginas Inform. descriptiva Dimensiones Material anexo
310(NR)	Periodicidad			a	Periodicidad
362(R)	Fecha de publicación			a	Fecha de publicación
490(R)	Mensión de series	0 sin asiento secund 1 con asiento secund	0 al 9	a n p v x	Título de la colección Nro. de lla parte Nombre de la parte Volúmen o número Nro. ISBN de la colección
500(R)	Notas			a	Nota general
502(R)	Nota de tesis			a	Nota de tesis
504(R)	Nota de bibliografía			a	Nota de bibliografía y refer.
520(R)	Nota de resumen			a	Resumen
653(R)	Descriptores	0	0	a	Descriptores (lenguaje libre)

Tabla 3.4: Campos MARC21 del CDJBV

Dentro del proceso de catalogación se almacenan también asientos secundarios,

para los cuales en el centro de documentación se reservan ciertos campos como lo muestra la tabla. 3.5.

Campo	**Descripción**	**1er. Ind**	**2do. Ind**	**subcampo**	**Ejemplo**
700(R)	Autor Secundario	0 Nombre directo 1 Apellido Nombre de familia		a b d e	Apellido, Nombre Numeración Fecha de nacim. y muerte Término de relación
710(R)	Autor corporativo	0 Nombre invertido 1 Jurisdicción 2 Directo		a a b e g	Autor institucional Entidad subordinada Entidad subordinada Sigla
773(R)	Título documento fuente		0 al 9	t g	Título de la revista Nro. páginas
856(R)	Archivo PDF (Tesis digital)			a u	Portada PDF Archivo PDF
900(NR)	Existencias			a f k l m n o p q r y	Nro. de volumenes Tipo adquisición Documentalista Nro. ejemplares General,Limitada Cod. Ingreso Ubicación Fecha entrega Precio Nro. días préstamo Nro. inventario

Tabla 3.5: Asientos Secundarios MARC21 CDJVB

Relación con otras Aplicaciones

El módulo de catalogación del sistema ABCD es uno de los principales sistemas del centro de documentación, sobre el cual se generan procesos adicionales. El material bibliográfico una vez adquirido pasa a ser catalogado para posteriormente estar disponible en préstamos, reservas o préstamos interbibliotecarios, esta relación se

observa en la figura 3.11

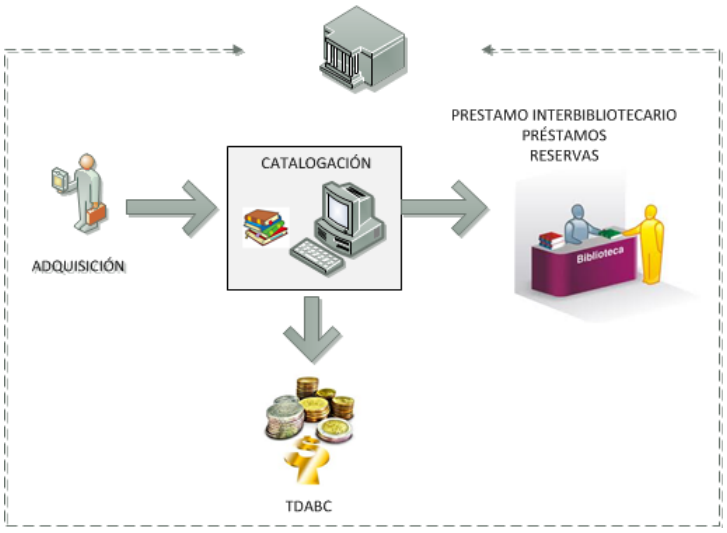


Figura 3.11: Relación de aplicaciones con el módulo catalogación

Ubicación.- Aplicación propia de los procesos internos del centro de documentación

Responsable.- Departamento Informático del CDJBV (Centro de Documentación "Juan Bautista Vázquez")

Sistema Operativo.- Centos

Volumen.- Aproximadamente 100 MB con 157.000 registros.

Acceso a Datos.- Intranet CDJBV

3.3.2.4. Fuente de información: Préstamos

Descripción

ABCD dispone de un modulo para gestionar los préstamos, sin embargo en el Centro de Documentación "Juan Bautista Vázquez" se desarrolló uno propio que se ajuste a sus necesidades, cabe mencionar que este módulo esta integrado en el sistema ABCD.

Repositorio de Datos.- Utiliza MySQL como motor de base de datos.

Diagrama de Datos El diagrama de la base de datos se presenta en la figura 3.12.

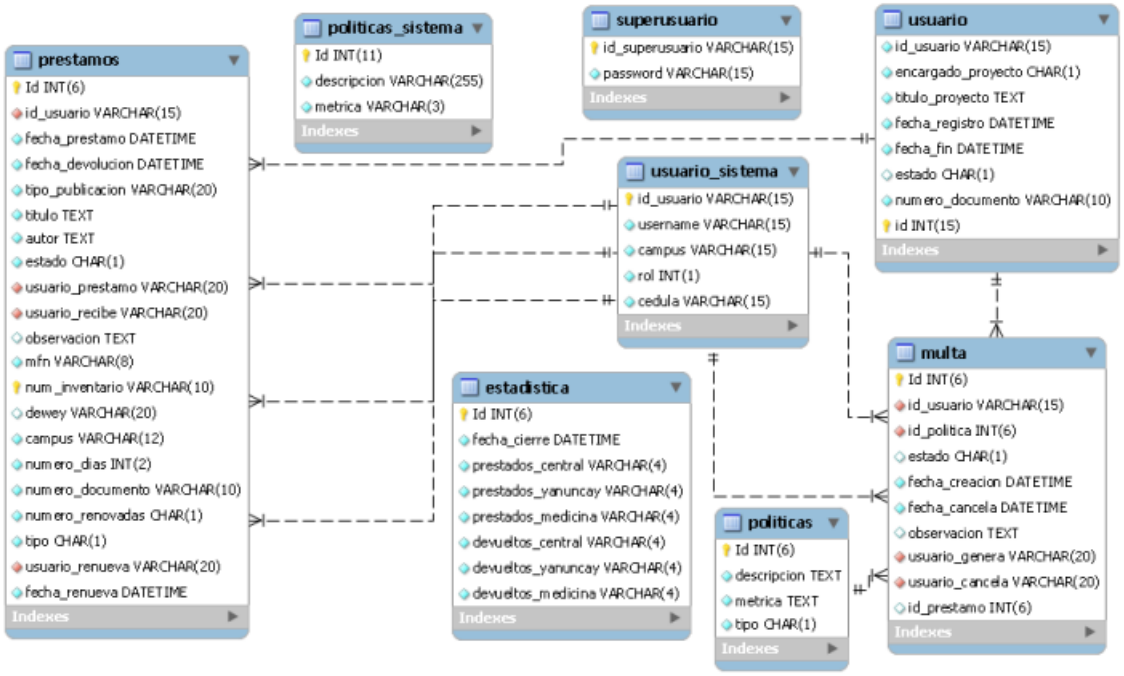


Figura 3.12: Diagrama Entidad-Relacion de Prestamos

Definición de Atributos

Un préstamo está relacionado con el usuario que solicita el servicio, el bibliotecario, datos generales del libro prestado, fecha del préstamo y la devolución entre otros. Tiene un estado, sea este P para cuando está prestado y D cuando el libro ha sido devuelto, además tiene un tipo N (Normal) o E (Especial) cuando existe un oficio que autorice la transacción.

Actualmente los sistemas que dispone el centro documentación no gestionan el cobro de multas, pero como lo muestra el diagrama de la base de datos de préstamos se está considerando implementarlo en el futuro. Los campos que se almacenarán en una tabla o en un archivo se indican en la figura 3.13:

Las multas están relacionadas con un préstamo y con una política, estás políticas tienen una métrica que es un valor entero que representa el valor monetario o el tiempo al que referencia la política. Las políticas son los parámetros a considerar para la gestión de préstamos por ejemplo:

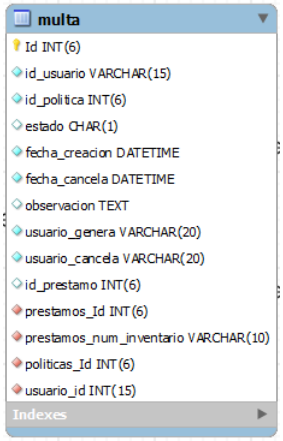


Figura 3.13: Atributos Multa

- Número máximo de préstamos para estudiantes
- Número máximo de préstamos para profesores
- Valor de la multa
- Días de gracia

Relación con otras Aplicaciones

El módulo de préstamos se relaciona con la base de datos Isis para acceder a la información del material bibliográfico y con la base de datos de RRHH que guarda información de los usuarios que solicitan el servicio de préstamos. Además, se debe considerar que el sistema TDABC se relaciona con todos los procesos dentro del centro de documentación y esté no es la excepción, como se observa en la figura 3.14.

Ubicación.- Fuente de datos local para el centro de documentación.

Responsable.- Ing. Mauricio Brito

Sistema Operativo.- Centos

Acceso a Datos.- Intranet

3.3.2.5. Fuente de información: TD-ABC

Descripción

El módulo TD-ABC es open source y utiliza el lenguaje PHP, HTML, CSS, AJAX,

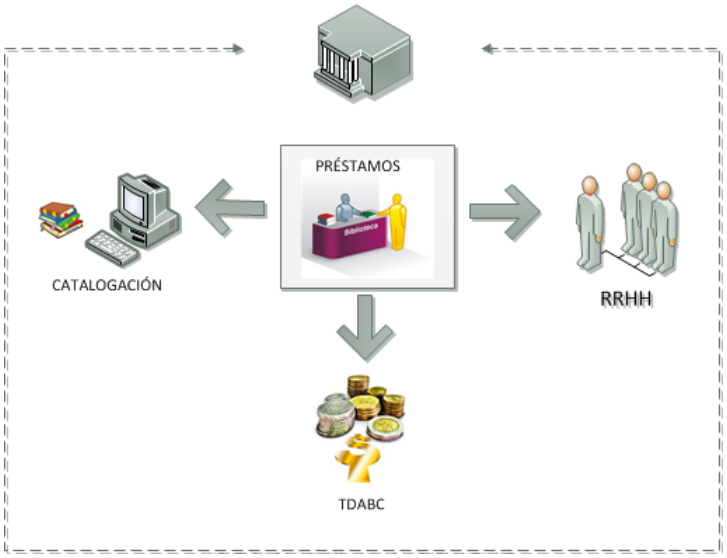


Figura 3.14: Relación de Préstamos con otras aplicaciones

JAVASCRIPT y API GOOGLE. Es un módulo adicional del sistema ABCD, es un sistema de costeo basado en actividades realizadas en un determinado tiempo y utiliza MySQL como motor de base de datos. La aplicación permite llevar un control de los costos operacionales de la biblioteca, considerando los recursos que se asignan a las diferentes actividades bibliotecarias para ayudar a la toma de decisiones administrativas, sobre todo las relacionas con el presupuesto (Ordoñez y Cabrera, 2012).

El sistema TD-ABC implementado en el Centro de Documentación "Juan Bautista Vázquez", hace el uso de dos parámetros para el análisis de costos:

1. Los costos de suministrar los recursos
2. Estimar el tiempo requerido para llevar a cabo una actividad

Repositorio de Datos.- MySQL

Diagrama de Datos

El diagrama entidad-relación de la base de datos del módulo TD-ABC se presenta a continuación en la figura 3.15

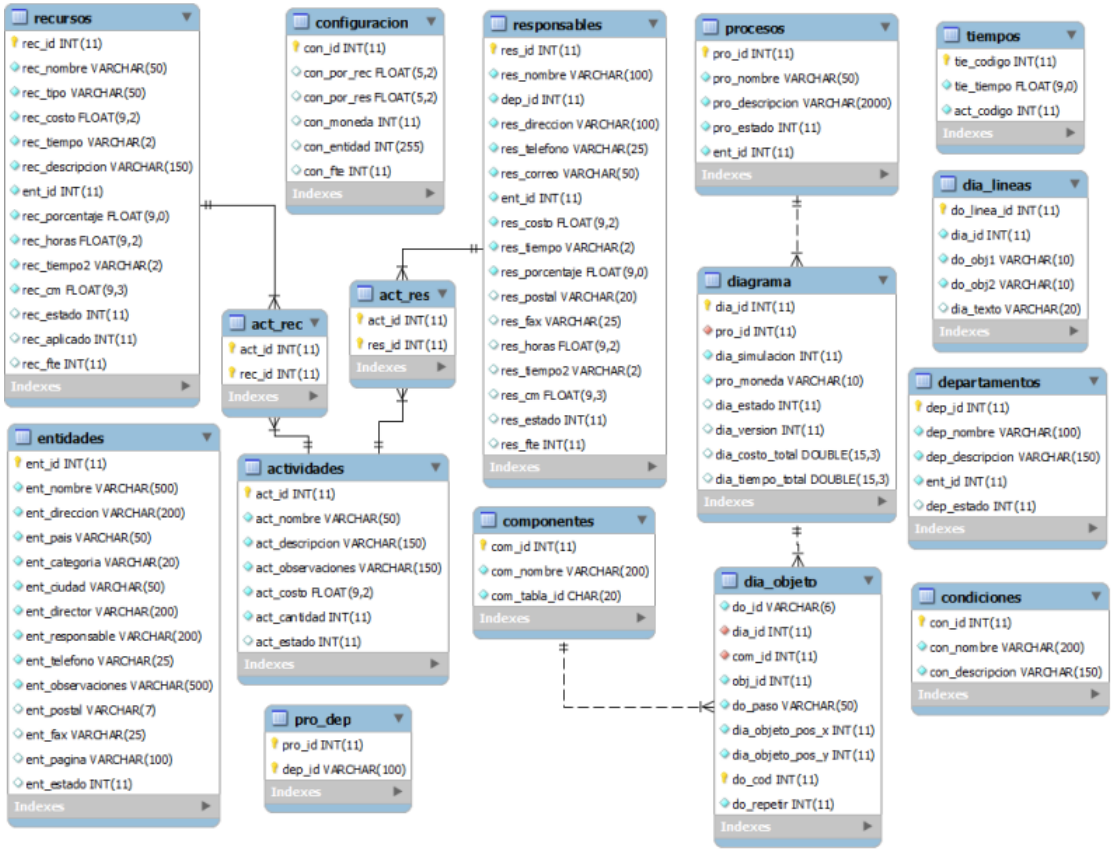


Figura 3.15: Modelo Entidad-Relacion TD-ABC

Definición de Atributos

En el centro de documentación se manejan procesos, estos procesos constan de actividades que tienen asignados recursos y responsables. Estas actividades tienen un costo monetario y una duración individual. En la tabla *diagrama* se almacena el costo total y el tiempo total de un proceso. Como ejemplo uno de los procesos de la biblioteca es la adquisición del material bibliográfico para lo cual se realizan las siguientes actividades:

- Recibir una solicitud de libros de los Decanos de cada Facultad, dichos libros vendrán de las necesidades de alumnos y profesores de las escuelas que forman la facultad

- Presupuestar el pedido y verificar si está dentro de la capacidad de gasto de la

biblioteca

- Autorizar o negar la compra
- Realizar el pedido del libro
- Pagar el monto del libro
- Codificar los libros de acuerdo a su ubicación y asignatura
- Ingresar los libros a la base de datos y catálogos digitales
- Colocar en la estantería

Cada una de las actividades mencionadas tiene un costo operacional, al sumar los costos de todas las actividades relacionadas a un proceso en particular se determina el costo operacional de un proceso. Las actividades de un proceso se definen mediante un diagrama de procesos en la aplicación web TD-ABC, como se muestra en la figura 3.16:

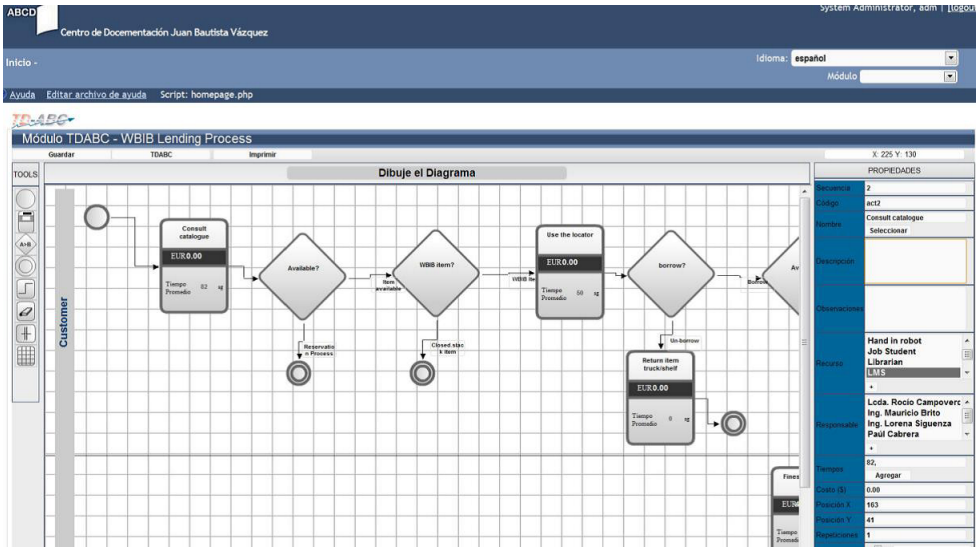


Figura 3.16: Definir Procesos TD-ABC

Relación con otras Aplicaciones

En la figura 3.18 se presenta la relación del módulo TD-ABC con las otras fuentes de información.

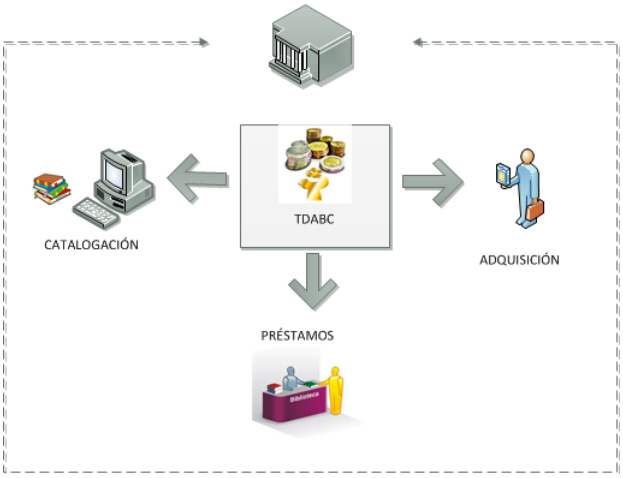


Figura 3.17: Relación de TD-ABC con otras aplicaciones

Ubicación.- Módulo aún no puesto en producción

Responsable.- Ing. Lorena Siguenza

Sistema Operativo.- Módulo aún no puesto en producción

Acceso a Datos.- Se dispone localmente el acceso al esquema de la base de datos, con datos de prueba.

3.3.2.6. Fuente de información: Reservas

No se gestiona reservas de material bibliográfico en el Centro de Documentación "Juan Bautista Vázquez" pero es un proceso muy común en este ámbito y un principal servicio para los usuarios por lo que debe ser considerado para que en un futuro sea implementado.

El proceso de reservas se lleva a cabo de la siguiente manera: al solicitar un ejemplar, el cual ha sido prestado a otro usuario y no se dispone físicamente en el centro de documentación se debe registrar una reserva, la cual asegura al usuario solicitante que recibirá algún tipo de alerta cuando el material biblográfico que necesita sea devuelto para que éste pueda pedir un préstamo inmediatamente.

Después de analizar algunos sistemas bibliotecarios que permitan la gestión de reservas, se considera que los principales datos a almacenar son los siguientes:

- Campus
- Id Ítem
- Código Dewey
- Fecha Reserva
- Prioridad
- Id Usuario que Solicita
- Correo Usuario
- Teléfono Usuario

3.3.2.7. Fuente de información: Préstamos Interbibliotecarios

El préstamo interbibliotecario es un tipo especial de préstamo que consiste en compartir material bibliográfico con otras bibliotecas. El Centro de Documentación "Juan Bautista Vázquez" no brinda este servicio a los usuarios pero se plantea un esquema con los principales atributos a considerar:

- Id Item
- Nivel de Servicio
- Medio Solicitado
- Biblioteca Externa
- Tipo Préstamo
- Fecha Préstamo
- Fecha Devolución

ESTADÍSTICAS DE USO

INSTITUCIONES DE EDUCACIÓN SUPERIOR	BASE DE DATOS	ENERO		FEBRERO		MARZO		ABRIL	
		Búsquedas	Sesiones	Búsquedas	Sesiones	Búsquedas	Sesiones	Búsquedas	Sesion
Universidad de Cuenca	CENGAGE LEARNING	2.350	526	338	136	860	368	5.513	1
	EBRARY	1.434		233		188		2.380	
	EBSCO	6.646	1.230	1.410	464	5.572	1.167	8.761	2
	ELIBRO	6.116		2.608		2.824		17.080	
	PROQUEST	10.176		11.088		83.095		73.927	
	SPRINGER	169		231		302		256	
	TAYLOR	182	168	139	102	325	280	316	
TOTAL POR MES		27.073	1.924	16.047	702	93.166	1.815	108.233	4.(

Figura 3.18: Estadísticas de acceso a las BD digitales

3.3.2.8. Fuente de información: Estadísticas de acceso a las BD digitales

El Centro de Documentación "Juan Bautista Vázquez" está suscrito a bases de datos digitales como Springer, Google Scholar entre otros. Los datos de acceso son obtenidos a través de un correo con las estadísticas de acceso a estos repositorios.

3.3.2.9. Fuente de información: DSpace

Descripción

Software creado en el año 2002 por Massachusetts Institute of Technology (MIT) conjuntamente con HP bajo licencia BSD, es un software de plataforma open source cuyo objetivo es la administración y almacenamiento de archivos digitales. Considerando como archivo digital a todo tipo de documento que contenga información y se presente en cualquier formato, este sistema sirve como un repositorio digital para instituciones y organizaciones.

Características

Las caracteristicas más relevantes son (Team, 2012):

- Almacenar y describir material digital.
- Distribuir material digital mediante búsquedas a través de un sistema web.
- Presentar reportes estadísticos de consultas y administración de los archivos digitales.

- Personalizar la presentación de resultados adaptando el código en base de las necesidades de cada organización.
- Permite el almacenamiento de diferentes tipos de formatos de archivos.
- Permite un control de acceso por parte de los usuarios hacia el documento digital.

Repositorio de datos.- Base de Datos (PostgreSql).

Metadatos

Los documentos digitales se almacenan en el sistema DSpace en base a un conjunto de atributos definidos como metadatos, las cuales permiten definir los datos de cada uno de los recursos que forman un documento digital. Estos documentos digitales pueden ser recuperados fácilmente, permitiendo compartir información con otros repositorios digitales externos a la organización.

Existen tres tipos de metadatos en DSPACE (Team, 2012) basados en el formato Dublin Core:

- Metadatos Descriptivos: metadatos que permite describir cada uno de los documentos digitales.
- Metadatos Administrativos: metadatos que permite describir la preservación, procedencia y los datos de las políticas de autorización.
- Metadatos Estructurales: metadatos que permite estructurar el formato de presentación de los ítems almacenados en el repositorio digital.

Diagrama de la base de datos

La figura 3.19 muestra las tablas y las relaciones que se utilizan en este trabajo.

Definición de Atributos

La base de datos se encuentra estructurada considerando básicamente seis ámbitos para un adecuado manejo de archivos digitales las cuales se describen a continuación:

- Community
- Collection

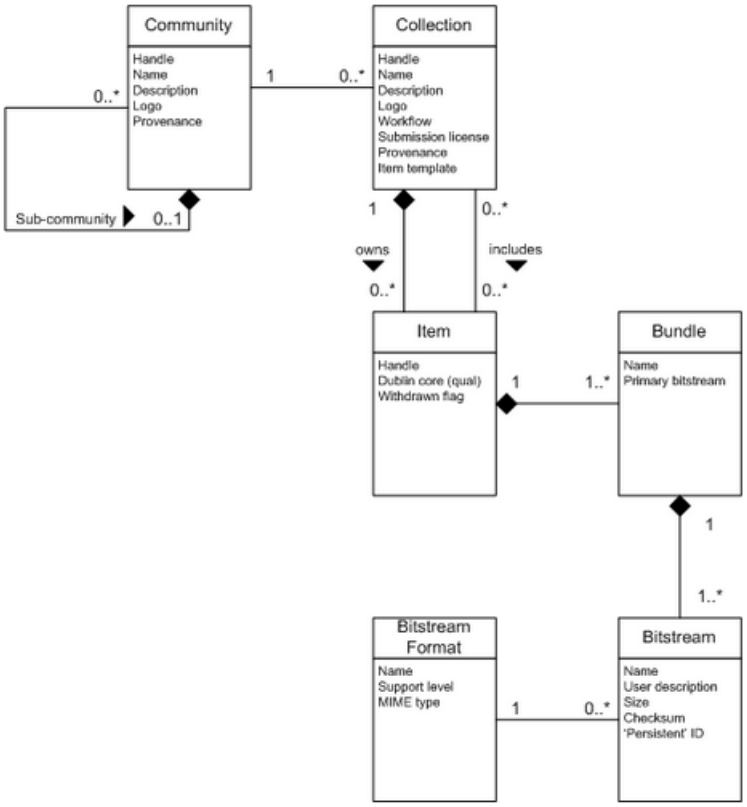


Figura 3.19: Diagrama Entidad - Relación de categorías DSPACE

- Item
- Bundle
- Bitstream
- Bitstream Format

Para un mejor análisis se toma un ejemplo del Sistema DSPACE implementado en el Centro de Documentación "Juan Bautista Vásquez" de la Universidad de Cuenca (figura 3.20).

Community

Toda institución se basa en una división orgánica para una adecuada administración, cada uno de estos departamentos es representado en DSPACE como una *Comunidad*, ásta en DSpace es un espacio o entidad mayor que agrupa unidades administrativas.

Figura 3.20: Ejemplo de DSpace Universidad de Cuenca

Una *comunidad* puede contener *subcomunidades* y *colecciones*. Según nuestro ejemplo, Community es "Facultad de Ingeniería" representada por "A"'. Una Sub Community es "Especializaciones", "Ingeniería Civil" e "Ingeniería de Sistemas" representada por "B".

Collection

Una colección son agrupaciones de ítems que contienen información relacionada ya sea por tipo de registro, materia, especialidad, etc.

Según el ejemplo planteado, Collection es "Documentos de la Facultad de Ingeniería", "Tesis de Pregrado de Ingeniería de Sistemas", que están representados por "C".

Item

Item es el propio archivo que se define un en repositorio digital ya sean estos documentos, libros, trabajos de investigación, etc, los cuales están formadas por un conjunto simple de archivos denominados Bitstream (figura 3.21).

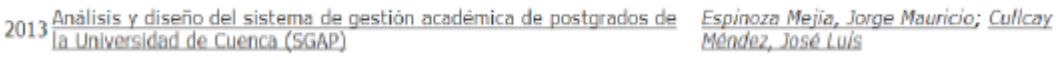


Figura 3.21: Ejemplo Item DSPACE

Bundle

Un grupo de Bitstreams que pertenecen a un mismo Item.

Bitstreams

Son archivos interrelacionados cuya unión forman un Item.

Formato de Bitstream

Cada uno de los Bitstreams posee un *Formato de Bitstream* para que pueda ser interpretado al momento de presentarse el ítem en una consulta. Cada ítem tiene asignado su propietario o un grupo de propietarios que crearon el documento, el cual está representada por la *tabla "bi-2-dis"* dentro de la base de datos del sistema DSPACE como se presenta en la figura 3.22.

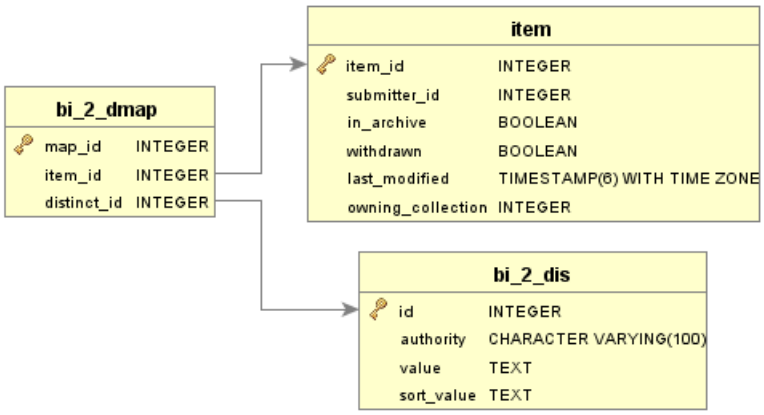


Figura 3.22: Relación tabla Bitstream DSPACE

Además del autor, cada ítem contiene el usuario o grupo de usuarios responsable de la creación o modificación del registro digital en el repositorio, el cual está representado por la *tabla "eperson"* dentro de la base de datos como indica la figura 3.23.

Relación con otras Aplicaciones

DSPACE se relaciona directamente con el sistema TDABC, debido a que el registro de un archivo digital en DSPACE también implica un costo operacional (Ver figura 3.24), similar a la Catalogación en el sistema ABCD.

Ubicación.- El sistema DSPACE forma parte de los procesos internos del centro de documentación.

Responsable.- Ing. Andrés de los Reyes

Sistema Operativo.- Centos

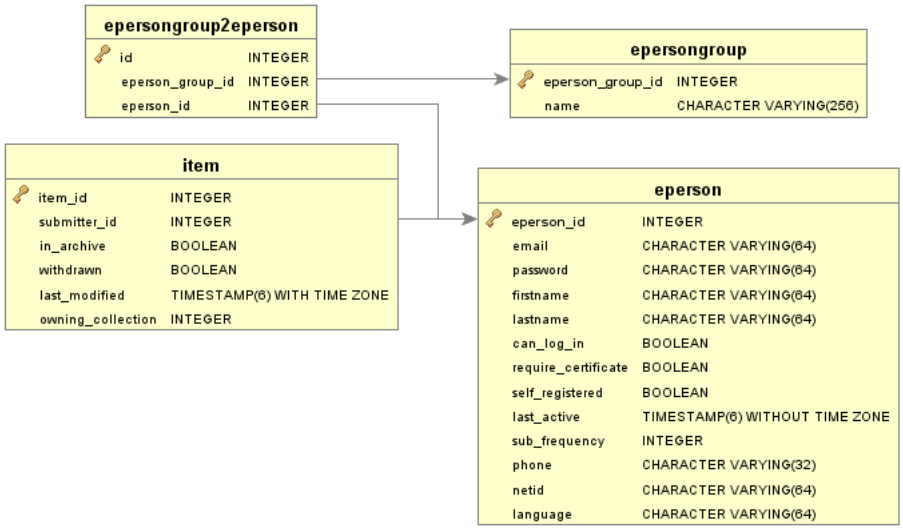


Figura 3.23: Relación tabla eperson DSPACE

Acceso a Datos.- Intranet y mediante el sistema Web DSPACE de la Universidad de Cuenca.

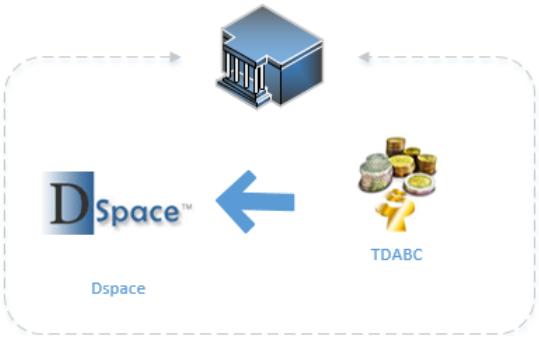


Figura 3.24: Relación de aplicaciones con el módulo DSpace

3.3.2.10. Fuente de información: Encuestas

Descripción

Este proceso se centra en evaluar el servicio prestado en el centro de documentación, el cual relaciona directamente al bibliotecario y la calificación asignada al mismo por los usuarios.

Características

El sistema de encuestas implementado y puesto en producción en el centro de documentación analizado mide la satisfacción del usuario por los servicios recibidos. Un bibliotecario está asignado en cada piso y este activa el sistema de encuestas en el ordenador asignado con su nombre de usuario, todo tipo de calificación sean estos relacionados a la calidad de servicio prestado por el bibliotecario o simplemente una evaluación a la infraestructura del centro de documentación afecta directamente al bibliotecario ya que el está a cargo ese momento y la calificación almacenada se asigna al bibliotecario. El sistema de encuestas no categoriza que a qué tipo de tarea o actividad se está evaluando pero permite tener un enfoque de la satisfacción de los usuarios sobre los servicios recibidos.

Repositorio de datos.- Base de Datos (MySQL).

Diagrama de la base de datos

EL diagrama de la base de datos del sistema de encuestas se presenta en la figura 3.25

Definición de Atributos

La *tabla "encuesta"* como se presenta en la figura 3.26, almacena la fecha y hora en la que se realizó la actividad, la terminal desde la cual se efectuó la evaluación y su correspondiente calificación que es puntuada en el rango de 1 a 5.

El proceso de Encuestas en el "Centro de Documentación Juan Bautista Vázquez" es asignada a un bibliotecario en cada piso en horarios rotativos, por lo cual cada bibliotecario debe iniciar sesión en el sistema de Encuestas. La información de acceso es almacenada en la tabla "logueo" con la fecha y la terminal de acceso.

Relación con otras Aplicaciones

El proceso de Encuestas está estructurado en dos procesos: el proceso de *"Atención al Cliente"*, el proceso de *"Encuestas"* y el módulo LibQUAL+ como se presenta en la figura 3.27.

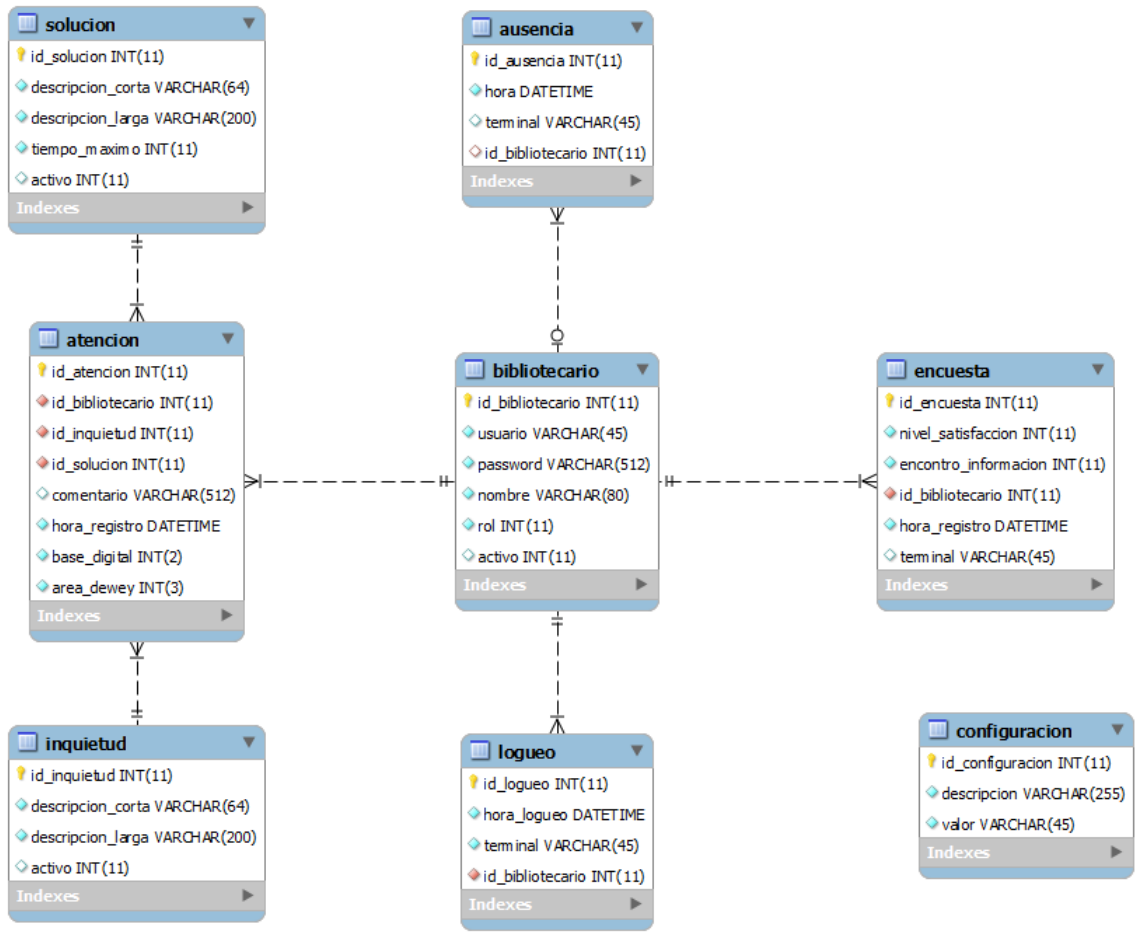


Figura 3.25: Modelo Entidad-Relacion ENCUESTA

Ubicación.- El proceso de Encuestas, forma parte de los procesos internos del centro de documentación.

Responsable.- Ing. Andrés de los Reyes
Sistema Operativo.- Centos

Acceso a datos

Intranet y mediante el sistema WEB de Encuestas de la Universidad de Cuenca.

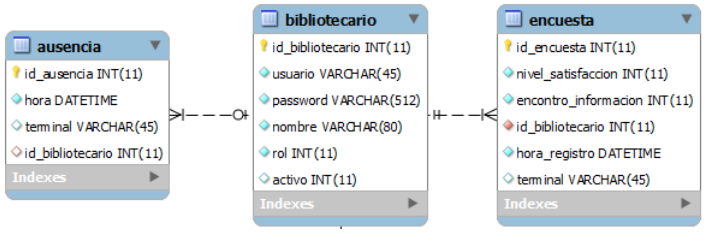


Figura 3.26: Diagrama Relacional: tabla encuesta

Figura 3.27: Relación de aplicaciones con el módulo de Encuestas

3.3.2.11. Fuente de información: Evaluación LibQuaL

Descripción LibQual+ es una metodología que permite medir la calidad de los servicios prestados en un centro de documentación o biblioteca. Esta metodología consiste en realizar encuestas a los usuarios que tienes relación directa con el centro de documentación, las cuales evalúan los servicios proporcionados.

En la actualidad la aplicación de esta metodología para la evaluación de servicios es muy común en los centros documentales, en este proyecto de tesis se incluye este tipo de evaluación a pesar de que no esté implementado. Razón por la cual, no se dispone de un historial o registros de evaluación y se considerara datos de muestras disponibles en hojas de Excel.

Para la evaluación se dispone de tres puntajes:

- **Nivel de servicio mínimo (VM):** es el valor por debajo del cual el usuario no tolera un puntaje al ítem que se evalúa.
- **Nivel observado (VO):** el valor que observa o percibe respecto al ítem que

se evalúa.

- **Nivel de servicio deseado (VD):** es el valor máximo que el usuario identifica como una evaluación deseada sobre el ítem evaluado.

LibQual+ evalua 22 preguntas clasificadas en tres categorías:

VALOR AFECTIVO DEL SERVICIO

1. AF 1 El personal le inspira confianza.
2. AF 2 El personal le ofrece atención personalizada.
3. AF 3 El personal es siempre amable.
4. AF 4 El personal muestra buena disposición para responder a las preguntas planteadas.
5. AF 5 El personal tiene conocimiento y es capaz de responder a las preguntas que se le formulan.
6. AF 6 El personal es atento con las necesidades del usuario.
7. AF 7 El personal comprende las necesidades de sus usuarios.
8. AF 8 El personal manifiesta voluntad de ayudar a los usuarios.
9. AF 9 El personal muestra fiabilidad en el tratamiento de los problemas del servicio manifestadas por los usuarios.

LA BIBLIOTECA COMO ESPACIO

1. ES 1 El espacio de la biblioteca ayuda al estudio y al aprendizaje.
2. ES 2 El espacio de la biblioteca es tranquilo para el trabajo individual.
3. ES 3 El espacio de la biblioteca es un lugar confortable y acogedor.
4. ES 4 El espacio de la biblioteca es un lugar para el estudio, el aprendizaje o la investigación.
5. ES 5 Existen espacios colectivos para aprendizaje y estudio en grupo.

CONTROL DE LA INFORMACIÓN

1. CI 1 El acceso a los recursos electrónicos es factible desde mi casa o despacho.
2. CI 2 El sitio Web de la biblioteca permite encontrar información por uno mismo.
3. CI 3 Los materiales impresos de la biblioteca cubren las necesidades de información que tengo.
4. CI 4 Los recursos digitales cubren las necesidades de información que tengo.
5. CI 5 El equipamiento es moderno y me permite un acceso fácil a la información que necesito.
6. CI 6 Los instrumentos para la recuperación de información son fáciles de usar y me permiten encontrar por mí mismo lo que busco.
7. CI 7 Puedo acceder fácilmente a la información para usarla y procesarla en mis tareas.
8. CI 8 Las revistas en versión electrónica y/o impresa cubren mis necesidades de información.

3.3.2.12. Fuente de información: Atención al Cliente

Descripción

Este proceso se inicia en el momento que el usuario del centro de documentación tiene alguna inquietud como por ejemplo al momento de realizar una búsqueda de material bibliográfico o por necesidad de algún tipo de información que necesite ayuda por parte de un bibliotecario.

Repositorio de Datos.- Base de Datos (MySQL).

Diagrama de la base de datos

Debido a que *atención al cliente* es parte del proceso de Encuestas, posee el mismo diagrama de base de datos. (Ver figura 3.25)

Definición de Atributos

De momento el "Centro de Documentación Juan Bautista Vázquez" como se observa en la figura 3.28 maneja 20 tipos de inquietudes posibles, presentado en la *tabla "inquietud"*, las cuales están manejados por un estado (activo 1 e inactivo 0).

*	id_inquietud	descripcion_corta	descripcion_larga	activo
1	27	a) No encuentra material en estantería	a) No encuentra mate...	1
2	28	b) No encuentra información en el catálogo	b) No encuentra infor...	1
3	29	c) No puede acceder a internet wifi	c) No puede acceder a...	1
4	30	d) No puede leer un archivo digital (tesis, etc.)	d) No puede leer un ar...	1
5	31	e) Solicitud de instrucción sobre el uso del catálogo	e) Solicitud de instruc...	1
6	32	f) Solicitud de instrucción sobre el uso de las bases digitales	f) Solicitud de instruc...	1
7	33	g) Solicitud de ayuda en el uso de los recursos tecnológicos del	g) Solicitud de ayuda ...	1
8	34	h) Requerimiento de información especializada	h) Requerimiento de i...	1
9	35	i) Solicitud Información para presentación de tesis	i) Solicitud Informació...	1
10	36	j) Solicitud de revisión de material bibliográfico prestado	j) Solicitud de revisión...	1
11	37	k) Solicitud de visitas guiadas	k) Solicitud de visitas ...	1
12	38	l) Solicitud de ayuda e instalación de equipos	l) Solicitud de ayuda e...	1
13	39	m) Solicitud de utilización de la mediateca	m) Solicitud de utilizac...	1
14	40	n) Solicitud de reportes de material especializado de acuerdo a	n) Solicitud de reporte...	1
15	41	o) Solicitud de asesoramiento para la asignación de palabras cla	o) Solicitud de asesor...	1
16	42	p) Solicita información sobre 60 horas	p) Solicita información...	1
17	43	q) Solicita información general	q) Solicita información...	1
18	44	r) Solicita ayuda para ubicación de objetos perdidos/olvidados	r) Solicita ayuda para ...	1
19	45	s) Otro (campo de ingreso manual)	s) Otro (campo de ingr...	1
20	46	t)No existe el material bibliográfico en la biblioteca	t)No existe el material...	1

Figura 3.28: Inquietudes administradas en el centro de documentación

Del mismo modo se tiene definido 25 tipos de soluciones posibles representadas en la *tabla "solucion"*, la cual consta de un tiempo máximo en resolver la solución del problema presentado y un estado (activo 1 e inactivo 0) como se presenta el la figura 3.29.

En la *tabla "Bibliotecario"* se almacena todos los nombres de los bibliotecarios y los nombres de usuario de los mismos.

Una vez dada la solución al problema presentado, este se almacena en la *tabla "atencion"* con la fecha y hora en la que se dio la solución.

En la figura 3.30 se observa las relaciones del sistema de Encuestas.

*	id_solucion	descripcion_corta	descripcion_larga	tiempo_maximo	activo
1	11	a) Orientación y ayuda al usuario a ubicar el material bibliográ	a) Orientación y ayuda al usuario a ubicar el materi...	5	1
2	12	b) Se indica cuales son las mejores fuentes de información o ins	b) Se indica cuales son las mejores fuentes de infor...	5	1
3	13	b/e) Instrucción al usuario en el manejo adecuado del catálogo	b/e) Instrucción al usuario en el manejo adecuado ...	5	1
4	14	c) Se traslada inquietud a la unidad tecnológica	c) Se traslada inquietud a la unidad tecnológica	5	1
5	15	d) Comunicación a unidad de procesos o unidad tecnológica para r	d) Comunicación a unidad de procesos o unidad tec...	5	1
6	16	f) Instrucción en el uso de las Bases de Datos Digitales	f) Instrucción en el uso de las Bases de Datos Digit...	5	1
7	17	g) Información general sobre uso de los recursos tecnológicos de	g) Información general sobre uso de los recursos t...	5	1
8	18	g) Capacitación en sala de uso múltiple	g) Capacitación en sala de uso múltiple	5	1
9	19	h) Búsqueda y envío de información de las Bases de datos digital	h) Búsqueda y envío de información de las Bases d...	5	1
10	20	i) Se proporciona la información existente en la página web (fo	i) Se proporciona la información existente en la pá...	5	1
11	21	j) Revisión en el sistema de préstamos	j) Revisión en el sistema de préstamos	5	1
12	22	k) Visita guiada	k) Visita guiada	5	1
13	23	l) Se facilita la utilización e instalación de equipos y materi	l) Se facilita la utilización e instalación de equipos ...	5	1
14	24	m) Se autoriza el uso de la mediateca	m) Se autoriza el uso de la mediateca	5	1
15	25	n) Entrega vía mail de reportes de material especializado de acu	n) Entrega vía mail de reportes de material especi...	5	1
16	26	n) Se pasa la inquietud a la dirección	n) Se pasa la inquietud a la dirección	5	1
17	27	n) Se pasa la inquietud a la unidad tecnológica	n) Se pasa la inquietud a la unidad tecnológica	5	1
18	28	n) Se orienta en el uso del catálogo para la elaboración de repo	n) Se orienta en el uso del catálogo para la elabora...	5	1
19	29	o) Se instruye en el manejo del Tesauro Especializado para la as	o) Se instruye en el manejo del Tesauro Especializa...	5	1
20	30	p) Se proporciona información de la página web de la biblioteca	p) Se proporciona información de la página web de ...	5	1
21	31	q) Se brinda información general	q) Se brinda información general	5	1
22	32	r) Ayuda para encontrar objetos perdidos u olvidados	r) Ayuda para encontrar objetos perdidos u olvida...	5	1
23	33	r) Se solicita a unidad tecnológica revisión en cámaras	r) Se solicita a unidad tecnológica revisión en cáma...	5	1
24	34	s) Otro (campo de ingreso manual)	s) Otro (campo de ingreso manual)	5	1
25	35	t) Área en la que no existe material bibliográfico	t) Área en la que no existe material bibliográfico	5	1

Figura 3.29: Soluciones administradas en el centro de documentación

Figura 3.30: Relación de aplicaciones con el módulo de Atención al Cliente

Relación con otras Aplicaciones

Ubicación.- El proceso de Encuestas- Atención al Cliente, forma parte de los procesos internos del centro de documentación.

Responsable.- Ing. Andrés de los Reyes

Sistema Operativo.- Centos

Acceso a Datos.- Intranet y mediante el sistema WEB de Encuestas de la Universidad de Cuenca.

3.3.2.13. Fuente de información: Socioeconomica

Descripción

Es el encargado de la administración de los datos económicos de cada estudiante matriculado en la Universidad de Cuenca, los mismos que son ingresados de forma personal por cada alumno al inicio de cada periodo académico.

Características

Está conformado por ochenta y cinco tablas aproximadamente, de las cuales para el presente proyecto se utilizará siete tablas para extraer la información más relevante de la economía de un estudiante.

Repositorio de datos.- Base de Datos (ORACLE).

Diagrama de la base de datos En la figura 3.31 se observa las relaciones del sistema Socioeconomico con los demas módulos.

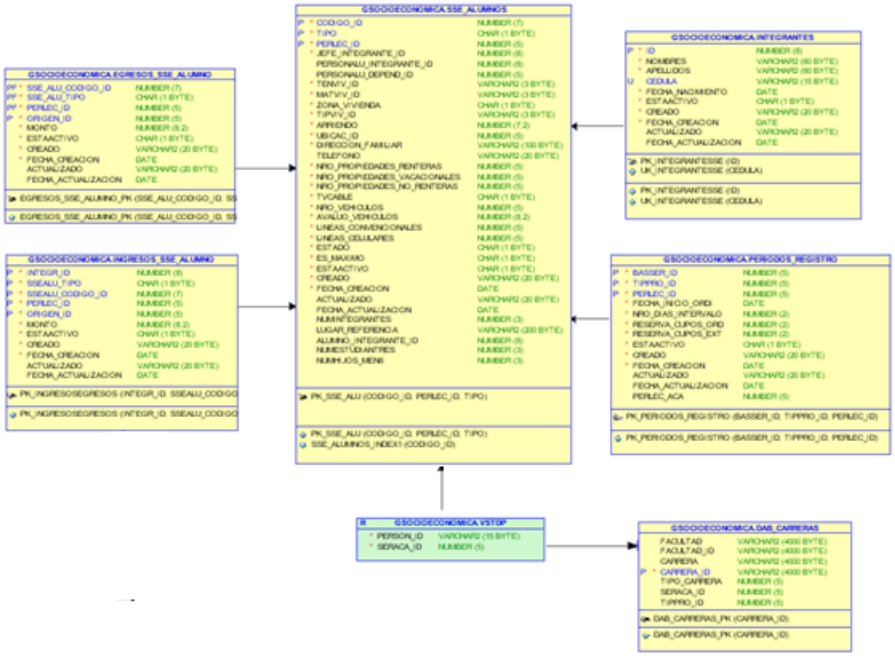


Figura 3.31: Diagrama Entidad - Relación del sistema GSocioeconomica

Definición de Atributos

El esquema mostrado previamente se divide en tres secciones para un mejor entendimiento:

1. Ingresos/Egresos Alumno

Estas tablas describen la situación económica de los miembros de la familia del estudiante, almacenando campos relevantes como: el monto de ingreso, el monto de egreso, código del estudiante (SSE-ALU-CODIGO-ID) y el periodo lectivo del mismo (PERLEC-ID) (Ver figura 3.32).

Figura 3.32: Tablas ingresos y egresos, Sistema GSocioeconomica

A partir de esta tabla se realiza el cálculo de los costos totales de ingreso y egreso de cada estudiante en cada periodo, considerando el campo monto tanto en INGRESO-SSE-ALUMNO como de EGRESO-SSE-ALUMNO, se calcula la suma de todos los ingresos relacionados con un alumno en un periodo dado para determinar el monto total de ingresos y de la misma forma para los egresos.

2. Integrantes

En la figura 3.33 se presenta la tabla integrantes que describe cada uno de las personas que conforman el grupo familiar de un estudiante. En esta tabla se almacena los datos principales de una persona como el nombre, apellido, cédula, fecha de nacimiento, etc.

3. Periodo Registro

Representada por la tabla PERIODOS-REGISTRO que almacena para cada período lectivo información relevante como la fecha de inicio del periodo y el código.

Relación con otras Aplicaciones

El sistema GSocioeconomica se relaciona directamente con el sistema *Academico* y el de *RRHH* como se observa en la figura 3.34.

Figura 3.33: Tabla Integrantes, Sistema GSocioeconomica

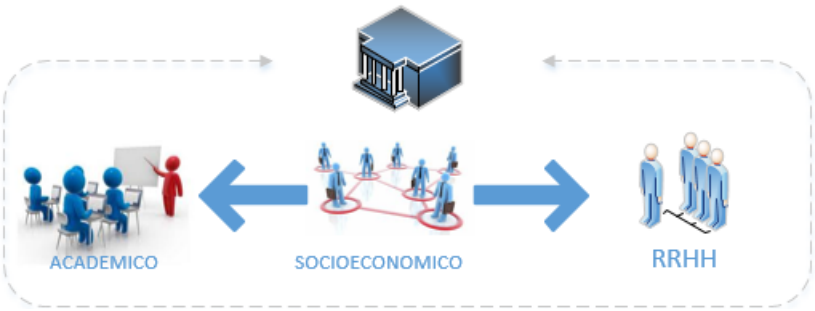


Figura 3.34: Relación de aplicaciones con el módulo Socioeconómico

Ubicación.- El proceso relacionado con el esquema GSocieconomica se encuentra administrada por el DTIC de la Universidad de Cuenca, siendo una fuente externa para el centro de documentación.

Responsable.- DTIC

Accesos a Datos.- Intranet y mediante el sistema WEB ESIUC para llenar la "Ficha Socieconómica"de la Universidad de Cuenca.

3.3.2.14. Fuente de información: Adquisición

Descripción

La parte de Inventarios de la Universidad de Cuenca está desarrollada en Oracle Database través del esquema "OLYMPOAF1" conformado por 792 tablas, de los cuales para el presente proyecto de tesis se utiliza seis tablas que son: AF-ACTIVOFIJO, AF-CLASE, AF-SECCION, AF-AREA, AF-PROVEE y DOCUACTIVOFIJO.

Características

Entre las características más relevantes están:

- Sistema administrado por el DTIC de la Universidad de Cuenca.
- El esquema OLYMPOAF1 almacena todos los bienes e inmuebles de la Universidad de Cuenca

Repositorio de datos.- Base de Datos (ORACLE).

Diagrama de Datos

El esquema de la base de datos del sistema Olympo se presenta en la figura 3.35.

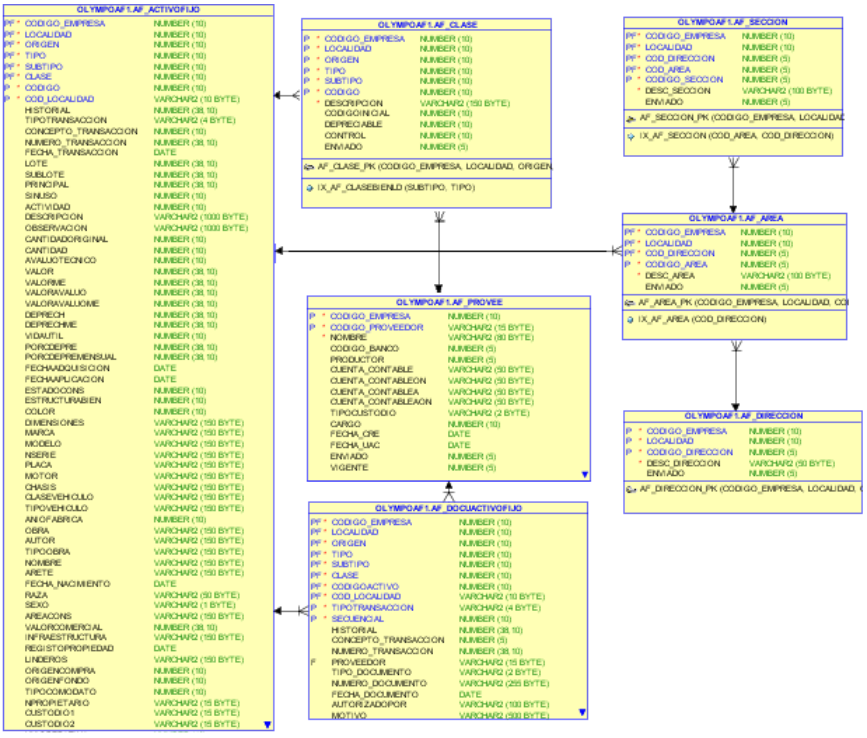


Figura 3.35: Diagrama Entidad - Relación del proceso Adquisición

Definición de Atributos

Para distinguir los diferentes tipos de material que existe, se usa la *tabla AF-TIPOS*

en la cual los campos “codigo” y “descripcion” indican el tipo de recurso (figura 3.36).

17)	11	BIENES MUEBGLES ADQUIRIDOS LIBROS	1	1	0
18)	10	BIENES MUEBLES ADQUIRIDOS	1	1	11
19)	9	LIBROS Y COLECCIONES	1	1	30
20)	8	BIENES ARTISTICOS Y CULTURALES	1	3	2
21)	7	EQUIPOS, SISTEMAS Y PAQUETES INFORMATICOS	1	1	86
22)	6	HERRAMIENTAS	1	1	1
23)	5	VEHICULOS	1	5	4
24)	4	MAQUINARIA Y EQUIPOS	1	1	66

Figura 3.36: Tablas Activo Fijo, sistema Adquisiciones

Para el desarrollo del proyecto de tesis se utiliza los del tipo *9*, que son registros de adquisiciones de material bibliográfico.
La clase del material bibliotecario adquirido se almacena en la *tabla “AF-CLASE”*, en la cual los campos “codigo” y “descripcion” ayudan en la identificación de la clase de cada registro en el esquema de inventario.
En la figura 3.37 se presentan los tipos de material bibliográfico que se dispone en la Universidad de Cuenca.

	TIPO	SUBTIPO	CODIGO	DESCRIPCION
585)	8	3	1	ESCULTURAS
586)	8	6	1	ADORNOS, ORNAMENTAS, UTENCILLOS VARIAS CULTURAS
587)	9	99	1	OTROS
588)	9	5	1	REVISTAS
589)	9	3	1	LIBROS TECNICOS PARA PROYECTOS
590)	9	4	1	MANUALES VARIOS
591)	9	1	1	COLECCIONES Y ENCICLOPEDIAS
592)	9	2	1	DICCIONARIO
593)	9	3	2	LIBROS
594)	10	18	1	SORTES SEPARADORES
595)	10	18	2	CANDADO

Figura 3.37: Detalle tabla tipo material bibliográfico

La relación “Af-SECCION”, “AF-AREA” y “AF-DIRECCION” permiten identificar que departamento dentro de la Universidad de Cuenca solicitó la compra del material bibliográfico. De esta manera la *tabla “AF-AREA”* describe el área educativa, la *tabla “AF-SECCION”* define los departamentos que conforman el área y la *tabla “AF-DIRECCION”* indica donde está ubicada el área (figura 3.38).

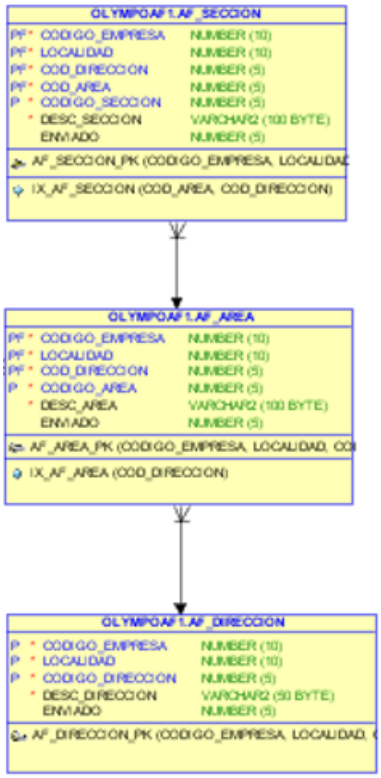


Figura 3.38: Relación tabla:Área, sistema Adquisición

"AF-PROVEE" describe al proveedor del material bibliográfico en la cual incluye información relacionada a la empresa y el representante empresarial en la Universidad de Cuenca como se presenta en la figura 3.39.

Relación con otras Aplicaciones

El sistema Olympo se relaciona indirectamente con el proceso Catalogación(figura 3.40).

Ubicación.- Se encuentra administrada por DTIC de la Universidad de Cuenca, siendo una fuente externa para el centro de documentación.

Acceso a Datos.- Intranet.

Figura 3.39: Tabla proveedor, sistema Adquisición

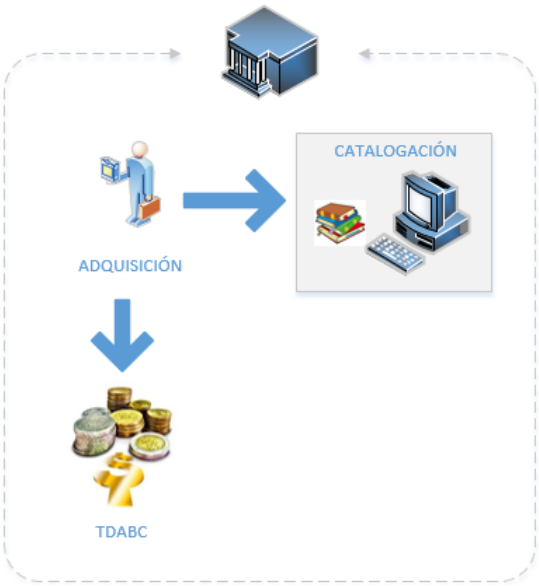


Figura 3.40: Relación de aplicaciones con el módulo de Adquisición

3.3.2.15. Fuente de información: Académico

Descripción

El sistema web del Registro de Matriculas de la Universidad de Cuenca está desarrollada en Oracle Database, formado por 85 tablas, de los cuales para el proyecto de tesis se utiliza cuatro tablas que son: SGA-MATRICULAS, SGA-PERIODOS-LECTIVOS, MATRICULAS-ASIGNATURAS y ASIGNATURAS.

Características

La características más relevantes es que el esquema ACADEMICO almacena todas las matriculas efectuadas por cada estudiante de la Universidad de Cuenca al inicio de cada periodo académico.

Repositorio de Datos.- Base de Datos (ORACLE).

Diagrama de la base de datos En la figura 3.41 se presenta el esquema de la base de datos del módulo academico.

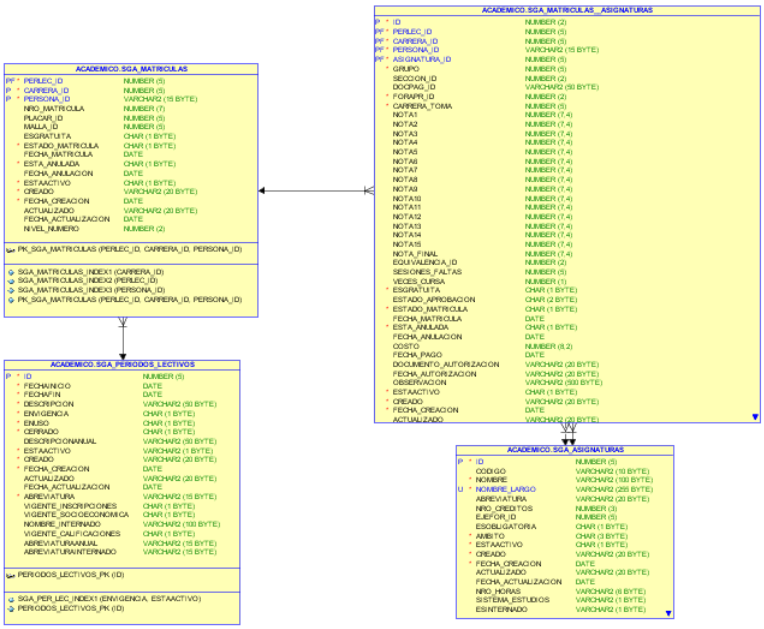


Figura 3.41: Diagrama Entidad - Relación del proceso ACADEMICO

Definición de Atributos

La *tabla "SGA-ASIGNATURA"* almacena las asignaturas que existen en la Universidad de Cuenca, conjuntamente con la facultad a la que pertenece y el tipo de sistemas de estudios que es (crédito, distancia, etc.).

La *tabla "SGA-PERIODOS-LECTIVOS"* se encarga de almacenar los periodos de cada ciclo lectivo, con su respectiva fecha y código.

La *tabla "SGA-MATRICULAS"* permite identificar que estudiante se matriculo en qué periodo lectivo y a qué carrera pertenece definida por los campos "persona-id", "perlec-id" y "carrera-id" respectivamente.

La *tabla "SGA-MATRICULAS-ASIGNATURAS"* integra las claves de matrícula y

asignatura, permitiendo identificar que estudiante se matriculo en que materias en un periodo académico.

Relación con otras Aplicaciones

El sistema ACADEMICO se relaciona directamente con el proceso GSOCIOECONOMICA y RRHH (figura 3.42).

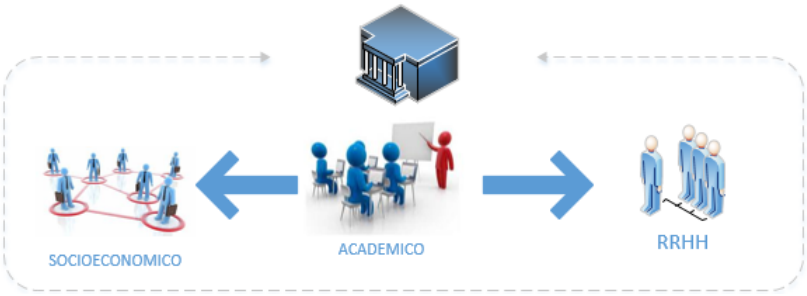


Figura 3.42: Relación de aplicaciones con el módulo Académico

Ubicación.- El proceso relacionado con el Sistema ACADEMICO se encuentra administrada por el DTIC de la Universidad de Cuenca, siendo una fuente externa para el centro documental.

Acceso a Datos.- Intranet.

3.3.3. Modelo Conceptual del Data Warehouse

Después de analizar los procesos que se llevan a cabo en el centro de documentación y las fuentes de datos existentes, se procede a diseñar el modelo lógico para el Data Warehouse basándose en el análisis holístico que consta de cuatro cuadrantes (Ver figura 3.43).

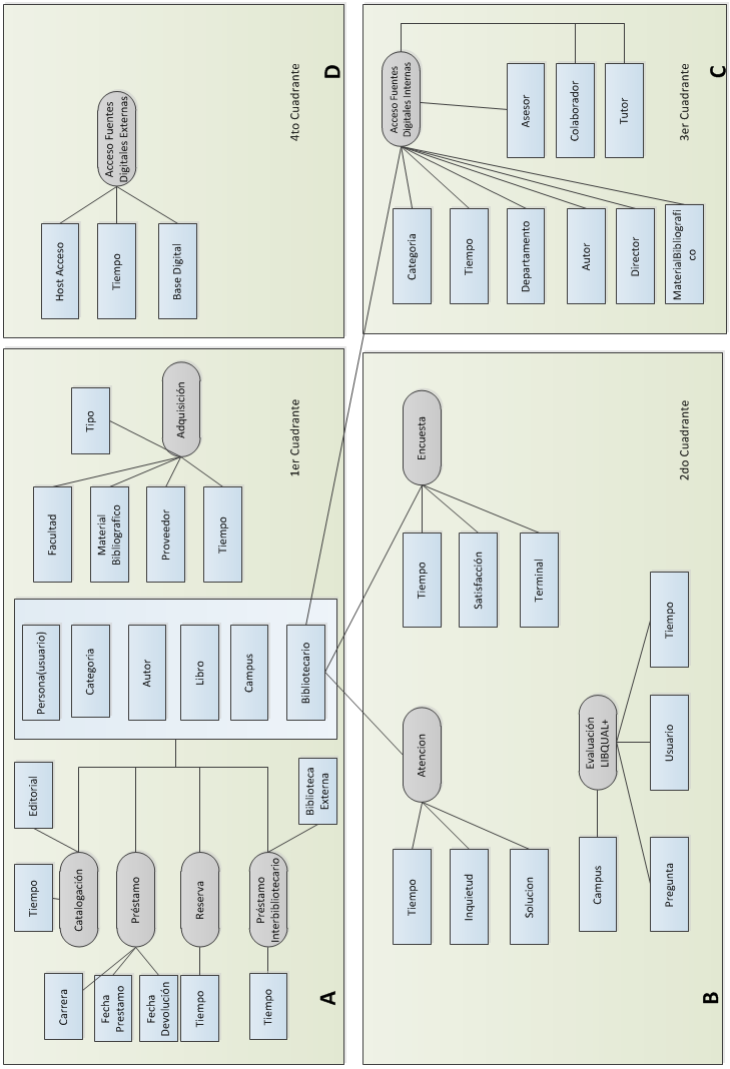


Figura 3.43: Modelo conceptual del Data Warehouse

En base a la sección *3.3.1. Definición de los requerimientos del negocio* se desarrolla el modelo conceptual que se explica a continuación.

- **Cuadrante I.-** En este cuadrante se tiene un total de 5 modelos conceptuales que representan a cada uno de los procesos relacionados al análisis de costos (Ver figura 3.43A). Los procesos de catalogación, reservas de libros, préstamos, préstamos interbibliotecarios y adquisiciones tienen en común perspectivas de costos, sean estos económicos o materiales que se consideró para agruparlo en este cuadrante.

- **Cuadrante II.-** En este cuadrante se tiene un total de 3 modelos conceptuales que son procesos relacionados a la evaluación de los servicios bibliotecarios (Ver figura 3.43B), entre los cuales existen: atención al cliente, evaluación y encuestas. Estos procesos están formados por perspectivas relacionadas a la evaluación de servicios.

- **Cuadrante III.-** En este cuadrante se encuentra un solo modelo conceptual que es *Acceso a fuentes digitales internas* (Ver figura 3.43C), el proceso DSpace está relacionado directamente en este cuadrante.

- **Cuadrante IV.-** Este cuadrante esta conformado por el modelo *Acceso a fuentes digitales externas* (Ver figura 3.43D).

3.4. Aplicación de la Metodología Hefesto para el proceso de Préstamos

Como se analizó anteriormente, este proyecto de tesis involucra varios procesos internos como externos, del cual el más complejo es el de préstamos debido a que integra los procesos de catalogación y reservas. Por esta razón, con el objetivo de implementar la metodología Hefesto en la creación del Data Warehouse se considera el proceso de préstamos. Los demás procesos involucrados en el Data Warehouse se presentan en los anexos.

3.4.1. Análisis de requerimientos

El primer paso de la metodología Hefesto comienza con la recolección de los requerimientos del personal involucrado en el desarrollo del Data Warehouse, que puede ser captada a través de variadas y diferentes técnicas como: entrevistas, cuestionarios, observaciones, etc. Estos requerimientos recolectados deben de ser captadas en forma de pregunta que involucren las necesidades del negocio desde diferentes puntos de vista de los involucrados en el proyecto.

Basado en el análisis holístico del proceso préstamos, este está incluido en el primer cuadrante junto con los procesos: adquisiciones, reservas, préstamos interbibliotecarios y catalogación.

3.4.1.1. Identicar preguntas

Luego de reuniones constantes con los Ingenieros encargados de los sistemas relacionados al proceso de Préstamos, con la Directora del centro de documentación y la mentora del proyecto del Data Warehouse y luego de entrevistas con los expertos del tema relacionados al proceso de catalogación, se analizó los requerimientos que estén soportados por una fuente.

A continuación se procedió a identificar las preguntas de negocio, las cuales son:

1. ¿Cuál es el número de préstamos de un libro de un determinado autor, de una determinada categoría a un tipo de usuario en la biblioteca en una unidad de tiempo?

2. ¿Cuál es el número de préstamos realizados por un determinado funcionario de la biblioteca a un usuario determinado en un determinado campus en una unidad de tiempo?

3. ¿Cuál es el número de devoluciones realizadas por un determinado funcionario de la biblioteca en una unidad de tiempo?

4. Total de multas y su valor, realizadas en un período de tiempo determinado

5. ¿Cuál es el costo operacional de los préstamos realizados en una unidad de tiempo?

3.4.1.2. Identicar indicadores y perspectivas

Luego de que se han establecido las preguntas se procede a identificar los indicadores (valores numéricos) y las perspectivas (objetivos) relacionadas al proceso Préstamos.

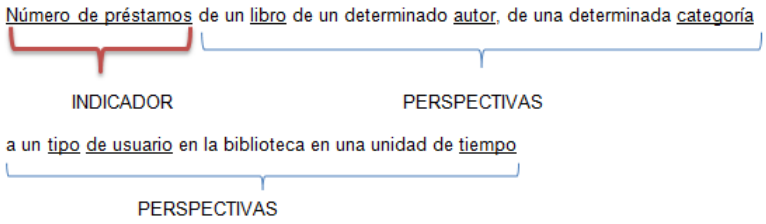


Figura 3.44: Indicadores y Perspectivas de la pregunta 1

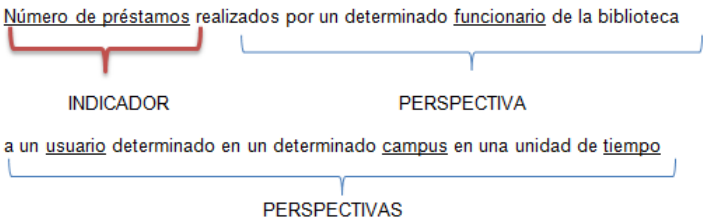


Figura 3.45: Indicadores y Perspectivas de la pregunta 2

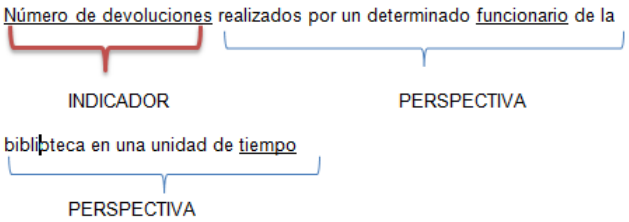


Figura 3.46: Indicadores y Perspectivas de la pregunta 3

Los indicadores que han sido identificados son:

- Número de préstamos
- Número de devoluciones

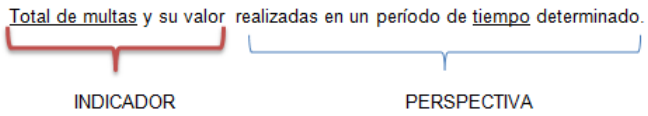


Figura 3.47: Indicadores y Perspectivas de la pregunta 4

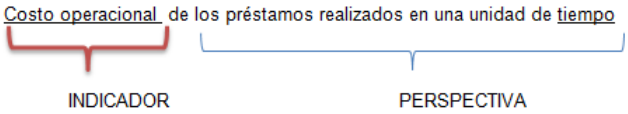


Figura 3.48: Indicadores y Perspectivas de la pregunta 5

- Total de multas
- Costo operacional

Las perspectivas que deben considerarse son:

- Libro
- Autor
- Categoría
- Usuario
- Fecha préstamo
- Bibliotecario
- Campus
- Fecha devolución

3.4.1.3. Modelo conceptual

Considerando los indicadores y perspectivas analizadas en el paso anterior se plantea el modelo conceptual, que permite observar con mayor facilidad el alcance del proyecto en base al proceso de Préstamos. El modelo conceptual del proceso préstamos se presenta en la figura 3.49

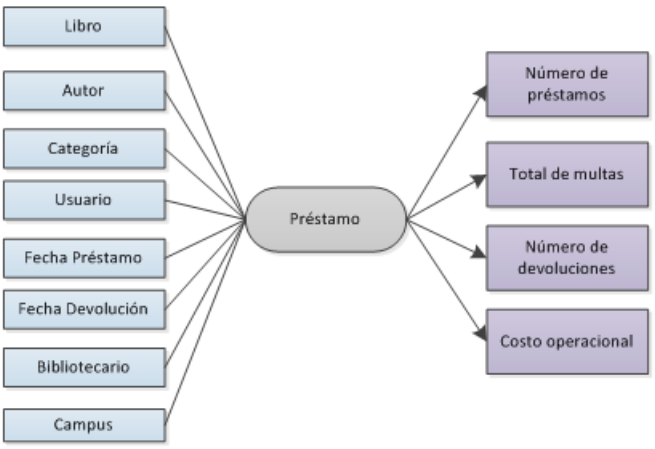


Figura 3.49: Modelo conceptual del proceso préstamos

3.4.2. Análisis de los OLTP

El análisis de las fuentes de información se describió de manera exhaustiva en la sección ***3.5 Análisis de las fuentes de datos***. En esta sección se continúa con la aplicación de la metodología Hefesto.

3.4.2.1. Conformar indicadores

En este paso se explica las operaciones para calcular los indicadores de cada tabla de hecho que forman parte del proceso de Préstamos. Los indicadores se calcularán de la siguiente manera:

Número de préstamos

- Hechos: cantidad
- Función de conteo: Count
- Aclaración: el indicador *número de prestamos* representa el conteo de la cantidad de préstamos que se han realizado,

Número de devoluciones

- Hechos: estado

- Función de conteo: Distinct count
- Aclaración: el indicador *número de devoluciones* representa el conteo de la cantidad de devoluciones que se han realizado, considerar que la variable estado puede tomar el valor 0 - Prestado y 1 - Devuelto en cada registro que se almacena en préstamos

Total de multas

- Hechos: multa
- Función de conteo: Sum
- Aclaración: el indicador *total de multas* representa la suma de los valores de multas asignados por cualquier motivo a un préstamos de un libro.

Costo operacional

- Hechos: costo-operacional
- Función de conteo: Sum
- Aclaración: el indicador *costo-operacional* representa la suma de los costos de las actividades relacionadas al servicio de préstamos.

3.4.2.2. Establecer correspondencias

Se examinó los OLTP involucrados en el proceso de préstamos para poder identificar y establecer las correspondencias de los mismos con el modelo conceptual. La correspondencia de indicadores se presenta en la figura 3.50 y la correspondencia de perspectivas se presenta en la figura 3.51.

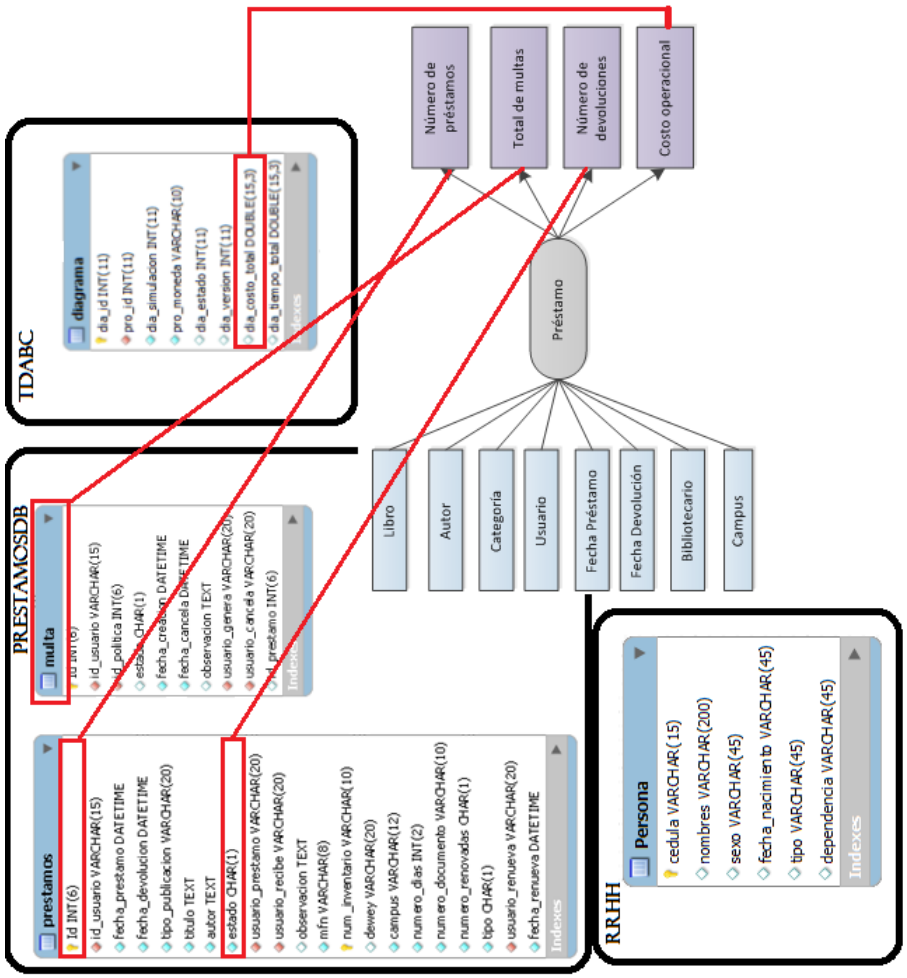


Figura 3.50: Correspondencia indicadores

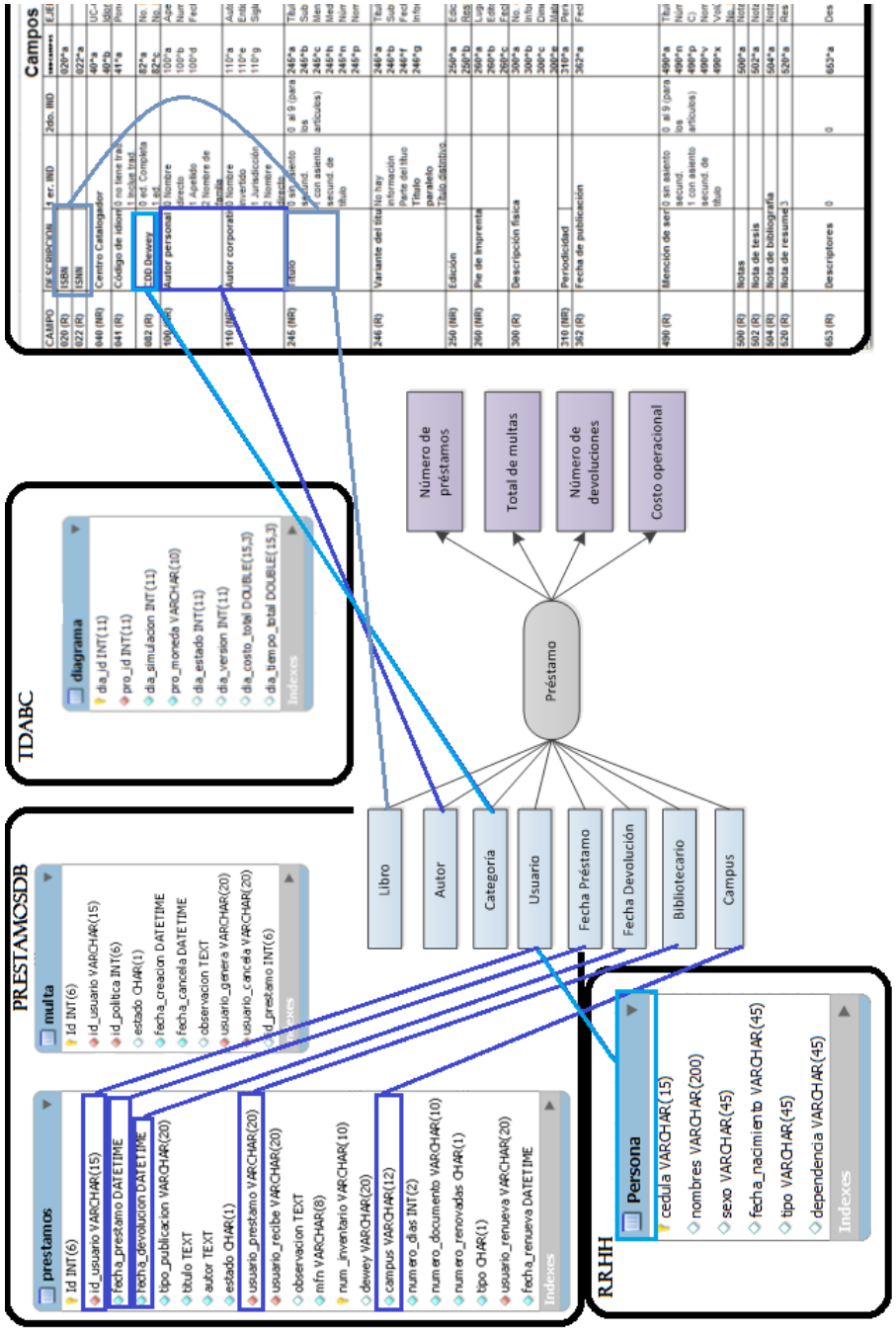


Figura 3.51: Correspondencia perspectivas

3.4.2.3. Modelo conceptual ampliado

En este paso se presenta los resultados obtenidos en los pasos anteriores, se ampliará el modelo conceptual, colocando bajo cada perspectiva los campos seleccionados.

Perspectiva Usuario: Se refiere al usuario que solicita un libro.

- Cedula: Número de cédula, ruc o pasaporte del usuario
- Nombres: Nombre completo del usuario
- Sexo: Cuyos valores serán F o M
- Fecha Nacimiento: Fecha de nacimiento de la persona
- Tipo: tipo de persona puede ser: estudiante, profesor, trabajador o empleado
- Dependencia: Facultad o dependencia a la que está relacionada la persona

Perspectiva Bibliotecario: Contiene información del bibliotecario que realizó el préstamo.

- Nombres: Nombre completo del bibliotecario
- Username: Nombre de usuario con el que accede al sistema

Perspectiva Autor: Contiene información del autor o los autores de un libro, los autores pueden ser de tipo personal o corporativo.

- Nombres: Nombre del autor
- Tipo: Personal o Corporativo

Perspectiva Campus: Contiene información del campus en el que fue realizado el préstamo.

- Nombre: Nombre del campus

Perspectiva Libro: Contiene información del libro o material bibliográfico.

- MFN: Identificador único para el libro
- Título: Título del libro
- ISBN: ISBN del libro en el caso de poseerlo
- ISSN: ISSN de la revista si dispone al tratarse de este tipo de material bibliográfico
- Idioma: Idioma en el que está escrito el libro
- Edición: Número de edición
- Cod-inventario: Código que relaciona al libro con la base de datos de adquisiciones
- Ubicación: Indica la ubicación del libro
- Lugar de publicación: Lugar donde fue publicado el libro
- Tipo: indica si se trata de un material bibliográfico monográfico o seriado

Perspectiva Categoría: Contiene información de la categoría a la que pertenece un determino libro

- Categoría: Nombre de la categoría
- Subcategoría: Nombre de la subcategoría
- Dewey: Código dewey del libro

Perspectiva Fecha Préstamo: Contiene información de la fecha de préstamo

- Año: Año del préstamo
- Mes: Mes en el que se realizó el préstamo
- Día: Día en el que se realizó el préstamo

Perspectiva Fecha Devolución: Contiene información de la fecha de devolución

- Año: Año de la devolución

- Mes: Mes en el que se realizó la devolución
- Día: Día en el que se realizó la devolución

El modelo conceptual ampliado del proceso de prestamos se presenta el la figura 3.52

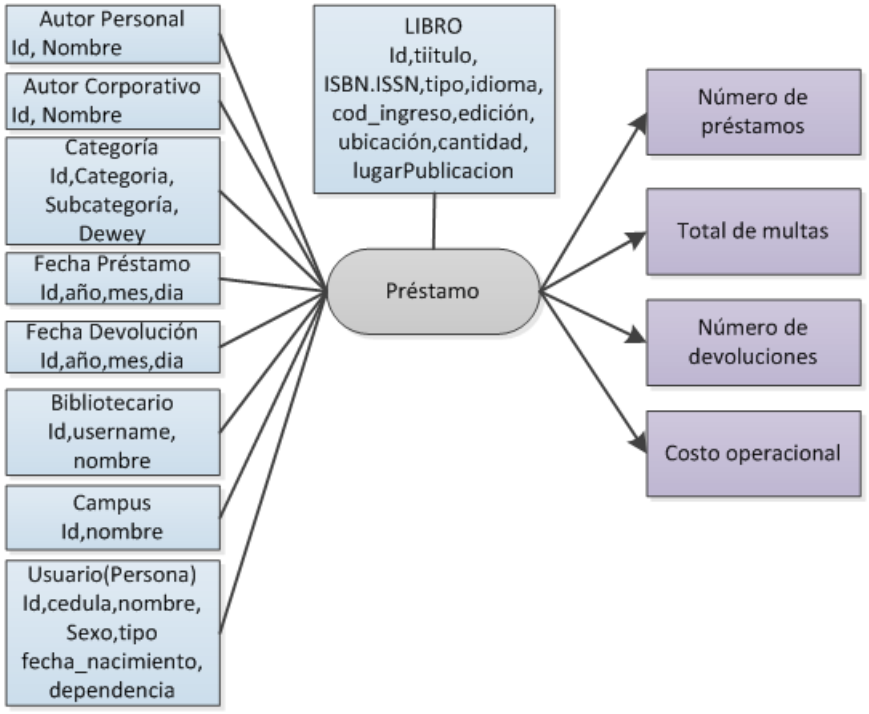


Figura 3.52: Modelo Lógico para préstamos

3.4.3. Modelo lógico del proceso préstamos

Tomando en cuenta el modelo conceptual ampliado del proceso de Prestamos, se realizó el mapeo de los OLTP al modelo lógico. El modelo lógico del proceso de prestamos se presenta el la figura 3.53

3.4.3.1. Tipo de Modelo Lógico del Data Warehouse

Se selecciona el tipo de esquema que más se adapte a los requerimientos a utilizar para la estructura del Data Warehouse. El esquema seleccionado para el desarrollo

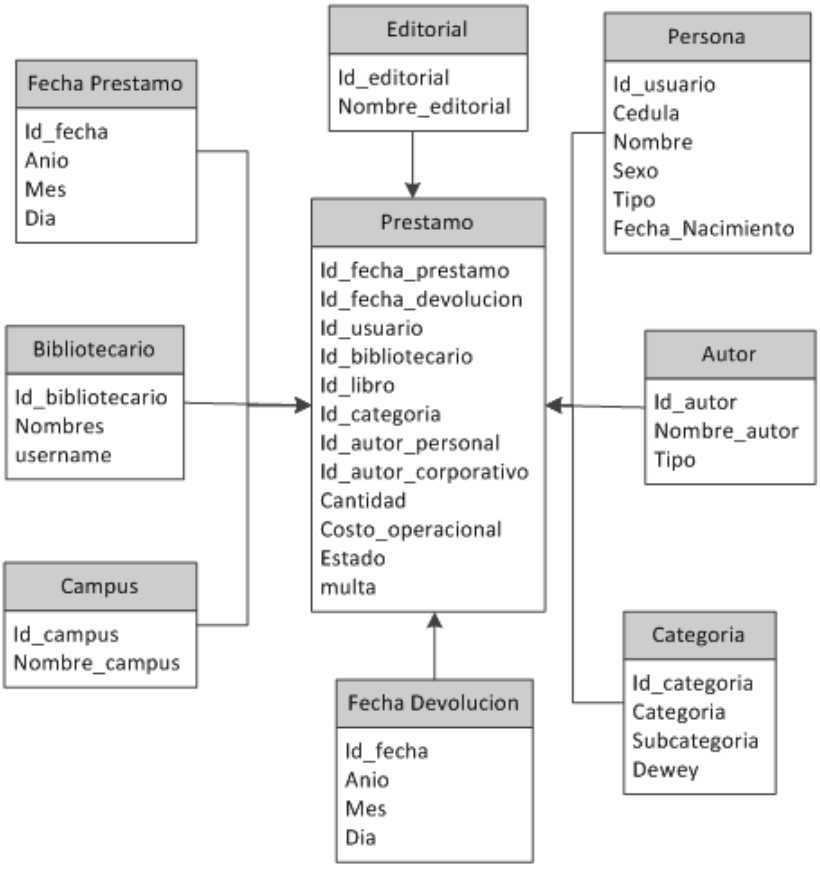


Figura 3.53: Modelo lógico del proceso préstamos

del Data Warehouse es el de estrella, debido a compatibilidad con la herramienta seleccionada y a la simplicidad de diseño.

3.4.3.2. Tablas de dimensiones

En este paso de diseñan las tablas de dimensiones que forman parte del proceso préstamos dentro del Data Warehouse, para la cual se debe de tomar en cuenta cual será el nombre de la tabla, añadir el campo que será la clave primaria y los respectivos nombres a los campos.

Perspectiva Persona: La nueva tabla de dimensión persona hace referencia a los usuarios que solicitan préstamos en el centro de documentación con los siguientes

atributos:

- Se le agregará una clave principal *id-usuario*
- Tendrá el número de *cédula*, ruc o pasaporte del usuario en el campo *cedula*
- Tendrá el campo *nombre* que es el nombre completo de la persona
- Se mantiene el atributo sexo que podrá tener los valore M o F
- Se incluye el campo fecha de nacimiento *fecha-nacimiento*
- Se mantiene el campo *tipo* que indica el tipo de persona, puede tomar uno de los siguientes valores: estudiante, profesor, trabajador o empleado
- Se modificará el campo dependencia por una dimensión llamada Carrera o dependencia.

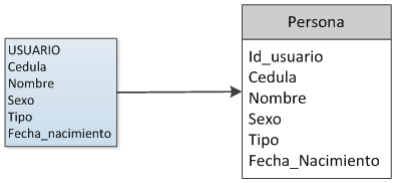


Figura 3.54: Tabla Dimensión Persona

Perspectiva Bibliotecario: Dimensión denominada dim-bibliotecario

- Se agrega el campo *id-bibliotecario* que es un identificador único para el bibliotecario
- Se mantiene el campo *nombres*
- Se modifica el nombre del campo usuario por *username*

Perspectiva Autor:

- Se agrega el campo *Id-autor*: Identificador único para el autor
- Se mantiene el campo *nombres*: Nombre del autor

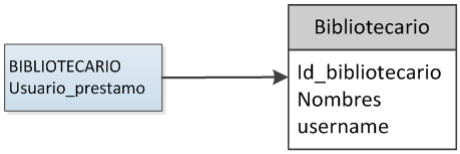


Figura 3.55: Tabla Dimensión Bibliotecario

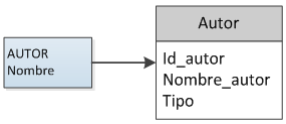


Figura 3.56: Tabla Dimensión Autor

- Se crea otra dimensión denominada tipo$_a$*utortipoparadistinguiralautor*.

  Perspectiva Campus:

  - Se agrega el campo *id-campus* que es un identificador único para el campus
  - Se modifica el nombre del campo *campus* a nombre-campus

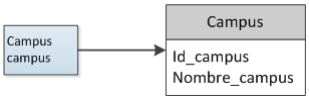


Figura 3.57: Tabla Dimensión Campus

  Perspectiva Libro:

  - Se agrega el campo *id-libro* que es un identificador único para el libro
  - Se mantiene el campo *título*
  - Se incluye el campo *isbn*
  - Tendrá el campo *issn* para el caso de las revistas
  - Tendrá el campo *idioma*
  - Se mantiene el dato *edicion*
  - Se mantiene el campo *cod-inventario*

- Se mantiene el campo *ubicacion*
- Se incluye el campo *lugar-publicación*
- Se incluye el campo *anio-publicación*
- Se incluye el campo *cantidad*
- Tendrá el campo *estado-registro*
- Se mantiene el campo *tipo-registro*
- Se incluye el campo *nivel-bibliográfico*
- Se incluye el campo *nivel-codificacion*

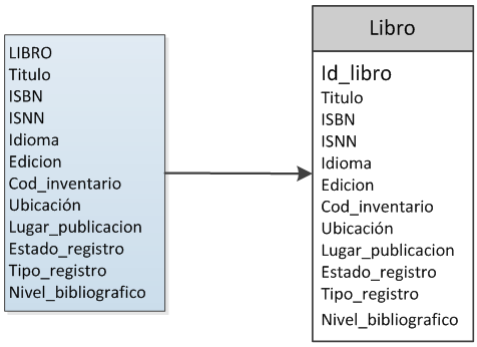


Figura 3.58: Tabla Dimensión Libro

Perspectiva Categoría:

- Se agrega el campo *id-categoría* que es un identificador único para la categoría
- Se conserva el campo *categoria*
- Se conserva el campo *subcategoría*
- Se mantiene el campo *dewey*

Perspectiva Fecha Préstamo:

- Se agrega el campo *id-fecha-prestamo* que es un identificador único para la fecha en la que fue prestado un libro.

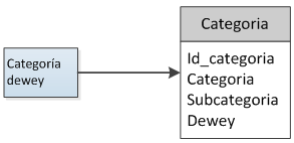


Figura 3.59: Tabla Dimensión Categoria

- El nombre de los campos no serán modificados.

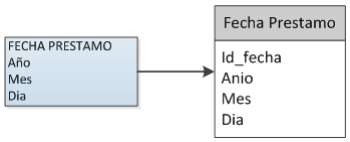


Figura 3.60: Tabla Dimensión Fecha Préstamo

Perspectiva Fecha Devolución:

- Se agrega el campo *id-fecha-devolución* que es un identificador único para la fecha en la que fue devuelto un libro
- El nombre de los campos no serán modificados

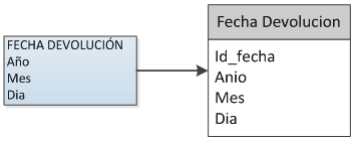


Figura 3.61: Tabla Dimensión Fecha Devolución

3.4.3.3. Tablas de hechos

En este paso se definen las tablas de hechos, que son las que contendrán los hechos a través de los cuales se construirán los indicadores de estudio. Para la cual, se debe de tomar en cuenta cual será el nombre de la tabla, añadir el campo que será la clave primaria y los respectivos nombres a los campos.

- La tabla de hechos tendrá el nombre PRESTAMO.

- Su clave principal será la combinación de las claves principales de las tablas de dimensiones antes definidas: id-usuario, id-libro, id-autor, id-categoria, id-campus, id-bibliotecario, id-fecha-devolucion e id-fecha-prestamo (Ver figura 3.62).
- Se crearón hechos, que se corresponden con los indicadores identificados anteriormente

Figura 3.62: Tabla de Hecho Prestamo

3.4.3.4. Uniones

Se realizó las uniones correspondientes entre sus tablas de dimensiones y sus tablas de hechos, en donde el resultado se ve en la figura 3.53.

3.5. Conclusión

Después del análisis de los requerimientos se diseñó el modelo del Data Warehouse, en este capítulo se aplicó la metodología de Hefesto para la creación del modelo lógico del proceso préstamos, los demás modelos se puede observar en Anexos.

Este procedimiento de diseño es independiente de la herramienta a utilizarse en la creación del Data Warehouse, la misma que se analiza en el siguiente capítulo.

Capítulo 4

Desarrollo e Implementación de un Data Warehouse para el Centro de Documentación "Juan Bautista Vázquez"

4.1. Introducción

Una vez diseñado el modelo multidimensional el siguiente paso es poblar los modelos con datos para lo cual, primero se debe instalar y configurar la herramienta de Pentaho Data Integration que permite realizar el proceso ETL (Extracción, Transformación, Carga). Estas acciones ayudan a poblar los datos en el Data Warehouse para que estos datos sean de calidad.

Se aplica la metodología Hefesto sobre el proceso de préstamos del material bibliográfico para realizar los procesos ETL como una muestra de todo el trabajo realizado para el Centro de Documentación "Juan Bautista Vázquez". Los ETL de los demás procesos del centro de documentación son expuestos en la sección de anexos.

4.2. Selección de las herramientas tecnológicas para el sistema de ayuda a la toma de decisiones

En el presente proyecto se consideran las herramientas open source que permitan realizar todo el proceso del Data Warehouse, siguiendo las políticas gubernamentales que determinan la utilización de software de código abierto se seleccionan herramientas libres, entre las más destacados están JasperReports y Pentaho.

4.2.1. Comparación de herramientas para implementar el Data Warehouse

En la tabla 4.1 se realiza una comparativa más exhaustiva sobre cada una de las herramientas de BI mencionadas previamente.

Producto	**Jasper**	**Pentaho**
Licencia	JasperReports LGPL V3 iReport GPL V3 Modelo normalizado	Pentaho Reporting LGPL V2.1
Lenguaje de desarrollo	Java, Perl	Java
Plug-in	Eclipse Plug-in Available	Eclipse Plug-in Available
JDBC Driver	6	37
Compilación de Reporte	SI	NO
Plataforma	Windows, Linux, Mac OS X	Windows, Linux, Mac OS X
Servidor de Aplicaciones	JBoss	JBoss
Servicios web	Tomcat	Tomcat
Reporte/Gráficos	Si (JasperReport)	Si (Pentaho Report Designer)
Cuadros de Mando	Si (JFreeChart)	Si (JFreeChart)
ETL	Si (JasperETL)	Si (Pentaho Data Integration)
Data Mining	No	Si (Weka)
KPI (Indicadores Clave de Desempeño)	No	Si

Tabla 4.1: Comparación de herramientas para el Data Warehouse

Jasper maneja un modelo de negocio del tipo comercial de código abierto, ofreciendo soluciones al análisis, y servicios de integración de datos con una arquitectura flexible construida en un modelo escalable para que sea integrable con otras aplicaciones.

Pentaho BI es una suite de software integrado que permite a la empresa desarrollar soluciones orientadas al problema. Su plataforma se basa en flujos de trabajos, procesos y definición de procesos que pueden ser integrados fácilmente. Pentaho ofrece una serie de productos como: Mondarían, JFreeReport, Kettle, Weka, etc. Debido a que es una completa gama de programas integrados, la arquitectura de Pentaho se basa en servidores, motores y componentes, muchos de ellos estándares ofreciendo una plataforma de BI escalable y sofisticada.

Basándose el la tabla comparativa de las herramientas tecnológicas analizadas anteriormente para la implementación del Data Warehouse de este proyecto de titulación, la herramienta seleccionada es Pentaho, ya que permite realizar todo el proceso ETL.

Pentaho es una herramienta que combina muchos componentes que facilitan una adecuada administración del Data Warehouse y descubrir nuevo conocimiento en los procesos OLAP y Bibliomining con los componentes que posee.

4.3. Instalación y configuración de las herramientas de desarrollo de Pentaho

Pentaho es una suite de software completa con una gama de programas integrados, entre los componentes a utilizarse en este proyecto de tesis están: Mondrian, Kettle, Weka y BI server.

4.3.1. Instalación y Configuración de Kettle Data Integration

Pentaho Data Integration (PDI) o Kettle, es una poderosa, intuitiva y eficiente herramienta, para la realización de procesos de Extracción, Transformación y Carga (ETL).

Entre las principales características se exponen:

- Multiplataforma
- Herramienta gráfica
- Open Source

Kettle consiste principalmente de las siguientes aplicaciones:

- **Spoon:** Es la herramienta gráfica que permite diseñar Jobs y Transformacions ETL, permite conectar diversos orígenes de datos y transformarlos para ser cargardos dentro de la estructura del Data Warehouse.
- **Kitchen:** Es un programa que permite ejecutar *Jobs* diseñados en Spoon, permitiendo programarlos y ejecutarlos en modo batch.
- **Pan:** Permite ejecutar *transformations* diseñados en Spoon, e incluso ejecutarlos desde línea de comandos.

A continuación se describen los pasos a seguir para la instalación y configuracion de Kettle

1. Para la descarga de la ultima version de Kettle, en este caso la versión 5.0.1. se debe de dirigirse a la direcion web: `http://sourceforge.net/projects/pentaho/files/Data%20Integration/5.0.1-stable/pdi-ce-5.0.1.A-stable.zip/download`. Obtendremos un archivo pdi-ce-5.0.1-stable.zip para Windows y el archivo pdi-ce-5.0.1-stable.tar.gz para Linux.
2. Se crea un directorio en el disco y se descomprime el archivo descargado, por defecto se obtiene el directorio data-integration. (Ver figura: 4.1)
3. Si se trabaja con bases de datos diferentes a MySQL, es necesario descargar los respectivos archivos .jar del JDBC y copiarlos a la ruta *..,/data-integration/libext /JDBC*. (Ver figura: 4.2)
4. Para ejecutar el programa entrar al directorio (.../data-integration) y ejecutar el archivo Spoon.bat en Windows. Si se trabaja en Linux, previamente se debe asignar permiso de ejecución al archivo Spoon.sh antes de ejecutarlo.

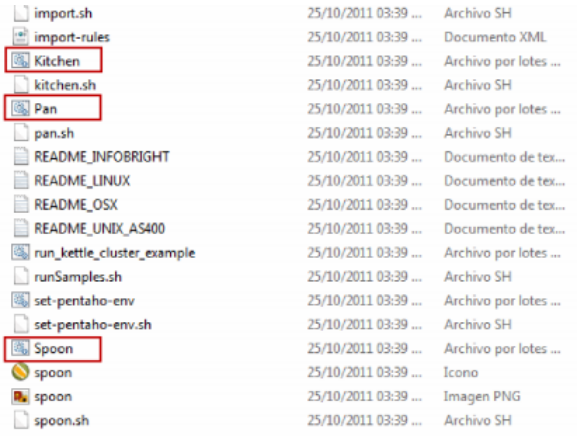


Figura 4.1: Archivos Kettle

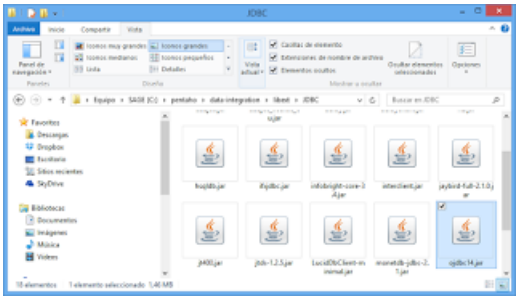


Figura 4.2: Kettle directorio JDBC-Drivers

5. La ventana de inicio permite conectarse a un repositorio predefinido o crear uno nuevo, como se observa en la figura 4.3

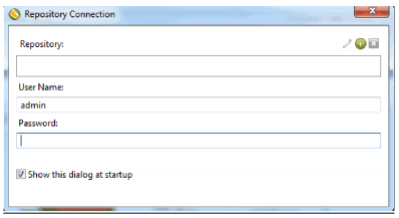


Figura 4.3: Repositorio Kettle

6. La pantalla inicial del Kettle se muestra en la figura 4.4

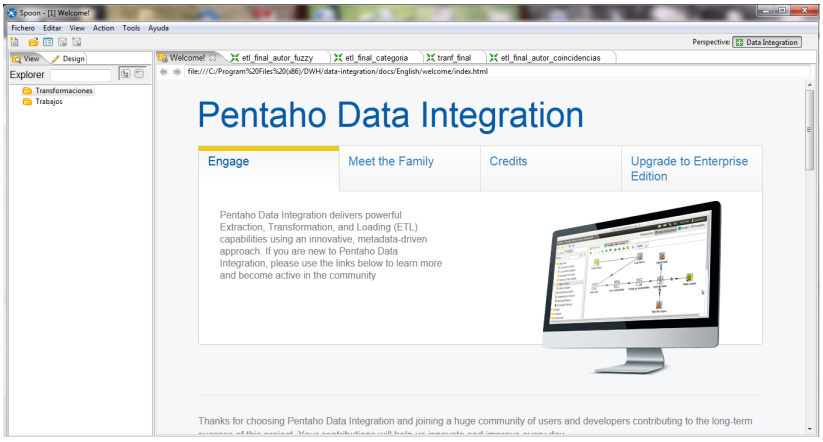


Figura 4.4: Pantalla inicial de Kettle

4.3.2. Instalación y Configuración de Mondrian

Mondrian Schema Workbench es una interfaz de diseño que permite crear y probar esquemas de cubos OLAP visualmente. Estos archivos son modelos de metadatos XML que se crean en una estructura específica y pueden ser consideradas estructuras de forma de cubo que utilizan tablas de hechos y de dimensiones existentes en un RDBMS.

A continuación se describen los pasos a seguir para la instalación y configuración del Schema Workbench:

1. Para la descarga de la última versión de Mondrian, en este caso la versión 3.6.1. se debe de dirigirse a la dirección web: *http : //sourceforge.net/projects /mondrian/files/schema%20workbench/3,6,1 − stable/psw − ce − 3,6,1.zip/ //download*. Se obtiene un archivo psw-ce-3.6.1.zip para Windows y el archivo psw-ce-3.6.1.tar.gz para Linux.

2. Se crea un directorio en el disco y se descomprime el archivo descargado, por defecto se obtiene el directorio schema-workbench. (Ver figura 4.5)

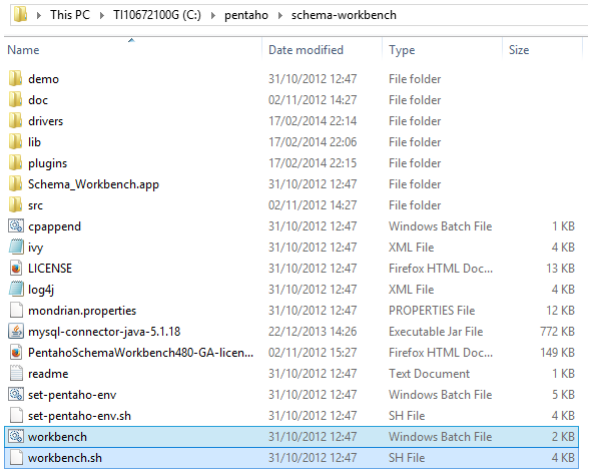


Figura 4.5: Directorio Schema Workbench

3. Si se trabaja con bases de datos diferentes a MySQL, es necesario descargar los respectivos archivos .jar del JDBC y copiarlos a la ruta *../schema-workbench/lib/ JDBC*. (Ver figura: 4.6)

4. Para ejecutar el programa entrar al directorio (.../schema-workbench) y ejecutar el archivo workbench.bat en Windows. Si se trabaja en Linux, previamente se debe asignar permiso de ejecución al archivo workbench.sh antes de ejecutarlo.

5. La pantalla inicial del schema-workbench se muestra en la figura 4.7

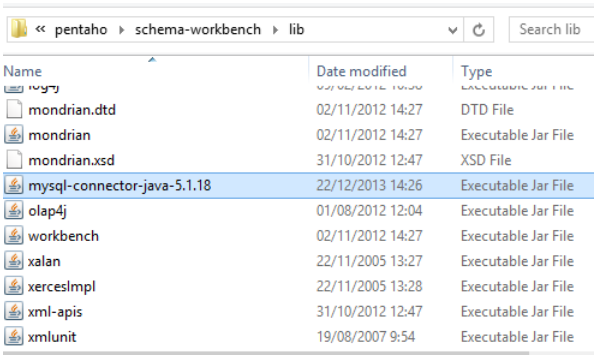


Figura 4.6: Directorio JDBC - Schema Workbench

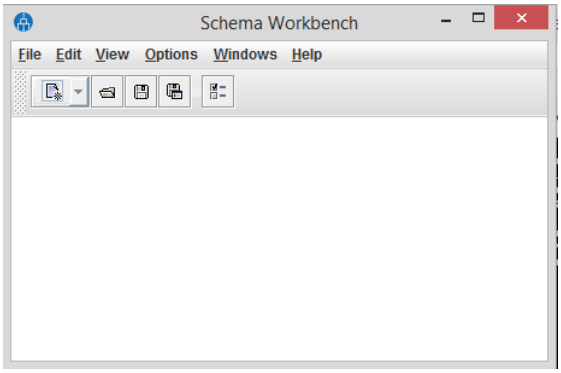


Figura 4.7: Pantalla inicial Mondrian Schema Workbench

4.3.3. Instalación y Configuración de Business Intelligence Server

Pentaho Business Intelligence Server permite publicar en la web reportes generados por sus herramientas que fueron utilizadas con anterioridad para la creación de los cubos, cabe recalcar que BI Server tiene la capacidad de aumentar su funcinalidad al instalar plugins incluso permitiendo generar Cuadros de Mando Integral, entre otros.

A continuación se describen los pasos a seguir para la instalación y configuracion de Business Intelligence Server

1. Descargar BI Server la última versión, en este caso biserver-ce-5.0.1-stable.zip para Windows y biserver-ce-5.0.1-stable.tar.gz para Linux, para la cual se debe dirigirse a la direcion web: `http://sourceforge.net/projects/pentaho/files/latest/download?source=files`. Se obtiene un archivo biserver-ce-5.0.1-stable.zip para Windows y el archivo biserver-ce-5.0.1-stable.tar.gz para Linux

2. Se crea un directorio en el disco y se descomprimir el archivo descargado, por defecto se obtiene el directorio (../biserver-ce)

3. Ejecutar el archivo start-pentaho.bat en Windows o start-pentaho.sh en Linux con las mismas consideraciones explicadas para el caso del Kettle.

4. Abrir el navegador en la dirección: `http://localhost:8080/` se puede ver la pantalla inicial como lo muestra la figura 4.8

5. Los datos para el ingreso por defecto son:

   Usuario: Admin **Password**: password
   Todos los permisos

   Usuario: Suzy **Password**: password
   No puede crear data-sources

6. Pantalla de inicio en el servidor.

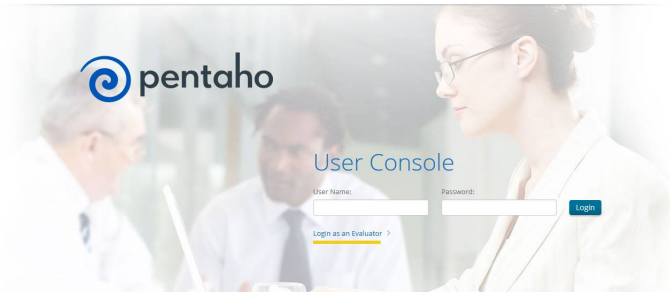
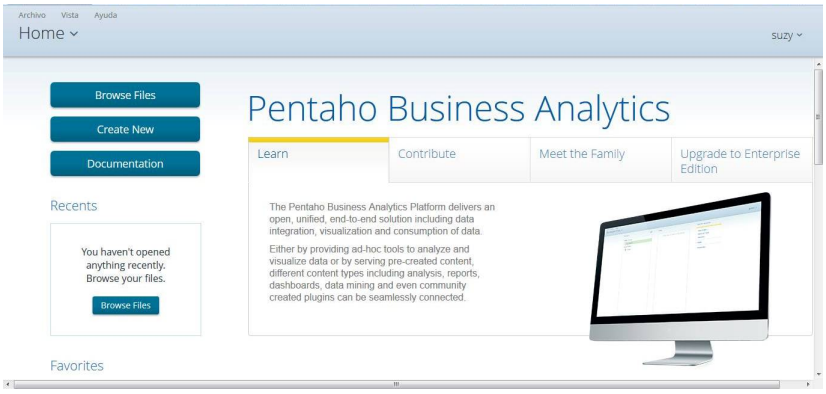


Figura 4.8: Pantalla inicial del BI-Server



4.4. Implementación del componente de integración de información para base de datos documentales

Pentaho posee acceso directo a bases de datos relacionales y también para archivos de datos pero no dispone de una conexión para las bases de datos documentales, a continuación se explica la alternativa que se utilizó para acceder a esta información.

4.4.1. Problema de acceso a la base de datos documental Isis

Pentaho no dispone de componentes que permita acceder a base de datos documentales, por lo que se procedió a crear el archivo MARC de los datos que dispone el centro de documentación del material bibliográfico como primer paso. Debido a que los archivos de tipo MARC son de extención .mrc se buscaron herramientas que permitan acceder y crear este tipo de formatos desde java.

4.4.1.1. Herramientas de acceso a archivos MARC

4.4.1.1.1. MARC4J .-

Es una Biblioteca de código abierto para trabajar con registros MARC y MARCXML en Java. La biblioteca MARC4J incluye:

- Una interfaz fácil de usar que puede manejar grandes conjuntos de registros.
- Implementación y apoyo independiente a través de XML, JAXP y SAX2, es una interfaz XML de alto rendimiento.
- Soporte para conversiones entre MARC y MARCXML.
- Documentación Javadoc.

Aunque MARC4J está diseñado principalmente para el desarrollo de Java se puede utilizar las utilidades de línea de comandos org.marc4j.util.MarcXmlDriver y org.marc4j.util.XmlMarcDriver para convertir entre MARC y MARCXML.

Permite la conversión de archivo ISO-2709 [1] a archivos MARCXML y viceversa, gracias a las funciones que posee. Entre las principales funciones están:

- MARC4J4R::Writer
- MARC4J4R::Record
- MARC4J4R::ControlField
- MARC4J4R::DataField
- MARC4J4R::SubField

4.4.1.1.2. JAVAMARC .-

Es una API de Java que permite controlar los registros MARC mediante una interfaz de aplicación genérica a los registros con el tipo de formato ISO 2709. Esta herramienta permite controlar registros MARC sin la necesidad de conocer la estructura del formato ISO 2709.

La principal desventaja de esta herramienta es que no posee documentación suficiente sobre el uso de la librería.

4.4.1.1.3. FRBR .-

FRBR (Funcional Requirements for Bibliographic Records) es una recomendación de la Federación Internacional de Asociaciones de Bibliotecarios y Bibliotecas (IFLA) para reestructurar las bases de datos de catálogo y reflejar la estructura conceptual de los recursos de información.

La herramienta de visualización FRBR es un programa que transforma los datos bibliográficos que se encuentran en los archivos de registros MARC mediante la agrupación de los datos bibliográficos a un modelo conceptual.

4.4.1.2. Comparación de herramientas para manipular archivos MARC

Por las ventajas que dispone MARC4J se seleccionó está librería para la implementación de un programa que permita generar un archivo MARC de extensión .mrc, con formato MARC21 a partir de la base documental IsisDB.

[1] Estándar ISO para la descripción bibliográfica

Producto	MARC4J	JAVAMARC	FRBR
Lenguaje de desarrollo	Java	Java	Java
Archivo base	.mrc	.iso (ISO-2709)	
Conversión a MARCXML	SI	SI	SI
Interfaz	NO	NO	NO
Documentación	Poca	Mucha	Poca

Tabla 4.2: Comparación de herramientas para el acceso a archivos MARC

4.4.2. Implementación del software para generar un archivo MARC

4.4.2.1. Funciones principales de la librería MARC4J

Entre las principales funciones y clases que dispone de la librería MARC4J están : (Peters, 2007)

- MarcXMLWriter: Para escribir con formato MARCXML.
- MarcWriter: Escribe en formato MARC ISO2709.
- Record: Permite crear un registro MARC.

  addVariableField: Agrega un campo(dataField) a un registro.

  getLeader: Devuelve los caracteres de la cabecera.
- DataField: Crea un campo que se debe agregar al registro.

  addSubfield: agrega un subcampo para el registro.
- Leader: Representa la cabecera de un registro.

4.4.2.2. Tratamiento de la cabecera

Al generar el archivo MARC se encontraron una serie de errores como: el hecho de que los registros no tengan cabecera, la fecha almacenada en el campo de control 005 no debería tener subcampos y contenidos en los campos que almacenan la palabra **marc** sin razón alguna. Todas estas inconsistencias hacen que los registros no coincidan con el formato MARC21 por lo que se deben solucionar, para que los

datos generados sean de calidad.
La información del material bibliográfico que dispone el Centro de Documentación "Juan Bautista Vázquez" tiene una serie de inconsistencias que hacen que la mayor parte de los datos no sean válidos. Estos problemas se han acarreado desde la migración de datos de un sistema WinIsis que fue modificado de acuerdo a las necesidades de este centro de documentación al sistema ABCD utilizado hasta la actualidad. El mayor problema consiste en que los registros no contienen información en el leader (cabecera), esta información podría ser agregada en el caso de que los campos expuestos a continuación tuvieran un valor no nulo según lo recomendado por los bibliotecarios del centro documental analizado:

- 3005 – Estado del registro.
- 3006 – Tipo de registro.
- 3007 – Nivel Bibliográfico.
- 3017 – Nivel de codificación.
- 3018 – Descripción de la forma de catalogación.

La cabecera o leader del registro esta formado por 24 caracteres que corresponden a la información mostrada en la tabla 4.3.

En el centro de documentación "Juan Bautista Vázquez" la muestra trabajada inicialmente contiene 149931 registros de material bibliográfico de los cuales 54755 son datos válidos, a los demás registros se tiene que dar un tratamiento para convertirlo en información útil y así trasladarlos al Data Warehouse.

Los 95176 registros restantes tienen el campo 3006 y 3007 vacíos lo cual no permite que se pueda reconocer el tipo de registro y el nivel bibliografico que poseen para la creación del leader del formato MARC21 que corresponden al carácter 05 y 06. En este centro de documentación al momento de migrar los datos dejaron en el campo 007 información válida que ayuda a identificar el tipo de registro que pertenece al carácter 06 del leader y el nivel bibliográfico que corresponde al carácter 07 del leader, se describe en la tabla 4.4 la información que se encuentra en el campo 007.

Posición	Descripción
00-04	Longitud de registro
05	Estado del registro
06	Tipo de registro
07	Nivel bibliográfico
08	Tipo de control
09	Código del esquema
10	Conteo de indicadores
11	Conteo del código de subcampo
12-16	Dirección base para los datos
17	Nivel de codificación
18	Forma de la catalogación descriptiva
19	Requisito del registro ligado
20	Longitud de la porción que da la longitud del campo
21	Longitud de la porción que da la posición del carácter de inicio
22	Longitud de la porción definida de la implementación
23	No definida

Tabla 4.3: Definición del leader

4.4.2.3. Tratamiento de los campos

Campo 005

El campo de control 005 almacena la fecha de la última modificación del registro en 16 caracteres que cumple el formato: aaaammddhhmmss.ms = 20131122102341.0 (22 de Noviembre del 2013, 10:23:41)

El formato encontrado en este campo de los datos es muy variado como lo muestra la figura 4.9. El total de registros que poseen información siguiendo el formato mencionado son 84154, de los faltantes 65138 pueden ser tratados y 639 registros tienen el campo vacío.

WINISIS	DESCRIPCION	ABCD	Observaciones	Nro. Registros
M / m / m / c	Identifica una colección monográfica	3006: s, 3007: m	El campo 490 no será nulo	80
M / m / m	Identifica un material textual monográfico(libro)	3006:a, 3007:m		53463
S / a / as	Identifica una revista	3006: s, 3007: b	Campo 362 no nulo	71
S / m/ ms	Identifica una revista monografica seriada	3006: s, 3007: m	Campo 362 no nulo	22
S/ s / s	Identifica una revista	3006: s, 3007: s	Campo 362 no nulo	134

Tabla 4.4: Valores campo 007

Campo 041

Es un campo repetible que guarda la información del idioma del material bibliográfico en el siguiente formato: 041 ## ˆa codˆh cod2 Los cod siguen el formato ISO 369-1 utilizado para códigos de representación de nombres de idiomas utilizando 3 caracteres. Existen 55754 registros que siguen el formato adecuado, mientras que con los 84626 restantes se hará una limpieza de datos, estos registros se encuentran almacenados (Ver figura 4.10) de una de las siguentes formas:

- 041 Español = No posee indicadores ni etiqueta de subcampo
- 041 ˆa eng = No posee indicadores

Además, los restantes 9546 son valores nulos.

Campo 500

Campo utilizado para almacenar notas general siguiendo el formato: *500 ## ˆa Notas* Existen 4500 datos almacenados correctamente y 108127 que no tienen el campo vacío pero uno de los errores encontrados en este campo es que el sistema ABCD guarda la palabra marc después del valor del campo por ejemplo:

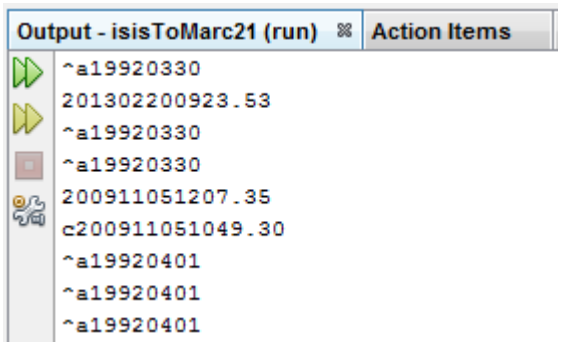


Figura 4.9: Tratamiento de campos: Valores del campo 005

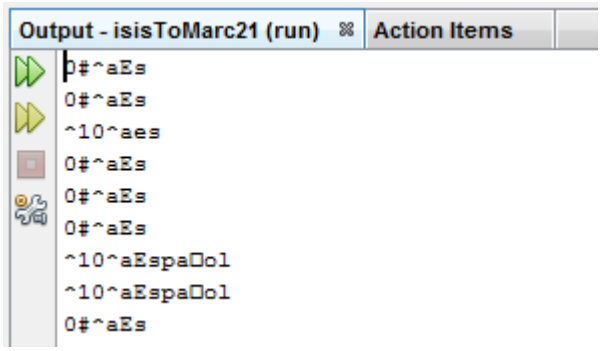


Figura 4.10: Tratamiento de campos: Valores del campo 041

- 500 ## marc
- 500 ## libro compradomarc
- 500 ## libro donadomarc

Una muestra de datos encontrados en el campo 500 se muestra en la figura 4.11

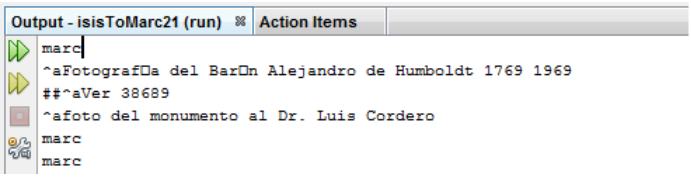


Figura 4.11: Tratamiento de campos: Valores del campo 500

En la siguiente tabla se presentan los valores de cada uno de los campos a los que se dio un tratamiento para las correcciones iniciales.

Campo	Nro de Registros
005	65138
041	84626
500	108127
007	53770

Tabla 4.5: Tratamiento de Campos

4.4.2.4. Implementación del Software IsisToMarc-Java

Una vez modificados los datos en los que se encontró error, se genera un archivo de extensión .mrc con formato MARC21, considerando todo el tratamiento de errores indicado anteriormente en las secciones 4.4.2.2 y 4.4.2.3 se obtiene esta herramienta que conversión denominado IsisToMarc-Java. El archivo generado es interpretado desde Pentaho Data Integration leyendo los datos en el formato (MFN-Campo-Subcampo-Dato) como se presenta en la figura 4.12, que además indica el proceso realizado.

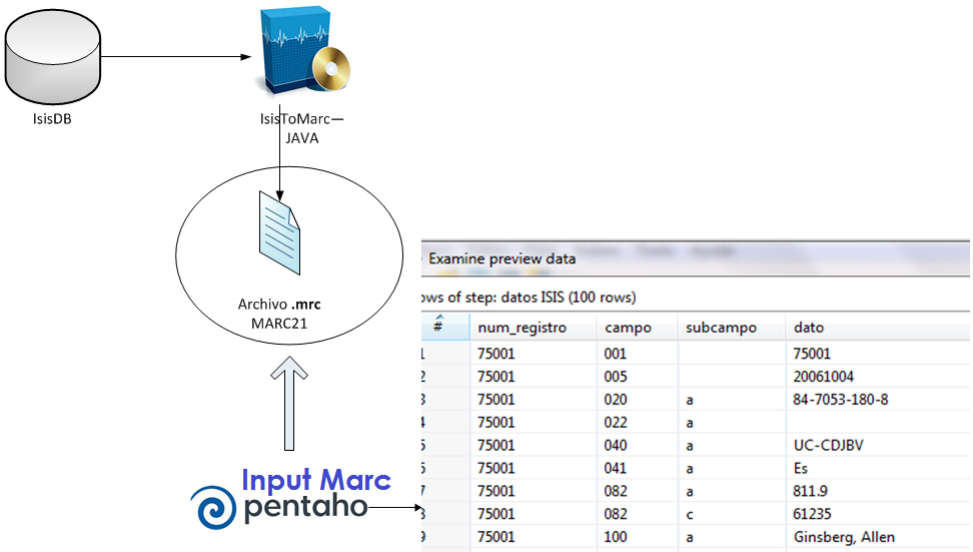


Figura 4.12: Extracción IsisDB hasta lectura mediante Input Marc Pentaho

Al identificar los problemas en los datos de cabecera como en los campos y después de haber hecho las limpiezas correctas sobre los registros leídos desde la base de datos documentales Isis, se procede con el desarrollo del software.

Al desarrollar el software se consideró dos tipos de procesos que se pueden ejecutar:

- **Carga iniciales.-** En este tipo de proceso se transforma todos los registros de la base *Isis* a un archivo *.mrc*. Al aplicar este tipo de ejecución el costo de procesamiento es muy alto, donde el tiempo aproximado de ejecución fue de 120 horas para completar el proceso de carga inicial.

- **Actualizaciones.-** Debido a que los tiempos de ejecución eran altos, se tuvo que plantearse una solución para las actualizaciones del Data Warehouse, por lo que se definió parámetros de ingreso adicionales a los de la carga inicial, éstos son el rango de fechas de ejecución. Con estos parámetros definidos se filtran los registros a transformar en un archivo mrc, esto se logra con la comparación del "campo 005" con el rango de fechas definidas por el usuario.

Figura 4.13: Interfaz de software de conversión de datos isis a registros mrc

En la figura 4.13 se presenta la interfaz del software implementado en la que se aplica la carga inicial de datos. Al mismo que se debe ingresar dos parámetros:

el directorio en la que está ubicado la carpeta de la base de datos Isis y el path de ubicación del archivo *.PFT*. El PFT es el encargado de definir las reglas y la estructura de lectura de una base de datos Isis. En caso de que se quiera hacer una actualización, se debe seleccionar la opción *Actualización en intervalo de tiempo* y definir el rango de fechas en la que se va hacer la lectura de los registros Isis.

4.5. Procesos ETL

Los procesos ETL(Extract, transform and load) permiten copiar los datos de una fuente y cargarlos en otra. Se extraen los datos a partir de las fuentes de datos OLTP, posteriormente se realiza una limpieza de datos y se transforman los mismos para que coincidan con los modelos en los que serán cargados finalmente. (Davenport, 2008). Este proceso se muestra en la figura 4.14

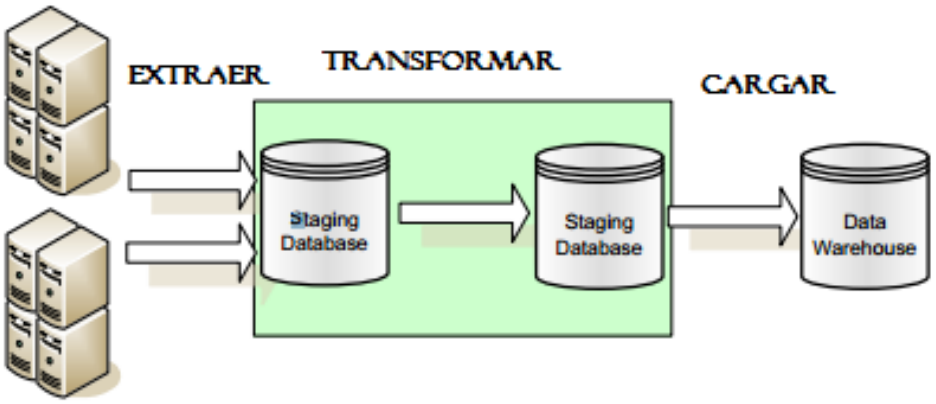


Figura 4.14: Procesos ETL

4.5.1. Extracción y Transformación

El primer proceso ETL es la extracción de datos de las distintas fuentes de información sean estás bases de datos relacionales o archivos planos. A los datos extraídos en la fase anterior se realizan una serie de transformaciones para que sean cargados a la nueva base de datos. Las mayoría de las transformaciones realizadas para el

proceso de préstamos se aplican sobre la base de datos documental que se presentan a continuación.

4.5.1.1. Tranformación para la dimensión Categoría

Se utiliza el sistema de codificación decimal Dewey explicado anteriormente para clasificar el material bibliográfico, en la base de datos se almacena el código de la categoría.

El proceso después de extraer los datos consiste en realizar un mapeo de valores para agregar la categoría y la subcategoría a la que pertenece cada material bibliográfico para así realizar consultas directamente sobre estos datos sin necesidad de conocer este sistema de codificación. Los datos utilizados para realizar el mapeo se muestran en la figura 4.15.

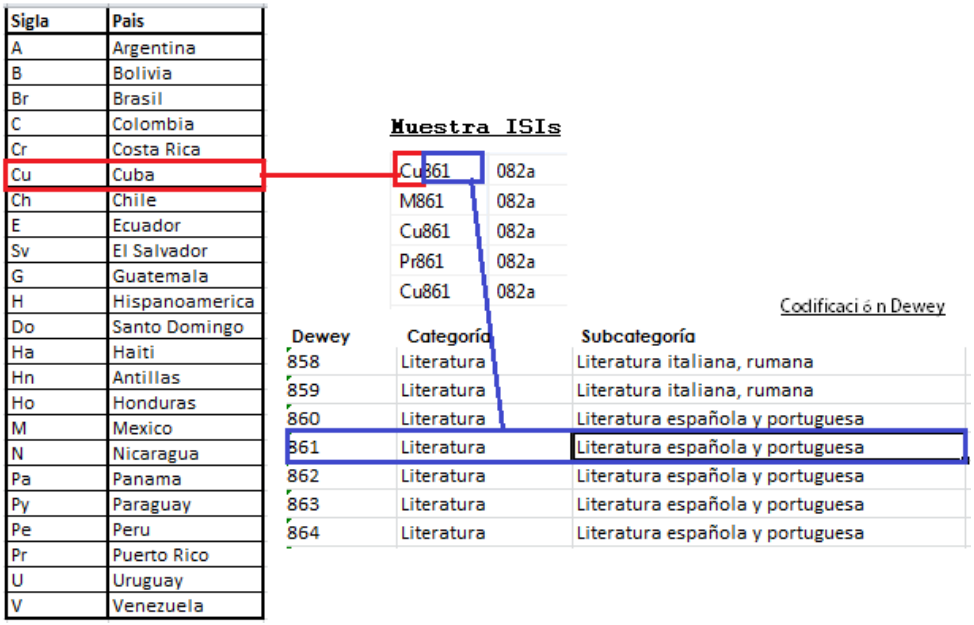


Figura 4.15: Mapeo Dewey

En el sistema de codificación Dewey los códigos del 860 al 869 están reservados para la literatura española y portuguesa, pero es necesario realizar una modificación a estos campos para especificar el país hispanoamericano al que referencia, para esto se manejan una serie de siglas expuestas en la figura 4.16

Sigla	Pais
A	Argentina
B	Bolivia
BR	Brasil
C	Colombia
CR	Costa Rica
CU	Cuba
CH	Chile
E	Ecuador
SV	El Salvador
G	Guatemala
H	Hispanoamerica
DO	Santo Domingo
HA	Haiti
HN	Antillas
HO	Honduras

Figura 4.16: Siglas Literatura Hispanoamericana

Los códigos almacenados en el campo 082c cuando hacen referencia a Literatura española o portuguesa tiene el siguiente formato: "SiglaLiteratura" + "Código Dewey", por ejemplo: E864.4 que hace referencia a la categoría Literatura, subcategoría Literatura española y portuguesa y la sigla E que indica que es Literatura ecuatoriana.

Código	Categoría	Subcategoría
TA	Tesis	Arquitectura
TV	Tesis	Artes Visuales
TS	Tesis	Informática
TI	Tesis	Ingeniería
TO	Tesis	Topografía
TQ	Tesis	Ciencias Químicas
TN	Tesis	Ingeniería Industrial
TT	Tesis	Tecnología
TDI	Tesis	Derecho Internacional/Territorial
TRBS	Tesis	Trabajo Social
TOF	Tesis	Orientación Familiar
TECO	Tesis	Economía
TAD	Tesis	Administración de empresas
TCON	Tesis	Contabilidad superior y auditoría
TLDA	Tesis	Desarrollo en derechos amazónicos
TSOC	Tesis	Sociología Gestión Social

Figura 4.17: Siglas Tesis Universidad de Cuenca

Adicionalmente, en el Centro de Documentación "Juan Bautista Vázquez" se manejan siglas para identificar las tesis que se realizan en las diferentes carreras ofertadas por la Universidad de Cuenca, para las cuales también se realiza un mapeo considerando las siglas proporcionadas por los bibliotecarios, un ejemplo de las siglas se muestra en la figura 4.17.

Finalmente el proceso ETL para la creación de la dimensión Categoría se muestra en la figura 4.18

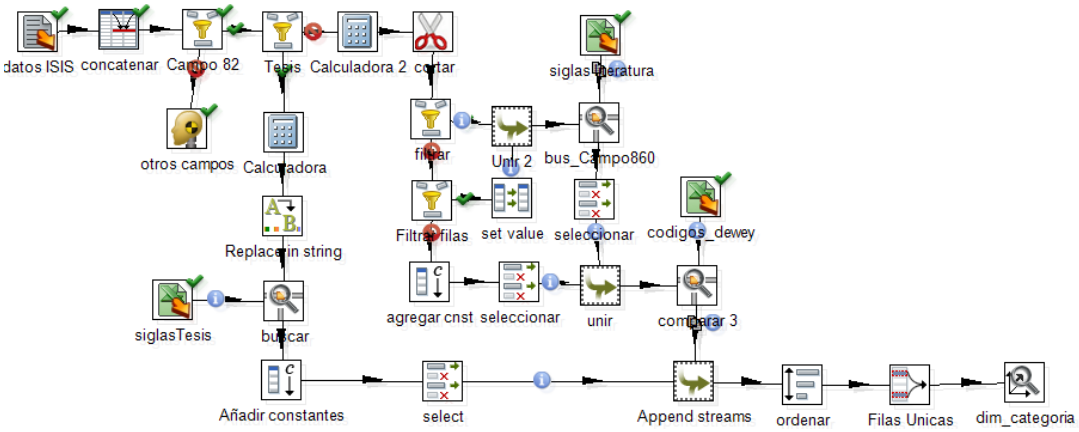


Figura 4.18: Transformación Categoría

4.5.1.2. Tranformación para la dimensión Bibliotecario

Al analizar los datos de los bibliotecarios que realizan los préstamos o que catalogan los libros se encontró que un mismo usuario dispone de varios nombres de usuario, para encontrar estos parecidos se utilizo el algoritmo Jaro Winkler para analizar el porcentaje de similitud entre dos cadenas de caracteres, el proceso ETL se muestra en la figura 4.19.

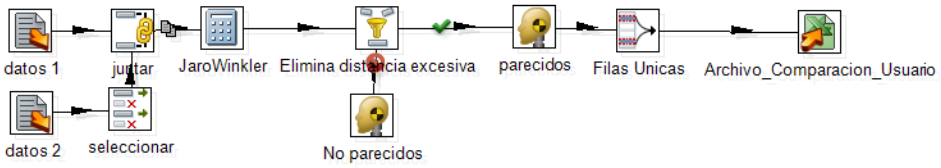


Figura 4.19: Comparación de nombres de usuarios de los bibliotecarios

Del ETL de comparación de usuarios se obtiene un archivo que debe ser analizado para seleccionar los parecidos razonables, tarea que debe ser desarrollada manualmente por alguien que tenga conocimiento sobre el personal del centro de documentación. En la figura 4.20 se puede observar una muestra de la tabla generada para el mapeo de datos para así definir un solo nombre de usuario para un mismo bibliotecario.

Estos datos fueron validados conjuntamente con el personal del Centro Documental "Juan Bautista Vázquez" específicamente con la Lic. Margarita Guitierrez

usuarioBien	usuario	comparacion
acriollo	acq&mkoq	,85
acriollo	acq	,71
acriollo	malicri	,70
bcabrera	ccabrera	,87
mquezada	dva-mqb	,87
mquezada	dva	,87
mquezada	d.v.a.	,80
emendez	edwinmendez	,80
eduran	edh	,73
epeñafiel	epe?afiel	,96
epeñafiel	ipenafiel	,80
fcriollo	fcm	,71
gmartinez	gamartinez	,94
gmartinez	gartinez	,97
gmartinez	gima	,73
gmartinez	gma	,84
gmartinez	gmartine	,98

Figura 4.20: Datos para el mapeo de los nombres de usuarios de los bibliotecarios

encargada de la catalogación y más actividades bibliotecarias.

4.5.2. Carga

Una vez que se ha realizado la limpieza de datos y cargado las dimensiones necesarias para el proceso de préstamos, se procede a realizar la carga de la tabla de hechos y las dimensiones fecha de préstamo, fecha de devolución y tipo usuario del proceso ETL que se muestra en la figura 4.21.

Inicialmente en el proceso se extraen los registros de la base de datos de préstamos provenientes del servidor del centro de documentación, para recuperar el id-libro se carga los datos del modelo multidimensional de catalogación *dim-libro*. Se realiza el join respectivo en el flujo para continuar con las búsqueda de los demás datos y obtener todos los campos necesarios para la tabla de hechos.

Siguiendo el proceso se realiza un filtro de datos ya que existen préstamos aún no devueltos, para los cuales no se necesita cargar la *fecha-devolución* ya que si se envia a cargar todos los datos, resulta un error debido a que los campos de fecha nulos no son reconocidos en el componente de entrada de tabla de la herramienta Kettle.

Posteriormente, se hace un mapeo de fechas para colocar los meses en un string cuyo dato será muy útil al el momento de generar los reportes.

Para recuperar el costo operacional del proceso préstamo que será un dato a considerar como hecho de la tabla se accede al esquema TDABC.

Debido a que los datos de préstamos solo disponen un identificador único del

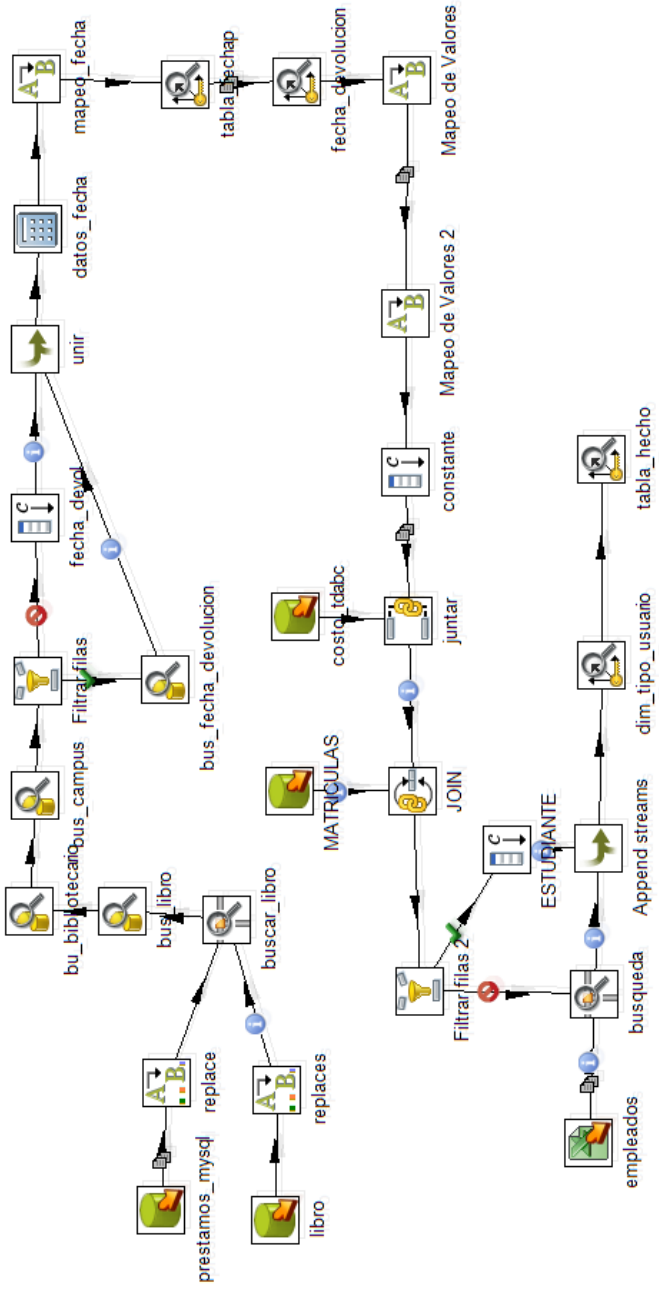


Figura 4.21: Transformación para el proceso Préstamos

usuario, y no se detalla mayor información se acceden a dos esquemas externos para extraer información. Al utilizar como parámetro la fecha del préstamo se puede determinar si un usuario es estudiante cuando conste en los registros de matrículas en esas fecha, o si es un funcionario si forma parte de la fuente de datos de empleados,trabajadores o pasantes.

Finalmente, se hace referencia a los identificadores de cada dimensión y a los datos numéricos recuperados para almacenarlos en la tabla de hechos denominada fact-prestamos.

4.6. Actualización

Antes de empezar a definir procesos de actualización se debe considerar el tipo de dimensión que se va ha actualizar, si es una dimensión de cambio lento se utiliza el componente *Lookup/Update in Dimension* el cual permite configurar un campo de comparación para los datos y definir los campos a actualizarse si éste encontró alguna coincidencia, considerando siempre que el campo clave no cambiará.

Para las dimensiones que tienen un incremento frecuente de los registros en la tabla se utiliza el componente *Lookup/Update in Combination* en el cual se eligen los datos de la tabla y siempre que se encuentre una tupla distinta a las ya ingresadas se almacena en la base de datos del modelo multidimensional, mientras que si al comparar todos los datos de la tupla hay uno en la base de datos ya no realiza ninguna operación y continua leyendo los registros. Es decir, esta herramienta hace una comparación total de los campos de los registros con el nuevo registro a ingresar mientras que la herramienta anterior permite mantener una clave de comparación específica.

La figura 4.22 muestra el Trabajo(Job) creado para la actualización del modelo multidimensional préstamo, se inicia el proceso y se repite cada 15 días inicialmente confirmando la conexión con el servidor que contiene la base de datos de préstamos, posteriormente se carga la trasformación autor, libro y editorial que tienen un tratamiento distinto de la información y finalmente se inicia la transformación de préstamo que almacena las dimensiones fecha-prestamo, fecha-devolucion, tipo-usuario y la tabla de hechos del proceso préstamos.

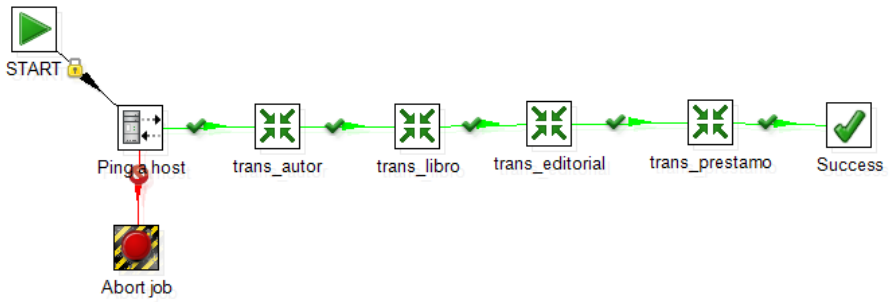


Figura 4.22: Job Préstamo

4.7. Generar los cubos

La generación de los cubos se realiza utilizando la herramienta Mondrian Schema Workbench en la que se crea el esquema con los cubos deseados y éstos a su vez se componen de las dimensiones con sus respectivos atributos almacenados en el modelo multidimencional. Además, se crea la tabla de hechos con los indicadores a los cuales se asigna la función matemática que realizarán. La figura 4.23 muestra la estructura del cubo de préstamos generado.

4.8. Pruebas

Después de generar los cubos y publicarlos en el BI Server se puede generar reportes seleccionando las dimensiones deseadas para responder las preguntas planteadas como requisitos.

Pregunta 1: ¿Cuál es el número de préstamos de un libro de un determinado autor, de una determinada categoría a un tipo de usuario en la biblioteca en una unidad de tiempo?

Al utilizar Saiku como herramienta para realizar los reportes dentro del BI ServerSaiku es un plugin del BI Server que debe ser descargado del Market Place para ser utilizado. se puede seleccionar de forma más sencilla las dimensiones que se desea obtener en el reporte e incluso filtrar valores específicos, la siguiente figura 4.24 muestra un reporte que indica los títulos de los libros disponibles en la biblioteca de los

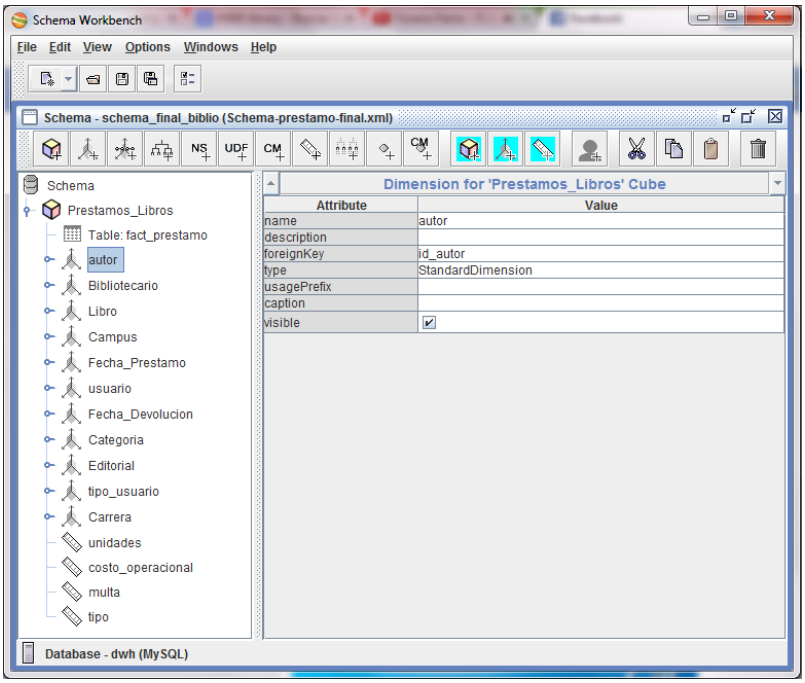


Figura 4.23: Cubo préstamo

cuales se han solicitado préstamos y se indica la cantidad de préstamos de este libro realizados por un bibliotecario.

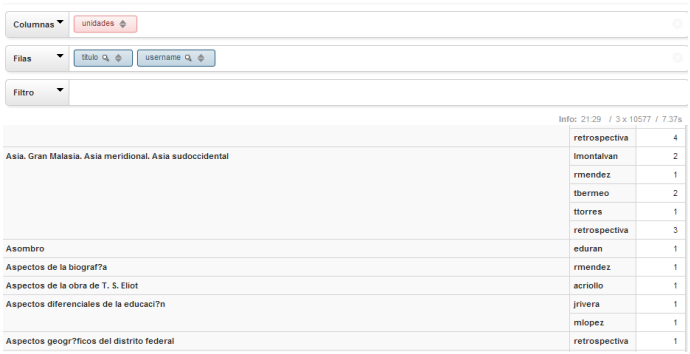


Figura 4.24: Respuesta del Cubo Préstamo Pregunta 1

Pregunta 2: ¿Cuál es el número de préstamos de un libro realizados por un determinado bibliotecario, a un tipo de usuario en la biblioteca en un determinado campus en una unidad de tiempo?

El reporte puede ser observado en forma de tabla con los datos o se puede generar un gráfico como el de la figura 4.25 que muestra la cantidad de préstamos realizados en cada uno de los campus del Centro Documental Juan Bautista Vázquez pudiendo observar que el campus central es el que realiza mayor número de transacciones.

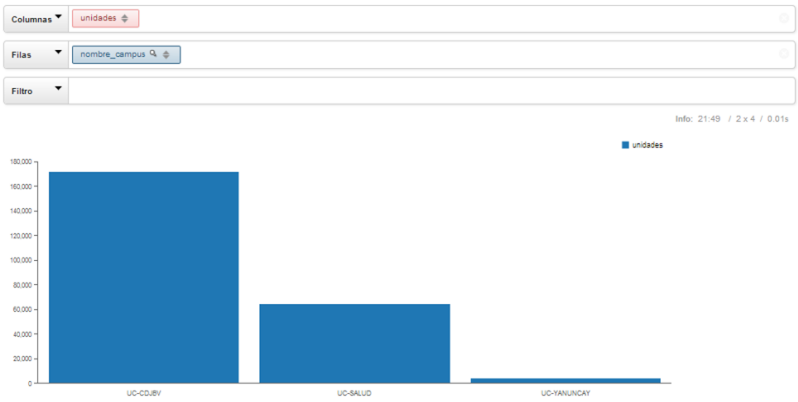


Figura 4.25: Respuesta del Cubo Préstamo Pregunta 2

Pregunta 3: ¿Cuál es el costo operacional de los préstamos realizados en una unidad de tiempo? La respuesta se muestra en la figura 4.26.

La figura 4.27 muestra un reporte con el costo operacional de los préstamos realizados en los diferentes campus realizando un filtro sobre los préstamos registrados en el año 2012.

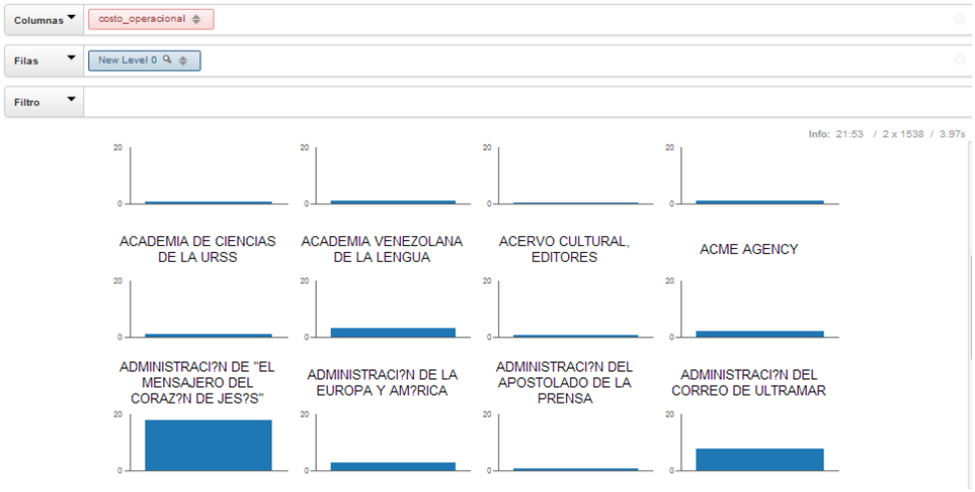


Figura 4.26: Respuesta del Cubo Préstamo Pregunta 3

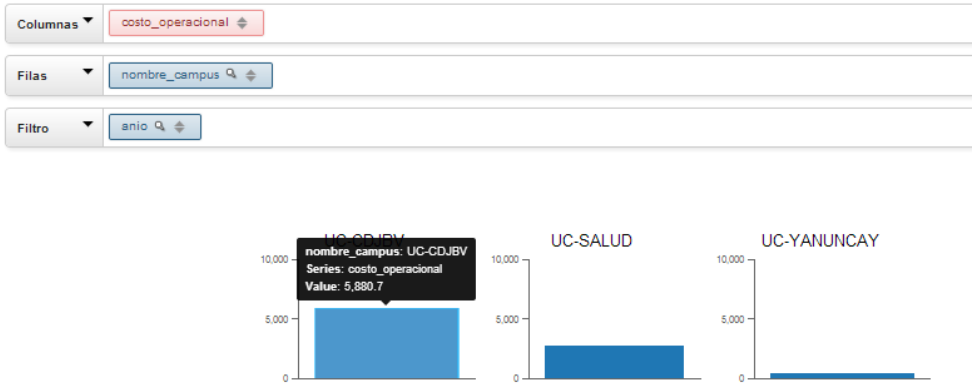
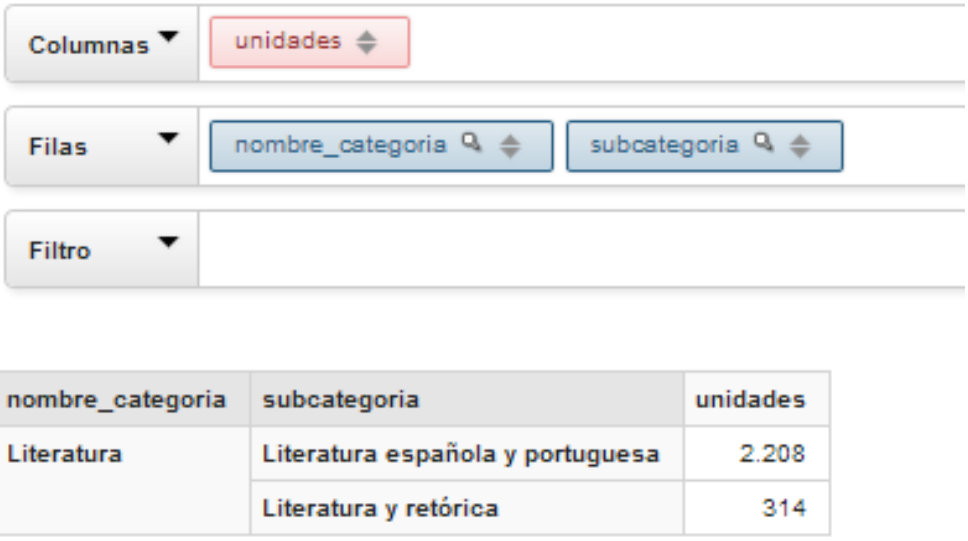


Figura 4.27: Respuesta del Cubo Préstamo Pregunta 3

La herramienta permite filtrar solamente los datos que se necesiten considerar en el reporte en este caso se muestra en la figura 4.28 específicamente la cantidad de préstamos realizados de libros que pertenecen a la categoría literatura, mostrando también la subcategoría.



nombre_categoria	subcategoria	unidades
Literatura	Literatura española y portuguesa	2.208
	Literatura y retórica	314

Figura 4.28: Respuesta del Cubo Préstamo

4.9. Conclusiones

Los ejemplos demuestran que el cubo realizado sobre el modelo multidimensional de préstamos cumple con los requerimientos solicitados permitiendo generar reportes dinámicos según las necesidades del centro de documentación.

Además vale recalcar que a pesar de que no se pueda generar cosultas sobre dos cubos a la vez, se puede encontrar la información deseada accediendo directamente a los datos del modelo multidimensional en la base de datos mediante consultas SQL.

Un ejemplo de lo mencionado se muestra en la figura 4.29 que son los datos resultantes al comparar 2 cuadrantes. Se muestra las catalogaciones realizadas por un determinado bibliotecario en el sistema ABCD(1er cuadrante) y en los repositorios digitales locales (3er cuadrante).

username	catalogación ABCD	catalogación DSPACE
acriollo	220	0
adelosreyes	30	32823
bcabrera	0	239
dnaula	8	392
gmartinez	250	0
epeñafiel	0	92
jmogollon	15	96
lmora	20	576
maguilar	850	0
mfeican	800	0
mbrito	0	10
mgutierrez	9764	357
mquezada	240	0
mmendez	20	924
npeña	100	0
rcampoverde	100	0
salvarado	200	0
ttorres	300	0
tbermeo	10	130

Figura 4.29: Reporte SQL de dos cubos

Este proceso de implementación se utilizó como base de publicación del paper denominado “DESIGN OF AN INTEGRATED DECISION SUPPORT SYSTEM FOR LIBRARY HOLISTIC EVALUATION” realizado por la Ing. Lorena Siquenza, Ing. Victor Saquicela y el Ing. Dirk Cattrysse.

Capítulo 5

Prototipo de Bibliomining

5.1. Introducción

El término KDD (Knowledge Discovery in Databases) descubrimiento de información en bases de datos, es el proceso no trivial de identificación de patrones válidos, nuevos y potencialmente útiles y comprensibles en los datos al aplicar algoritmos de descubrimiento y análisis de datos. (U et al., 1996)

La minería de datos aunque en teoría puede ser aplicada a cualquier tipo de información comúnmente es aplicada a grandes volúmenes de datos de las organizaciones con el objetivo de mejorar el rendimiento de procesos.

La palabra descubrimiento en el ámbito de la minería de datos está relacionado con el hecho de que mucha de la información valiosa es desconocida con anterioridad, por lo que en esta fase mediante técnicas existentes se puede confirmar cualquier sospecha.

En este capítulo se procede a la explotación del Data Warehouse desarrollado en los capítulos III y IV, la misma que es la base para Bibliomining donde se define algoritmos de Data Mining para el descubrimiento de información oculta y la definición de patrones de comportamiento por parte de los usuarios del centro de documentación.

5.2. Algoritmos de Data Mining

Un algoritmo de Data Mining es un conjunto de cálculos y reglas que permite crear un modelo de minería de datos a partir de los datos. Estos algoritmos analizan los datos de entrada, en busca de patrones, encontrando todas las conexiones posibles que pueda haber en toda la información. Para la cual se debe definir los parámetros de ingreso que se analizarán en el conjunto de datos para obtener como resultado patrones de comportamiento en base a los atributos analizados.

Entre la técnicas de Data Mining mas utilizadas estan los algoritmos de predicción, clasificación, clustering y asociación.

5.2.1. Algoritmos de Predicción

Son algoritmos que mediante técnicas y operaciones matemáticas dan como resultado un estimado de la verdad a corto plazo. Estos algoritmos pretenden predecir una o más variables continuas de un conjunto de datos en base a otros atributos o patrones del mismo conjunto.

La veracidad en las predicciones depende del conocimiento y habilidad del usuario, además del conjunto de parámetros utilizados para la predicción que logre la mejor simulación. Varios algoritmos evalúan distintos conjuntos de valores y en base a los resultados de las simulaciones se van mejorando dichos valores con simulaciones posteriores.

Se debe de considerar la utilización de un escenario bueno en un instante de tiempo para predecir qué es lo que pasará en los instantes de tiempo posterior, considerando algún criterio de selección y la combinación de los valores de los parámetros que permita converger hacia combinaciones de valores que den buenas simulaciones.

5.2.1.1. Algoritmo de Regresión Lineal

El algoritmo de regresión ayuda a calcular una relación lineal entre una variable independiente y otra dependiente, toma la forma de una ecuación líneal que mejor represente una serie de datos (Ver figura 5.1). Además de dar como resultado una expresión lógico-matemática que interpreta cómo es esa relación, permitir realizar predicciones de los valores que tomará una de las variables a partir de otra.

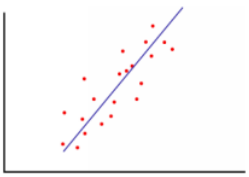


Figura 5.1: Regresión Lineal

Para el análisis de una regresión se considera una variable predictora X y otra variable de respuesta Y, que se modela por la ecuación lineal:

$$Y = \beta_0 + \beta_1 * X_1$$

Los dos parámetros de la ecuación de regresión lineal simple β_0 y β_1, son conocidos como el origen o constante y la pendiente del modelo respectivamente.

Una vez que sean conocidos los valores de β_0 y β_1 del modelo de regresión lineal simple, éste puede ser utilizado como modelo predictivo

Cada punto de datos del diagrama tiene un error asociado con su distancia con respecto a la línea de regresión. Los coeficientes β_0 y β_1 de la ecuación de regresión ajustan el ángulo y la ubicación de la recta de regresión. Puede obtener la ecuación de regresión ajustando β_0 y β_1 hasta que la suma de los errores asociados a todos los puntos alcance su valor mínimo.

5.2.2. Algoritmos de Clasificación

Los algoritmos de clasificación permiten el ordenamiento y disposición de un conjunto de datos en diferentes categorías basados en un criterio de similitud a partir de un patrón común entre los datos. Estas técnicas permiten clasificar un conjunto de objetos en unidades más pequeñas que facilitan su administración, compresión e interpretación.

Los algoritmos de clasificación son del tipo supervisado la cual permite obtener un modelo o regla general para la clasificación y así ayuda tratar casos futuros, donde el sistema sea capaz de aprender de lo que tiene para poder generalizar y tratar lo que no tiene.

5.2.2.1. Algoritmo de Naïve Bayes

Es uno de los algoritmos de clasificación más potente y utilizados en data mining debido a su simplicidad y rapidez. Es una técnica de clasificación y predicción supervisada que construye modelos que predicen la probabilidad de posibles resultados.

5.2.2.2. Algoritmo J48

Conocido como algoritmo C4.5, este permite la predicción y clasificacion basada en la teoría de la información de datos. Es un árbol multinivel que para su calculo realiza la comparación de los valores de información antes y después de cada uno de los posibles candidatos.

Permite trabajar con valores continuos para los atributos, separando los posibles resultados en dos ramas y escoger un rango de medida apropiada.

5.2.3. Algoritmos de Clustering

Los algoritmos clústeres llamadas algoritmos de clasificación o de aprendizaje no supervisado utilizan técnicas iterativas sobre datos de entrada para agrupar elementos de un conjunto de datos con similares características en base a atributos que se conozcan, los mismos que no disponen de un conjunto de entrenamiento por lo tanto no poseen conocimiento a priori. Estos algoritmos no poseen atributos que diferencie la clase a la que pertenece cada una de las instancias de entrada debido a que no poseen información inicial que valide la pertenencia a un determinado clúster.

En los algoritmos clústeres no se elige el campo de predicción para generar clases de agrupación. Lo que se debe de definir es el número posible de clases que se va ha obtener como resultado después del procesamiento.

El objetivo del agrupamiento es clasificar un conjunto de objetos en grupos, de forma tal que los objetos dentro de un grupo posean un alto grado de semejanza, mientras que los pertenecientes a grupos diferentes sean poco semejantes entre sí como se presenta en la figura 5.2.

Distancia entre datos

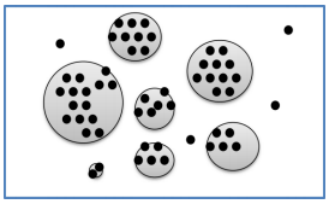


Figura 5.2: Clústeres

Los algoritmos de clústeres usan criterios de comparación de similitud o divergencia entre los datos analizados denominada distancia entre datos. Esta distancia entre dos datos ***i*** y ***j*** es una medida denotada por ***d(i,j)*** que mide el grado de semejanza o la desemejanza entre ambos datos en relación a un cierto número de características cuantitativa o cualitativas dentro de un conjunto. El valor de ***d(i,j)*** es siempre un valor no negativo, y cuanto mayor sea este valor mayor será la diferencia entre los individuos ***i*** y ***j***.

Toda distancia debe verificar, al menos, las siguientes propiedades:

1. No negatividad

$$d(i,j) > 0$$

2. Identidad

$$d(i,i) = 0$$

3. Simetría

$$d(i,j) = d(j,i)$$

4. Desigualdad triangular

$$d(i,j) < d(i,t) + d(j,t)$$

Donde ***i***, *j* y ***t*** pertenecen a un mismo conjunto ***I*** de elementos.

5.2.3.1. Algoritmo Canopy

El algoritmo canopy permite realizar agrupamientos en la cual su operación se basa en realizar cálculos sencillos para generar subgrupos de puntos en la cual cada dato puede pertenecer a más de un subgrupo. Después de la cual usa métodos de segmentación como el k-means con la restricción de no realizar cálculos de distancia entre dos puntos que no pertenecen al mismo subgrupo.

5.2.4. Algoritmos de Asociación

Estos algoritmos permiten descubrir reglas de asociación que ocurren en común entre elementos u objetos que pertenecen a un conjunto de datos. Para la cual se considera todas las posibles combinaciones de atributo-valor de todos los datos almacenados en el conjunto.

La problemática que se pretende resolver con estos problemas es : Dado un conjunto de registros, encontrar reglas que predicen la ocurrencia de un ítem, basándose en las ocurrencias de otros ítems en el registro

Un ejemplo clásico de estos algoritmos es el problema de la cesta de compras, en la cual los modelos de asociación se generan basándose en conjuntos de órdenes de compras que contienen identificadores para productos individuales. Para este caso un grupo de productos de un supermercado se denomina un conjunto de elementos. El modelo de asociación se compone de un conjunto de productos y de las reglas que describen cómo estos productos se agrupan en cada uno de los casos.

Los problemas de asociación se define como:

Sea $I = i_1, i_2, ..., i_n$ un conjunto de n atributos binarios llamados items.

Sea $D = t_1, t_2, ..., t_m$ un conjunto de transacciones almacenadas en una base de datos.

Cada transacción tiene un identificador único y contiene un subconjunto de items de $\boldsymbol{I}$. Una regla se define como una implicación de la forma: $X \Rightarrow Y$

Donde:

$X, Y \subseteq I$ y $X \cap Y = \phi$.

Los conjuntos de items $\boldsymbol{X}$ y $\boldsymbol{Y}$ se denominan respectivamente antecedente y consecuente de la regla respectivamente.

5.3. Instalación y Configuración de WEKA

Weka (Waikato Environment for Knowledge Analysis - Entorno para Análisis del Conocimiento de la Universidad de Waikato) es un software de código abierto publicado bajo la Licencia Pública General GNU codificado en java y desarrollada por la Universidad de Waikato. Weka es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Contiene herramientas para el pre procesamiento de datos, clasificación, regresión, clustering, reglas de asociación, y la visualización de información, útil en el desarrollo de nuevos sistemas de aprendizaje de máquina.

Entre las principales características están:

- Multiplataforma
- Herramienta gráfica
- Open Source

A continuación se describen los pasos a seguir para la instalación y configuracion del Weka.

1. Para la descarga de la ultima version de Weka, en este caso la versión 3.6.11 para desarrolladores se debe de dirigirse a la direcion web: `http://sourceforge.net/projects/weka/files/`. Se obtiene un archivo weka-3-6-11jre.exe para Windows y el archivo weka-3-6-11jre.tar.gz para Linux
2. Se procede con la ejecucion del instalador.
3. Si se trabaja con bases de datos diferentes a MySQL, es necesario descargar los respectivos archivos .jar del JDBC.

4. Antes de ejecutar el weka, se debe de crear el *CLASSPATH* en la que se indique la ruta del driver de conexión como se indica en la figura 5.3

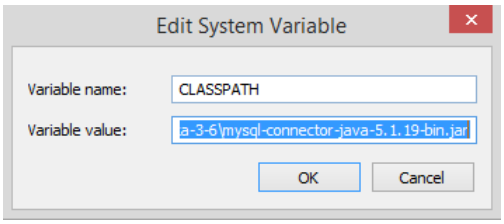


Figura 5.3: CLASSPATH

5. Una vez configurada se procede a ejecutar. La pantalla inicial de Weka se presenta en la figura 5.4:

Figura 5.4: Interfaz de Weka

5.4. Bibliomining: Aplicación de Algoritmos

A continuación se procede con la aplicación de los algoritmos estudiados, para el análisis de Bibliomining en el centro de documentación "Juan Bautista Vázquez", implementando un algoritmo de predicción, un algoritmo de clasificación, un algoritmo de asociación y un algoritmo de clustering.

Para este análisis partiremos de los datos obtenidos en los capítulos anteriores (Capitulo III y Capitulo IV) que dio como resultado el Data Warehouse del centro de documentación.

5.4.1. Predicción de costos en compras de libros

Definición del problema

Predecir el presupuestos de inversión en la adquisición o compra de libros por parte de la Universidad de Cuenca.

Este análisis se enfoca en el costo de la adquisición de cada libro, en donde los parámetro a analizar serán el "costo" frente al "tiempo". Los periodos de tiempo que se analizara para la predicción serán mensuales.

Elección del algoritmo

El problema planteado intenta obtener como resultado la predicción del presupuesto que necesita la Universidad de Cuenca en los siguientes meses para la adquisición de libros. Se utilizará el algoritmo más común en predicción que es el de "Regresión Lineal" analizado previamente.

Obtención de Datos

Los datos a analizar serán extraídos directamente del esquema multidimensional del Data Warehouse, este esquema esta almacenado en MySQL como gestor de base de datos.

Para la obtención de los datos se define la consulta SQL que se muestra en la figura 5.5, utilizando la base de datos "datawarehouse" del esquema adquisición presentado en la imagen 5.6.

```
SELECT da.fechaAdquisicion, sum(fa.valor) as valor
FROM datawarehouse.fact_adquisicion fa,
     datawarehouse.dim_fecha_adquisicion da
where fa.id_fecha_adquisicion=da.id_fecha and anio>1999
group by da.anio,da.mes;
```

Figura 5.5: SQL de extracción de datos

Como se observa en la consulta, se tiene dos variables a analizar:

- **fechaAdquisicion.-** Es la variable independiente que se proyecta en el tiempo para la predicción planteada, la predicción será realizada en intervalos de tiempo mensuales.

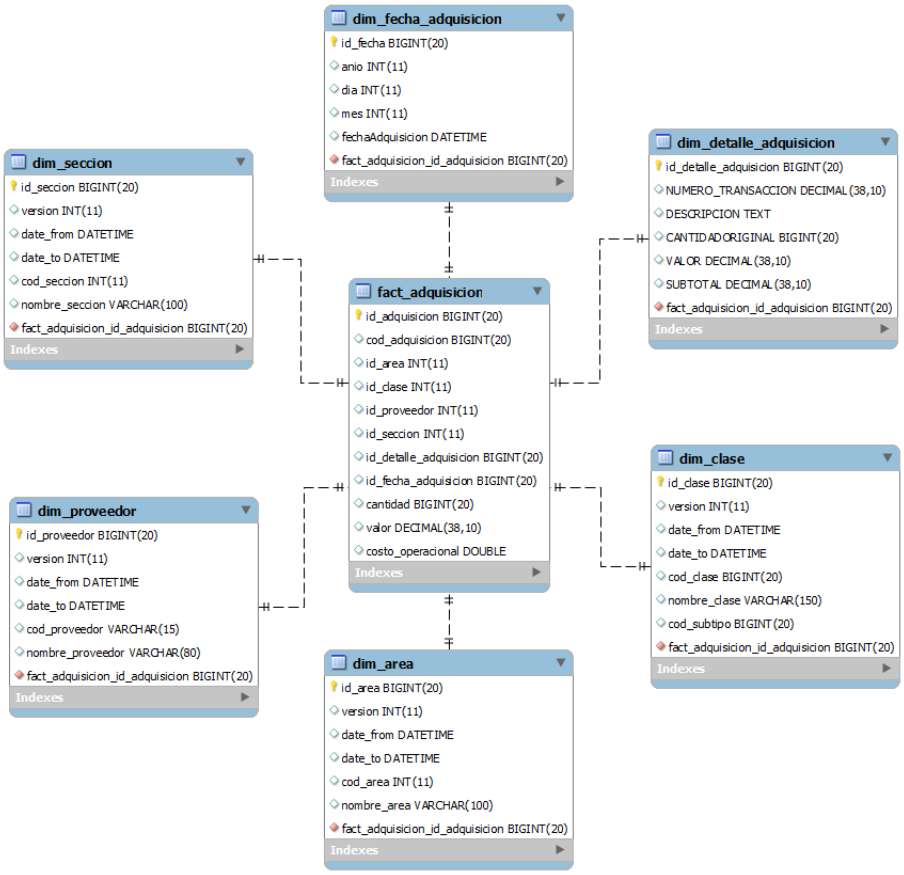


Figura 5.6: Modelo multidimencional: Adquisición

- **valor.-** Es la variable dependiente sobre la cual se predece los costos de compra de libros.

Minería de datos

Características principales del modelo:

- **Estructura a la que pertenece:** Esquema adquisición
- **Algoritmo a utilizar:** Regresión Lineal y Proceso Gausiano.
- **Objetivo del modelo:** Predecir los costos de compra de libros en los siguientes meses.
- **Atributos a predecir:** Costo

Aplicación de Algoritmo

Una vez identificado los atributos y después de tener codificada la consulta SQL se procede a la aplicación y ejecución del algoritmo seleccionado.

El primer paso a realizar es la conexión de WEKA con la base de datos, este procedimiento se presenta en la figura 5.7.

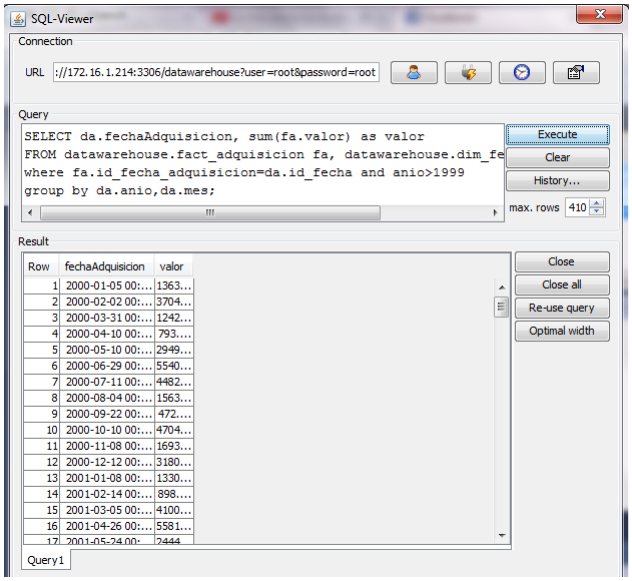


Figura 5.7: Conexión de WEKA con la base de datos

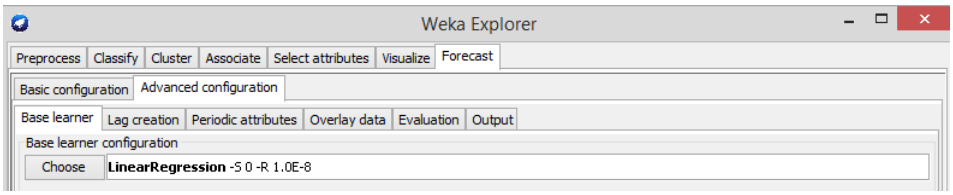


Figura 5.8: Selección de Algoritmo

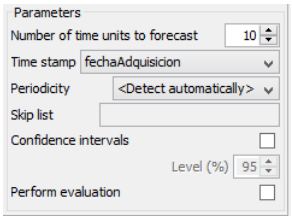


Figura 5.9: Definición de Parámetros

Despúes de establecida la conexión se procede a la selección del algoritmo (Ver figura 5.8) y la definición de los atributos en WEKA (Ver figura 5.9).

Los datos históricos a considerar para la predicción son a partir de Enero del 2000 hasta Abril del 2014. El costo de inversión a predecir será para los futuros 10 meses, es decir desde Mayo del 2014 hasta Febrero del 2015.

En la figura (5.10) se detalla los costos predecidos de inversión en la adquisición de libros por parte de la Universidad de Cuenca. En esta, la primera columna indica la fecha de predicción en el formato año y mes, mientras que la segunda y tercera columna respectivamente indica el valor predecido por cada uno de los algoritmos en la fecha indicada.

Periodo de Tiempo	Presupuesto	
	Regresión Lineal	Proceso Gaussiano
2014-05	12859.35	4823.28
2014-06	263.60	8037.73
2014-07	9081.40	9785.66
2014-08	15541.08	4128.52
2014-09	3393.00	4680.82
2014-10	11431.92	15416.38
2014-11	20736.48	9121.97
2014-12	7922.30	10646.79
2015-01	3061.48	11659.04
2015-02	11146.95	6855.52

Figura 5.10: Comparación de Resultados

La gráfica 5.11a presenta las predicciones utilizando el algoritmo de *Regresión Lineal* y la gráfica 5.11b presenta las predicciones utilizando el algoritmo del *Proceso Gaussiano*. Los datos históricos en el intervalo de tiempo mencionado previamente se representa con una línea continua y los costos predichos para los futuros 10 meses se muestran por líneas entrecortadas.

Evaluación del modelo

Una vez obtenido los resultados, el siguiente paso es la comprobación del algoritmo. Los algoritmos de Data Mining recomiendan que del 100 % de los datos totales se debe aplicar un porcentaje suficiente para preparar al algoritmo (entrenamiento) y reservar un porcentaje para realizar la evaluación del algoritmo (test).

En estos dos algoritmos utilizados para realizar las predicciones no se definen datos para el entrenamiento y el test, debido a que el método utilizado es el cross-validation que se encarga de realizar internamente el respectivo test con datos aleatorios del conjunto total.

5.4.2. Clasificación de un tipo de Usuario

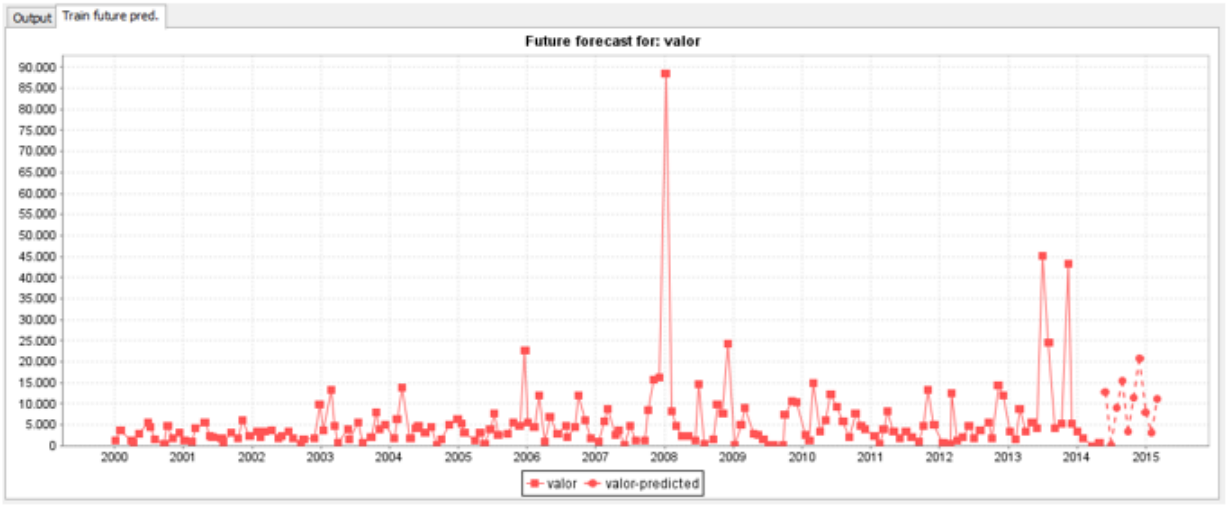
Definición del problema

Qué características tienen en común los usuarios del tipo estudiantes que devuelven o no los documentos bibliotecarios prestados por el centro de documentación en el rango de tiempo establecido.

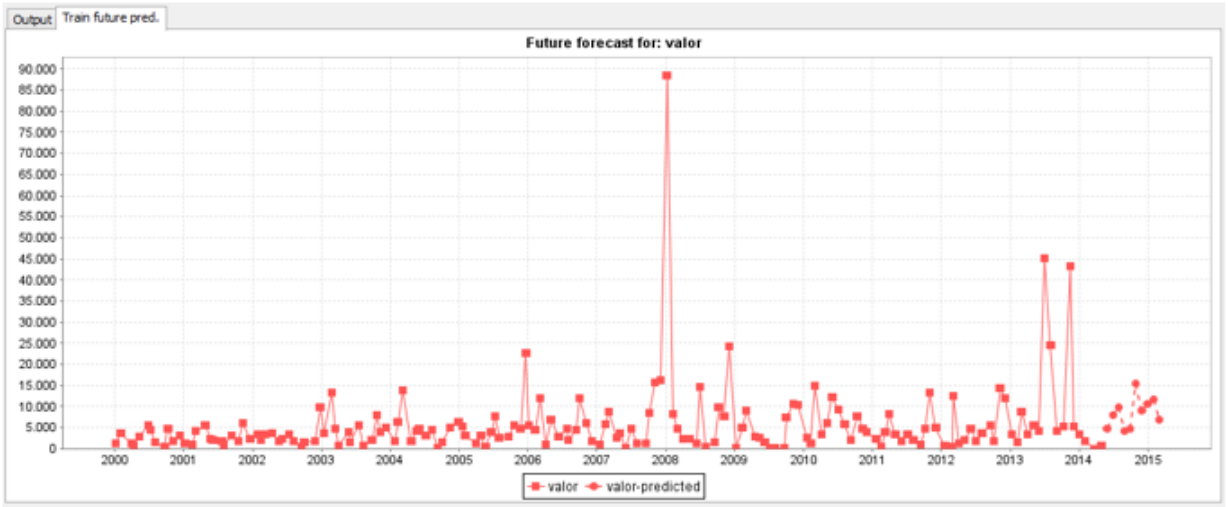
Este análisis se enfoca en clasificar un tipo de usuario, en este caso estudiante de una facultad, de una carrera determinada, con cierto puntaje académico y con un determinado ingreso económico. A este tipo de usuario se le clasificara como excelente, bueno o malo en base al tiempo promedio de devolución del documento prestado.

Elección del algoritmo

El problema planteado intenta obtener como resultado la clasificación de todos los estudiantes de la Universidad de Cuenca que han tramitado un préstamo en el centro de documentación. Este análisis intentara descubrir el patrón más adecuado que se ajuste con los datos históricos, por eso que, se utilizará el algoritmo más común en clasificación que es el de “NaiveBayes” y para la validación se utilizará el “J48” analizado previamente.



(a)



(b)

Figura 5.11: Resultado de Predicción: (a) Algoritmo de Regresión Lineal, (b) Proceso Gaussiano

.

Obtención de Datos

Para la obtención de los datos se define la consulta SQL que se muestra en la figura 5.22, en la cual los datos son extraídos de la base de datos "datawarehouse" del modelo multidimensional préstamos (Ver figura 3.53), socioeconómica (Ver figura 5.14) y académico (Ver figura 5.15).

```
select tbl1.ingresos, aca.PROMEDIO, aca.carrera, aca.facultad, pres.horas_prestamo
from (select se.id_socioeconomica, se.id_integrante, se.id_periodo_academico,
             se.id_carrera as id_carrera2, se.ingresos,per.cod_persona,
             paca.cod_periodo,paca.PERLEC_ID, se.id_carrera
    from fact_socioeconomica as se, dim_persona as per, datawarehouse.dim_periodo_academico as paca
    where se.id_integrante=per.id_persona
    and se.id_periodo_academico=paca.id_periodo_academico) tbl1 inner join academico as aca
on (aca.id_carrera=tbl1.id_carrera2
and aca.persona_id=tbl1.cod_persona
and aca.PERLEC_ID=tbl1.cod_periodo) inner join prestamo_periodo as pres
on (pres.cedula=cod_persona
and pres.cod_periodo=aca.PERLEC_ID)
```

Figura 5.12: SQL de extracción de datos

Como se observa en la consulta, existen cinco variables a analizar:

Variable	Esquema	Descripción
Facultad	Académico /Socioeconómica	Variable que almacena la facultad al que pertenece un usuario determinado.
Carrera	Académico /Socioeconómica	Variable que almacena la carrera al que pertenece un usuario determinado.
Puntaje académico	Académico	Variable que almacena la equivalencia de la calificación académica del usuario en cada uno de los periodos lectivos que se halle matriculado.
Ingreso económico	Socioeconómica	Variable que almacena la equivalencia de los ingresos económicos por parte de cada uno de los integrantes de la familia en un periodo académico que se halle matriculado.
Tiempo de Devolución	Préstamos	Variable que almacena el intervalo de tiempo entre el préstamo y la devolución de un material bibliográfico.

Figura 5.13: Análisis de variables

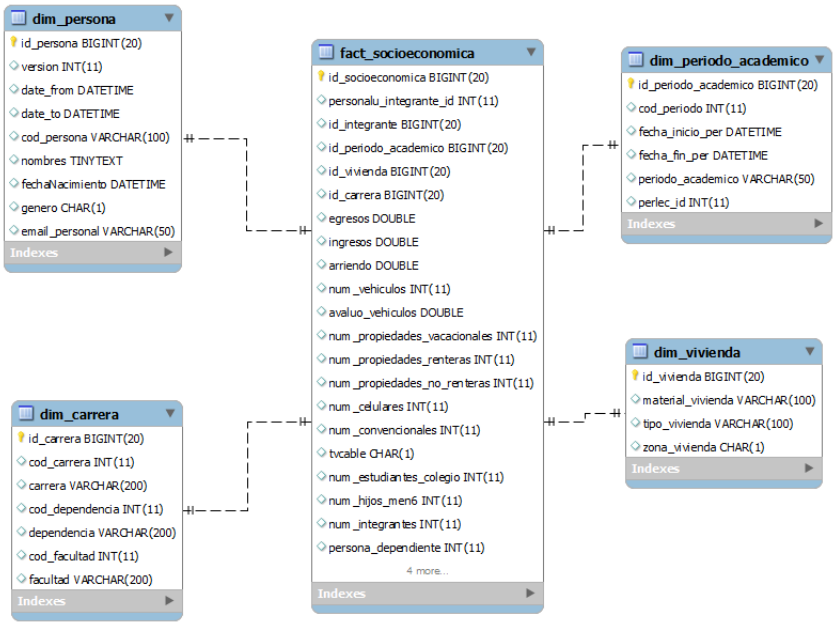


Figura 5.14: Modelo multidimencional: Socioecónomica

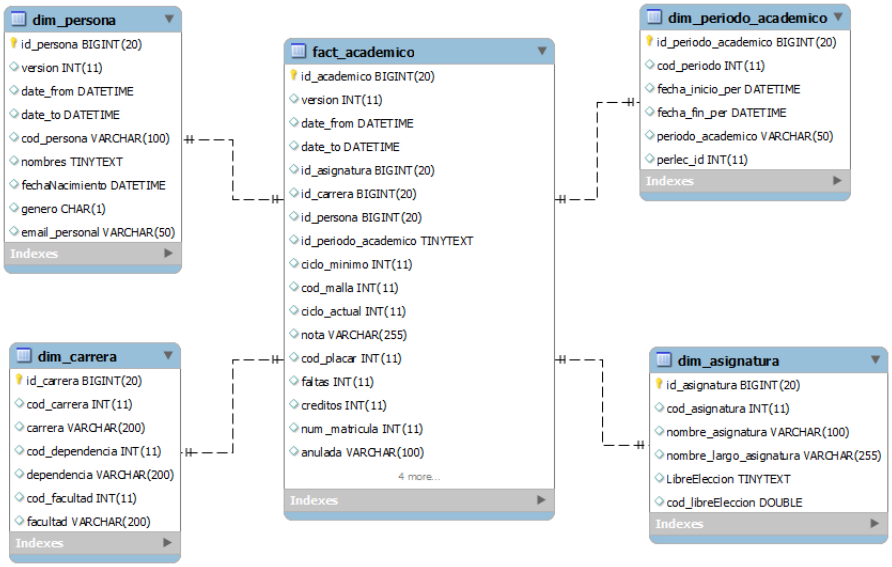


Figura 5.15: Modelo multidimencional: Académico

Minería de datos

Características principales del modelo:

- **Estructura a la que pertenece:** Modelo multidimensional académico, socieconómica y préstamos.
- **Algoritmo a utilizar:** NaiveBayes y J48.
- **Objetivo del modelo:** Clasificar al usuario estudiante del centro de documentación en base al tiempo de devolución del material bibliotecario pedido.
- **Atributos a predecir:** Facultad, carrera, puntaje académico, ingreso económico, tiempo de devolución.

Aplicación de Algoritmo

Una vez identificado los atributos y después de tener codificada la consulta SQL se procede a la aplicación y ejecución del algoritmo seleccionado.

El primer paso a realizar es la conexión de WEKA con la base de datos, este procedimiento se ve en la figura 5.16

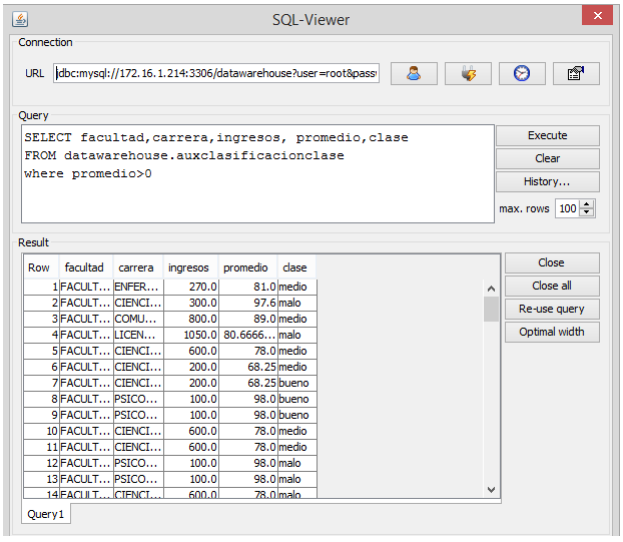


Figura 5.16: Conexión de WEKA con la base de datos

Una vez establecida la conexión se procede con la selección del algoritmo (Ver figura 5.17) y la definición de los respectivos atributos.

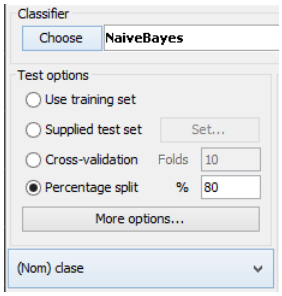


Figura 5.17: Selección de Algoritmo

Los datos históricos a considerar para la clasificación son a partir del periodo lecctivo *MARZO2010-AGOSTO2010* hasta el periodo *MARZO2014-AGOSTO2014*. Además, el tiempo de préstamo obtenido por la diferencia de la *fecha de devolución* y la *fecha de préstamo*, la misma que se mapea de una variable cuantitativa a una variable cualitativa (Ver figura 5.18). El mapeo se realiza considerando el tiempo máximo de devolución definida en el Centro de Documentación "Juan Bautista Vázquez".

Lower Bound	Upper Bound	Value
	72.0	bueno
72.0	120.0	medio
120.0		malo

Figura 5.18: Mapeo a variable cualitativa

.

La figura 5.19 presenta la clasificación utilizando el algoritmo de *NaiveBayes*. En la misma se puede apreciar la clasificación de una facultad frente a una clase de usuario, es decir, se categoriza a una facultad según el historial del tiempo de devolucción del material bibliográfico por parte de los estudiantes, los mismos que pueden ser malo, medio o bueno.

En la figura 5.20 se puede apreciar una clasificación más exhaustiva, esta categoriza a una carrera según el tiempo de devolución.

Evaluación del modelo

Para la evaluación del algoritmo mencionado previamente, se utiliza el algoritmo *J48*, de la misma manera se establece la conexión desde WEKA y la definición de parámetros respectivos. Como indica la figura 5.17, se elige el *método Percentage*

Classifier output

Attribute	Class medio (0.28)	malo (0.29)	bueno (0.43)
facultad			
FACULTAD DE CIENCIAS MÉDICAS	2867.0	1922.0	5240.0
FACULTAD DE FILOSOFÍA, LETRAS Y CIENCIAS DE LA EDUCACIÓN	1573.0	2067.0	2295.0
FACULTAD DE CIENCIAS ECONÓMICAS Y ADMINISTRATIVAS	2427.0	2186.0	3734.0
FACULTAD DE JURISPRUDENCIA	310.0	322.0	348.0
FACULTAD DE PSICOLOGÍA	487.0	592.0	724.0
FACULTAD DE CIENCIAS DE LA HOSPITALIDAD	212.0	252.0	261.0
FACULTAD DE ARTES	220.0	518.0	275.0
FACULTAD DE INGENIERÍA	1451.0	1756.0	1985.0
FACULTAD DE ARQUITECTURA Y URBANISMO	1296.0	1968.0	1873.0
FACULTAD DE CIENCIAS QUÍMICAS	809.0	905.0	1184.0
FACULTAD DE ODONTOLOGÍA	399.0	362.0	801.0
FACULTAD DE CIENCIAS AGROPECUARIAS	89.0	108.0	181.0

Figura 5.19: Resultado de Clasificación: Algoritmo de NaiveBayes, reporte facultad

.

Classifier output

Attribute	Class medio (0.28)	malo (0.29)	bueno (0.43)
carrera			
ENFERMERIA	268.0	122.0	416.0
CIENCIAS HUMANAS GESTION PARA EL DESARROLLO CULTURAL	5.0	6.0	7.0
COMUNES - ECONOMIA	1363.0	1137.0	2187.0
LICENCIATURA EN GENERO Y DESARROLLO	21.0	33.0	6.0
CIENCIAS DE LA EDUCACION EN LA ESPECIALIZACION DE LENGUA Y LITERATURA INGLESA	269.0	441.0	381.0
PSICOLOGIA EDUCATIVA EN EDUCACION BASICA	119.0	156.0	167.0
TURISMO	119.0	101.0	169.0
EDUCACION GENERAL BASICA	169.0	132.0	199.0
CIENCIAS DE LA EDUCACION EN LA ESPECIALIZACION DE LENGUA LITERATURA Y LENGUAJES AUDIOVISUALES	162.0	422.0	294.0
CIENCIAS DE LA COMUNICACIÓN SOCIAL MENCIÓN EN PUBLICIDAD Y RELACIONES PÚBLICAS	17.0	15.0	43.0
DISENO DE INTERIORES	33.0	127.0	51.0
NUTRICION Y DIETETICA	29.0	29.0	53.0
INGENIERIA ELECTRICA	332.0	384.0	449.0
INSTRUCCION MUSICAL	2.0	14.0	6.0
ARQUITECTURA	1296.0	1968.0	1873.0
MEDICINA Y CIRUGIA	2336.0	1595.0	4262.0
CONTABILIDAD Y AUDITORIA	411.0	331.0	589.0
BIOQUIMICA Y FARMACIA	133.0	157.0	238.0
CIENCIAS DE LA EDUCACION EN LA ESPECIALIZACION DE FILOSOFIA SOCIOLOGIA Y ECONOMIA	171.0	219.0	225.0
ODONTOLOGIA	399.0	362.0	801.0
CIENCIAS DE LA EDUCACION EN LA ESPECIALIZACION DE CULTURA FISICA	85.0	105.0	121.0
INGENIERIA CIVIL	571.0	891.0	819.0
INGENIERIA DE SISTEMAS	283.0	192.0	367.0
COMUNES - DISEÑO	1.0	12.0	1.0
IMAGENOLOGIA	14.0	18.0	20.0
PSICOLOGIA EDUCATIVA EN ORIENTACION PROFESIONAL	123.0	165.0	181.0
INGENIERA AMBIENTAL	250.0	301.0	388.0
CIENCIAS DE LA COMUNICACIÓN SOCIAL MENCIÓN EN PERIODISMO	11.0	30.0	35.0

Figura 5.20: Resultado de Clasificación: Algoritmo de NaiveBayes, reporte carrera

.

split con un valor de 80 %, la misma que define el 80 % de datos totales para el **entrenamiento** del algoritmo y el 20 % para el **test**.

El resultado del algoritmo J48 se observa en la figura 5.21, que se analiza una carrera frente al promedio académico por parte de los estudiantes de una facultad. Este promedio está dividida en cinco categorías que van de: 1 a 20.8, 20.8 a 40.6, 40.6 a 60.4, 60.4 a 80.2 y de 80.2 a 100.

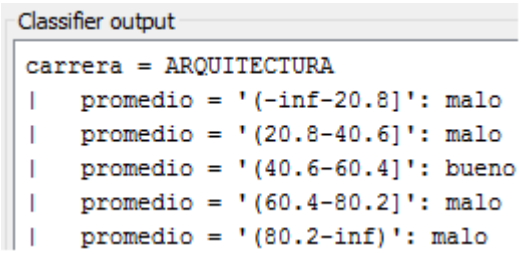


Figura 5.21: Resultado de Clasificación: Algoritmo J48

.

A simple vista se observa que para la mayor parte de categorías del promedio corresponde una clasificación malo, la misma que coincide con el resultado del algoritmo de NaiveBayes (Ver figura 5.20) para la Facultad de Arquitectura.

5.4.3. Clúster de un tipo de Usuario

Definición del problema

Que características tienen en común los usuarios del tipo estudiantes que piden prestados o documentos bibliotecarios del centro documental.

Este análisis se enfoca en agrupar o clusterizar a un tipo de usuario, en este caso estudiante, con cierto puntaje académico y con un determinado ingreso económico. A este tipo de usuario se clusteriza por el tiempo promedio de devolución del documento prestado.

Elección del algoritmo

El problema planteado intenta obtener como resultado patrones de clasificación de todos los estudiantes de la Universidad de Cuenca que han tramitado un préstamo en el centro de documentación, ajustandose a los datos históricos. Para el cual se

utilizará el algoritmo más común en clusterización que es el de "Canopy". Este algoritmo permite una mejor interpretación de los resultados obtenidos ofreciendo resultados sencillos de interpretar y sobre todo el tiempo de procesamiento de datos es muy corto a comparación de otros algoritmos de clustering.

Obtención de Datos

Para la obtención de los datos se plantea la consulta SQL que se muestra en la figura 5.22, en la cual los datos son extraídos de la base de datos "datawarehouse" del esquema adquisición préstamos (Ver figura 3.53), socioeconómica (Ver figura 5.14) y académico (Ver figura 5.15).

```
select tbl1.ingresos, aca.PROMEDIO, aca.carrera, aca.facultad, pres.horas_prestamo
from (select se.id_socioeconomica, se.id_integrante, se.id_periodo_academico,
             se.id_carrera as id_carrera2, se.ingresos,per.cod_persona,
             paca.cod_periodo,paca.PERLEC_ID, se.id_carrera
    from fact_socioeconomica as se, dim_persona as per, datawarehouse.dim_periodo_academico as paca
    where se.id_integrante=per.id_persona
    and se.id_periodo_academico=paca.id_periodo_academico) tbl1 inner join academico as aca
on (aca.id_carrera=tbl1.id_carrera2
and aca.persona_id=tbl1.cod_persona
and aca.PERLEC_ID=tbl1.cod_periodo) inner join prestamo_periodo as pres
on (pres.cedula=cod_persona
and pres.cod_periodo=aca.PERLEC_ID)
```

Figura 5.22: SQL de extracción de datos

Se tiene cinco variables a analizar:

Variable	Esquema	Descripción
Puntaje académico	Académico	Variable que almacena la equivalencia de la calificación académica del usuario en cada uno de los periodos lectivos que se halle matriculado.
Ingreso económico	Socioeconómica	Variable que almacena la equivalencia de los ingresos económicos por parte de cada uno de los integrantes de la familia en un periodo académico que se halle matriculado.
Tiempo de Devolución	Préstamos	Variable que almacena el intervalo de tiempo entre el préstamo y la devolución de un material bibliográfico.

Figura 5.23: Análisis de variables

Minería de datos

Características principales del modelo:

- **Estructura a la que pertenece:** Esquemas académico, socieconomica y prestamos.
- **Algoritmo a utilizar:** Canopy.
- **Objetivo del modelo:** Obtener como resultado patrones de clasificación de todos los estudiantes de la Universidad de Cuenca que han tramitado un préstamo en el centro documental.
- **Atributos:** Puntaje académico, ingreso económico, tiempo de devolución.

Aplicación de Algoritmo

Una vez identificado los atributos y después de codificar la consulta SQL se procede a la aplicación y ejecución del algoritmo seleccionado.

El primer paso a realizar es la conexión de WEKA con la base de datos, la que se ve en la figura 5.16

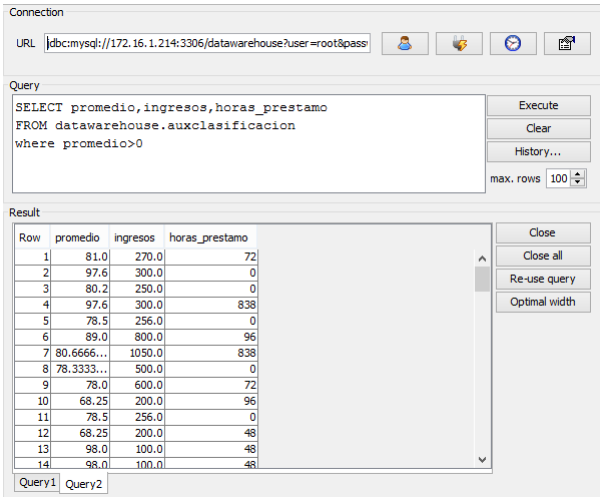


Figura 5.24: Conexión de WEKA con la base de datos

Una vez establecida la conexión se procede a la selección del algoritmo (Ver figura 5.25) y la definición de los respectivos atributos.

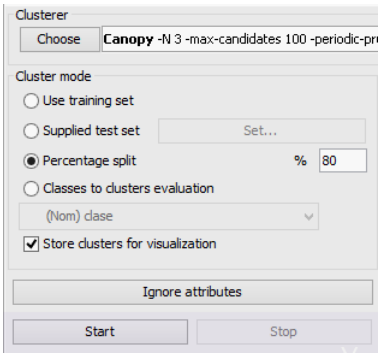


Figura 5.25: Selección de Algoritmo

Los datos históricos a considerar para este ejemplo son a partir del periodo lecctivo *MARZO2010-AGOSTO2010* hasta el periodo *MARZO2014-AGOSTO2014*. Además, el tiempo de préstamo obtenido por la diferencia de la *fecha de devolución* y la *fecha de préstamo*.

La figura 5.26 presenta los numeros de cluster generados utilizando el algoritmo de *Capony*. En esta figura se puede apreciar los tres cluster generados de un tipo de usuario.

En la figura 5.27 se puede apreciar una de manera detallada los tres cluster generados donde se cruza los préstamos con el promedio académico de los estudiantes.

Se puede observar que los grupos cluster 0 (azul) y cluster 2 (verde) presentan mejores promedios académicos en un periodo lectivo a comparación del cluster 1 (rojo), además el cluster verde son los que mas tardan en realizar las devoluciones del material pedido.

Evaluación del modelo

Para la evaluación del algoritmo como indica la figura 5.25, se elige el *método Percentage split* con un valor de 80 %, la misma que define el 80 % de datos totales para el **entrenamiento** del algoritmo y el 20 % para el **test**.

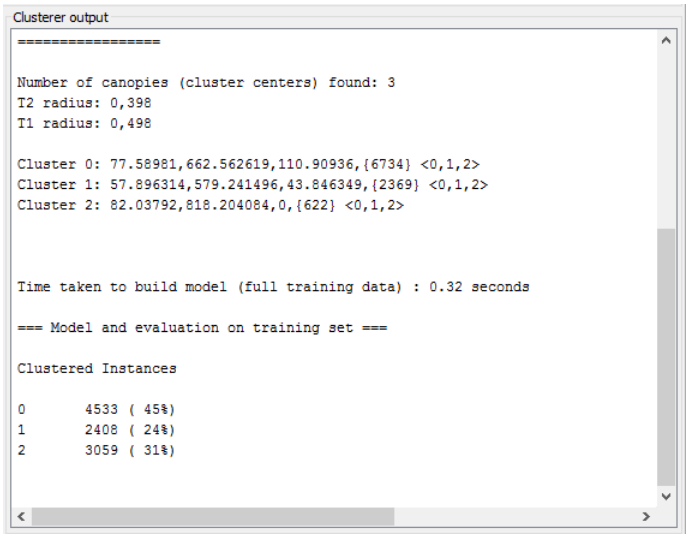


Figura 5.26: Resultado de Cluster: Algoritmo de Capony

.

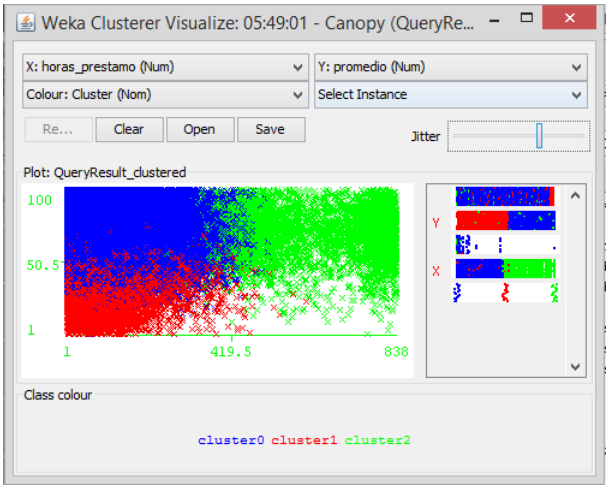


Figura 5.27: Resultado de Cluster: Algoritmo de Capony

.

5.5. Conclusiones

Al finalizar este capítulo se aprecia la utilidad de las diferentes técnicas de clasificación, predicción y de clustering aplicadas al centro de documentación, que basándose en Bibliomining permiten descubrir información desconocida para los encargados y directivos de dicha organización.

La necesidad de tener un conocimiento profundo de los datos y descubrir reglas de negocio que ayuden a una adecuada toma de decisiones, ha generado que se inicie en la investigación de Bibliomining en el centro de documentación "Juan Bautista Vázquez". Dichas investigaciones futuras se desarrollaran en base de este proyecto de titulación, ya que la base de datos creada será el punto de partida para la obtención de datos.

Capítulo 6

Conclusiones y Recomendaciones

El objetivo de este capítulo es exponer los resultados del proyecto realizado, desde un enfoque global. Se detallan las conclusiones de los resultados obtenidos y las recomendaciones que se debería considerar para la realización de un trabajo como éste, así como también se mencionan recomendaciones para el "Centro de Documentación Juan Bautista Vázquez".

6.1. Conclusiones

En este trabajo se demuestra la gran utilidad de realizar un análisis sobre los datos y el beneficio que se tiene al integrar la información y tener una visión más general. Las conclusiones generadas al finalizar el proyecto de tesis se exponen a continuación:

1. Hefesto es una metodología que proporciona los pasos necesarios para la implementación de un Data Warehouse de forma eficiente, existe mucha documentación al respecto. Además, se considera importante el hecho de que se parte de los requerimientos de los usuarios permitiendo así que los resultados persigan los objetivos del centro documental.

2. La suite de Pentaho facilita herramientas necesarias para la implementación de un Data Warehouse y ayuda en cada etapa de la creación. Cabe recalcar que el tiempo de carga inicial de los datos al ejecutar lo procesos ETL son muy altos,

cuando se desarrolle un Data Warehouse se debe considerar no cargar todos los datos a la vez, una solución es cargar los datos de acuerdo a un rango de tiempo para tratar de reducir el tiempo de procesamiento.

3. El uso de la herramienta BI Server para realizar los reportes es muy intuitiva y permite que fácilmente los usuarios generen sus reportes de acuerdo a las necesidades que aparezcan.

4. Al realizar un análisis de los datos almacenados en las diferentes fuentes de información se llego a descubrir una serie de inconsistencias que se estaban llevando a cabo en la catalogación de los libros lo cual permite reconocer que el datawarehouse no solo beneficia en la integración de la información sino que también permite *madurar los procesos* que se llevan a cabo diariamente.

5. Los procesos de reservas de libros, préstamos interbibliotecarios, manejo de multas no son llevadas a cabo en el centro de documentación "Juan Bautista Vázquez" pero son procesos que pueden implementarse en un futuro ya que son muy comunes en las bibliotecas, razón por la cual en la presente tesis de titulación se considero crear ciertas dimensiones y atributos básicos a tener presentes para estos procesos, lo que permite que el Data Warehouse implementado sea escalable.

6. Actualmente se dispone de tanta información que parece complicado sacarle provecho, el Data Warehouse que integra la información disponible permite ayudar de esta manera como soporte en la toma de decisiones. Sin embargo hay que considerar que su beneficio incrementa cuando se aplican sobre éstos técnicas de Data Mining.

7. El éxito de Data Mining depende del valor de la información que se disponga, una de las ventajas es que se puede aprovechar de la gran cantidad de datos para que nueva información sea descubierta con el uso de técnicas de Bibliomining que permite potenciar el ahorro de dinero, proporcionar programas más apropiados, resolver más de las necesidades de información del usuario, observar problemas de su colección y sirve como fuente de información más eficaz de sus usuarios.

6.2. Recomendaciones

Al terminar este proyecto de tesis se identifico ciertos procesos que pueden mejorarse en base a las siguientes recomendaciones:

1. El sistema ABCD que utiliza una base de datos ISISDB es un sistema del cual ya no hay nuevas actualizaciones desde hace algunos años y existe información limitada en cuanto al funcionamiento y conexión con los datos desde un enlace externo.

   El sistema no maneja validación de datos lo cual genera varios errores al momento de catalogar un libro: los datos son ingresados en ocaciones con mayúsculas, minúsculas, con tilde, sin tilde lo que ocaciona que no se pueda realizar búsquedas efectivas. Además, no se tiene un formato que determine exactamente como ingresar la información en cada campo y esto ocasiona que no se conozca los libros de un determinado autor, o de un determinado título porque se puede hacer referencia al mismo autor pero este tiene diferente descripción de la información ingresada.

   Razón por la cual se recomienda definir estándares de ingreso de la información para no seguir generando estos ingresos erróneos de información que no permiten realizar una exploración o análisis de datos con facilidad.

   Como solución a este problema se ha considerado cambiarse de sistema, lo cual solucionaría los problemas de ingreso de información desde que sea puesto en producción pero hay que considerar que para realizar una migración de datos es necesario realizar un limpieza de los mismos.

   Al desarrollar un componente java que permita exportar una base de datos ISISDB a un archivo con formato MARC21 y de extensión ".mrc" se llego a descubrir gran cantidad de errores en cuanto a los datos almacenados. En el Data Warehouse creado se ha pasado gran información despues de realizar una limpieza de datos de los campos más críticos que se necesitan para el análisis y la ayuda a la toma de decisiones. A pesar de esto, debe dedicarse más tiempo a la limpieza de datos con alguna persona experta en el proceso de catalogación que pueda ayudar a identificar la información correcta y así poder

realizar una migración exitosa si se llegase a cambiar el sistema ABCD por otro. Considerando que en el centro documental analizado ya se realizo un cambio de WinIsis a ABCD pero los errores no fueron identificados y se acarrearon a la base de datos actual.

2. La base de datos MySql que se utiliza para almacenar los préstamos realizados diariamente no permite un fácil enlace a los datos provenientes de la base de datos documentales IsisDB debido a que no existe un identificador único para el material bibliográfico que ayude a identificarlos individualmente.

   Los datos almacenados en IsisDB no permiten identificar el número de ejemplares que se dispone de cada material bibliográfico es por eso que se recomienda catalogar un material bibliográfico almacenando un solo identificador único en algún campo del formato MARC21 y dejar algún otro campo para identificar el número de ejemplares disponibles.

3. Tanto la base de datos de préstamos albergada en MySql como la base de datos Olympo con un motor de base de datos Oracle no guardan referencias similares para poder unir estos datos, los procesos se llevan independientemente pero seria muy recomendable que se llegue a definir un ingreso similar de la información en las dos bases de datos para que se pueda enlazar los datos y aprovecharlos de mejor manera.

4. En cuanto a lo relacionado con el proceso de evaluación de servicios se sugiere que se realice una evaluación por diferentes perspectivas ya que actualmente se recibe una calificación por parte del usuario de los servicios en el centro documental pero no se tiene claro el parámetro de evaluación, es decir es incierto conocer si evalúa la atención recibida, o las instalaciones, o el material encontrado. Por esta razón, se sugiere que se desarrolle una encuesta más detallada que ayude a determinar las fortalezas y debilidades del centro documental.

   En la presente tesis se considera el proceso de evaluación de servicios LibQual que consta de 22 preguntas y permite evaluar la satisfacción del usuario en base a 3 persectivas: el espacio físico, acceso y el personal permitiendo tener una visión más general de la opinión de los usuarios para tomar decisiones en cuanto a lo que se pueda mejorar.

Cabe recalcar que el Data Warehouse desarrollado tiene modelos para recibir datos de éstas evaluaciones, incluso da la facilidad de que puedan crearse otros ámbitos de evaluación y agregarse preguntas.

5. Durante el desarrollo del presente proyecto la mayor cantidad de tiempo se dedico al análisis de los datos, para reducir este tiempo se recomienda que algún integrante de la institución en la que se realice el Data Warehouse participe activamente si ésta persona tiene todo el conocimiento sobre la empresa y ha adquirido experiencia en los procesos que se realizan.

Finalmente, una vez concluido el trabajo se dice que el Data Warehouse realizado esta en la capacidad de ayudar a la toma de decisiones gerenciales por la cantidad de información que posee y además sirve de base para la aplicación de Data Mining y generación de consultas mediante SQL para aprovechar la gran cantidad de datos disponibles.

Capítulo 7

Anexos

7.1. Transformación de Dimensiones

7.1.1. Dimensión Carrera-Facultad-Departamento

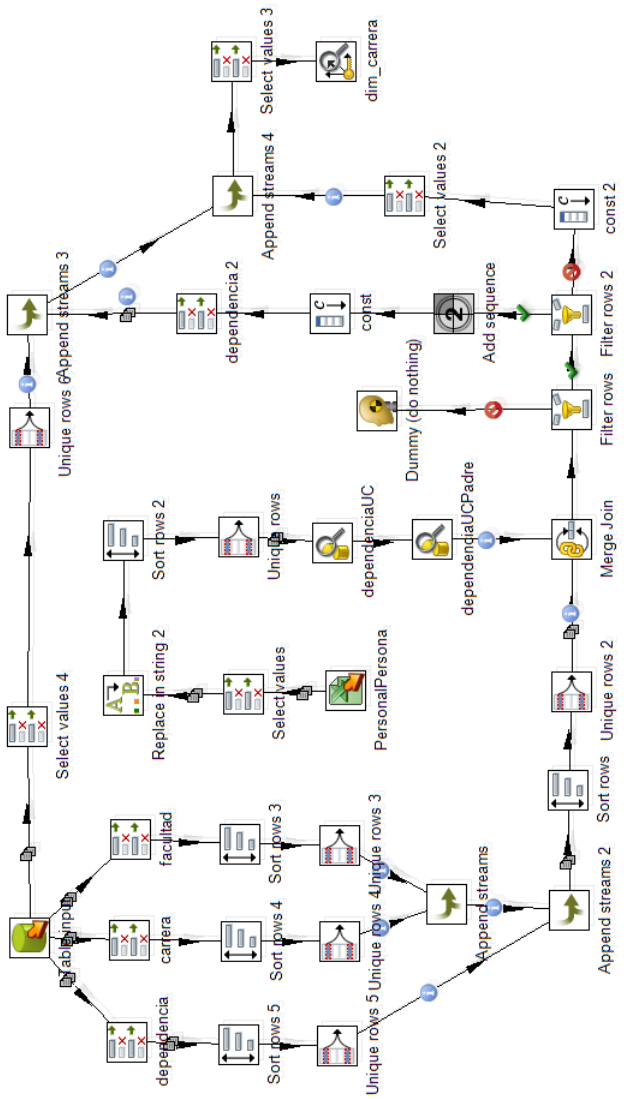


Figura 7.1: Transformación: Dimensión Carrera-Facultad-Departamento

7.1.2. Dimensión Persona

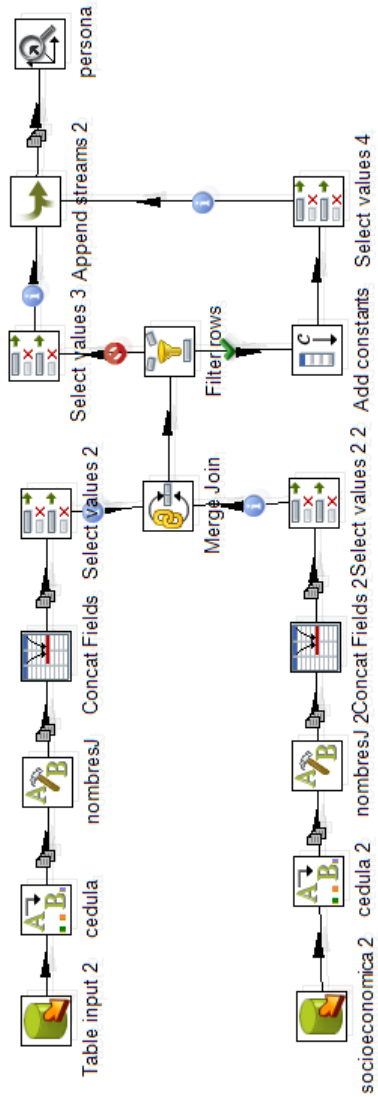


Figura 7.2: Transformación: Dimensión Persona

7.1.3. Dimensión Libro

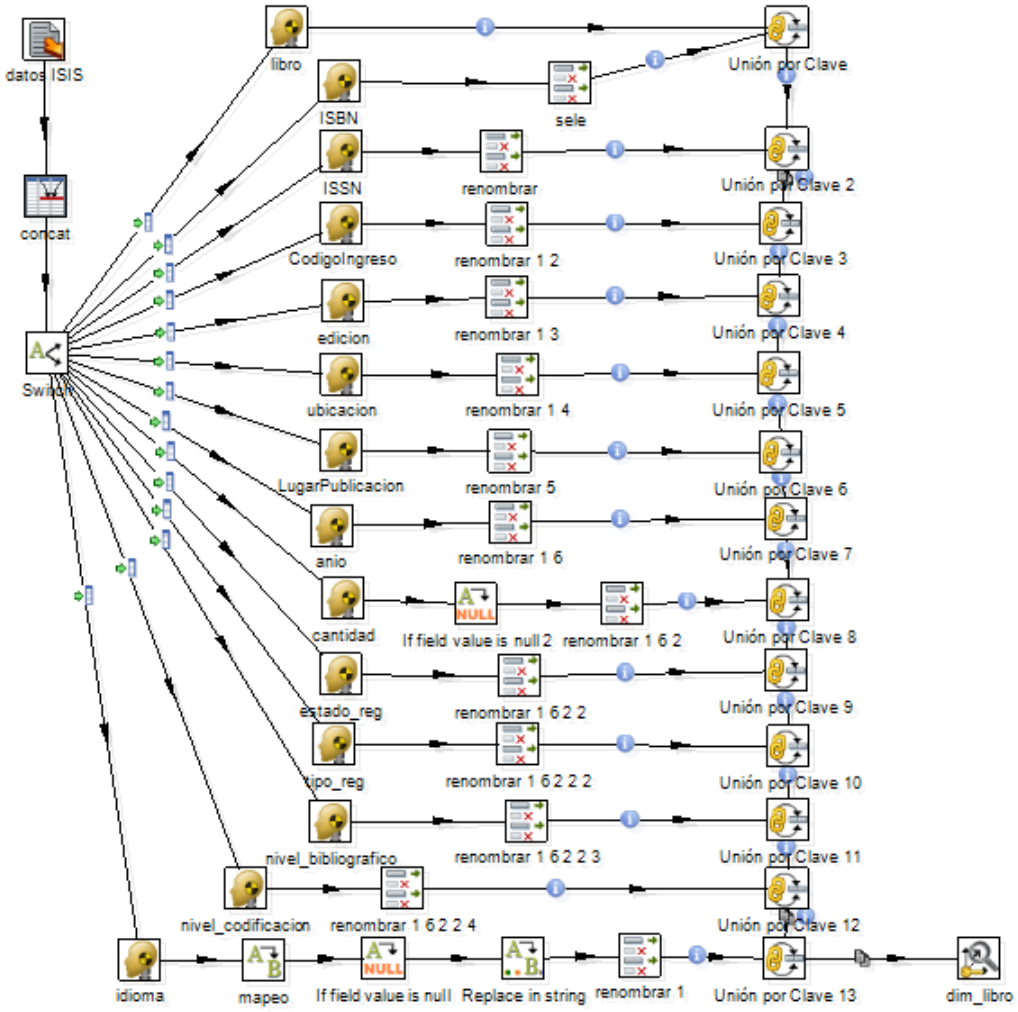


Figura 7.3: Transformación: Dimensión Libro

7.2. Proceso: Catalogación

7.2.1. Identicar preguntas

1. ¿Cuántas publicaciones fueron catalogados en una determinada categoría en una unidad de tiempo?

2. ¿Cuántos registros de catálogos fueron ingresados de un determinado autor en una unidad de tiempo?

3. ¿Cuántas publicaciones fueron catalogados por una determinada persona y su costo de operación en una unidad de tiempo?

4. ¿Cuántas publicaciones fueron eliminados por una determinada persona y su costo de operación en una unidad de tiempo?

5. ¿Cuántas publicaciones fueron modificados por una determinada persona y su costo de operación en una unidad de tiempo?

7.2.2. Modelo Lógico

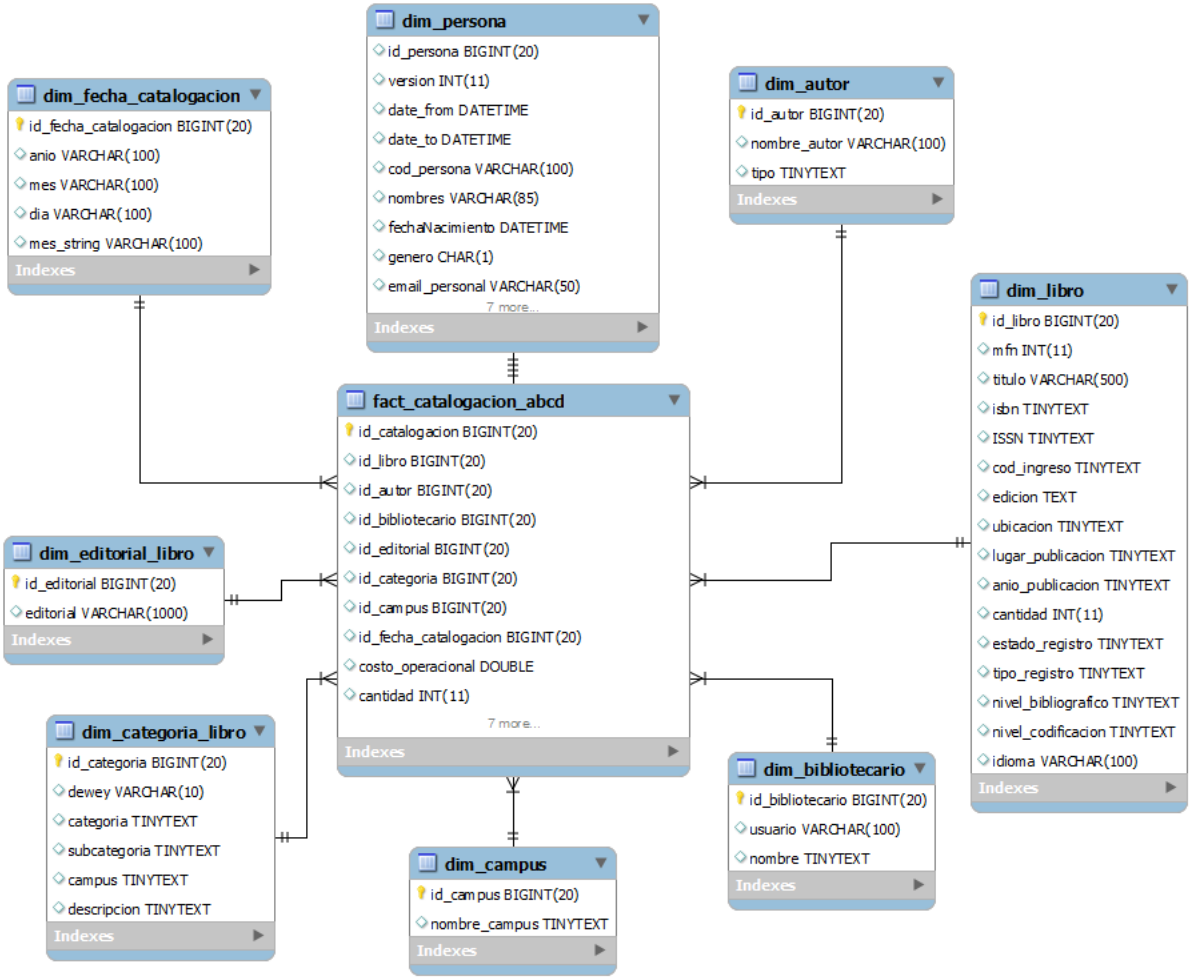


Figura 7.4: Modelo Lógico: Catalogación

7.2.3. ETL

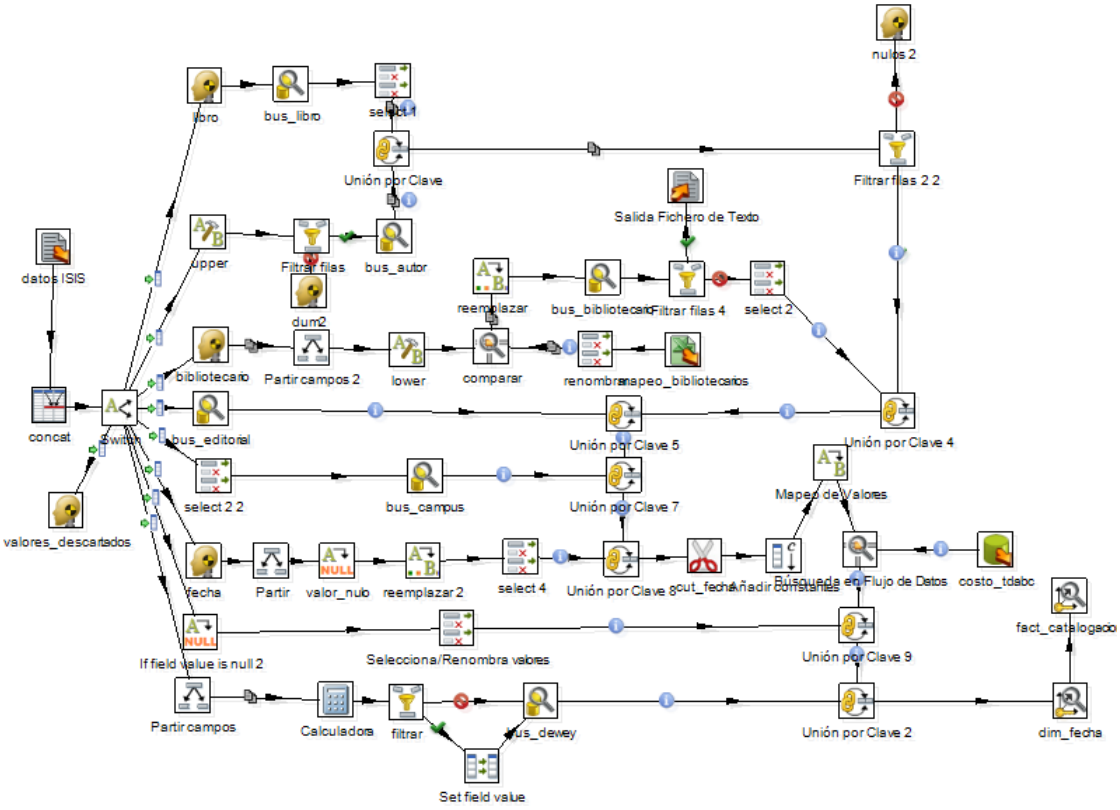


Figura 7.5: Transformación: Catalogación

7.2.4. Cubo

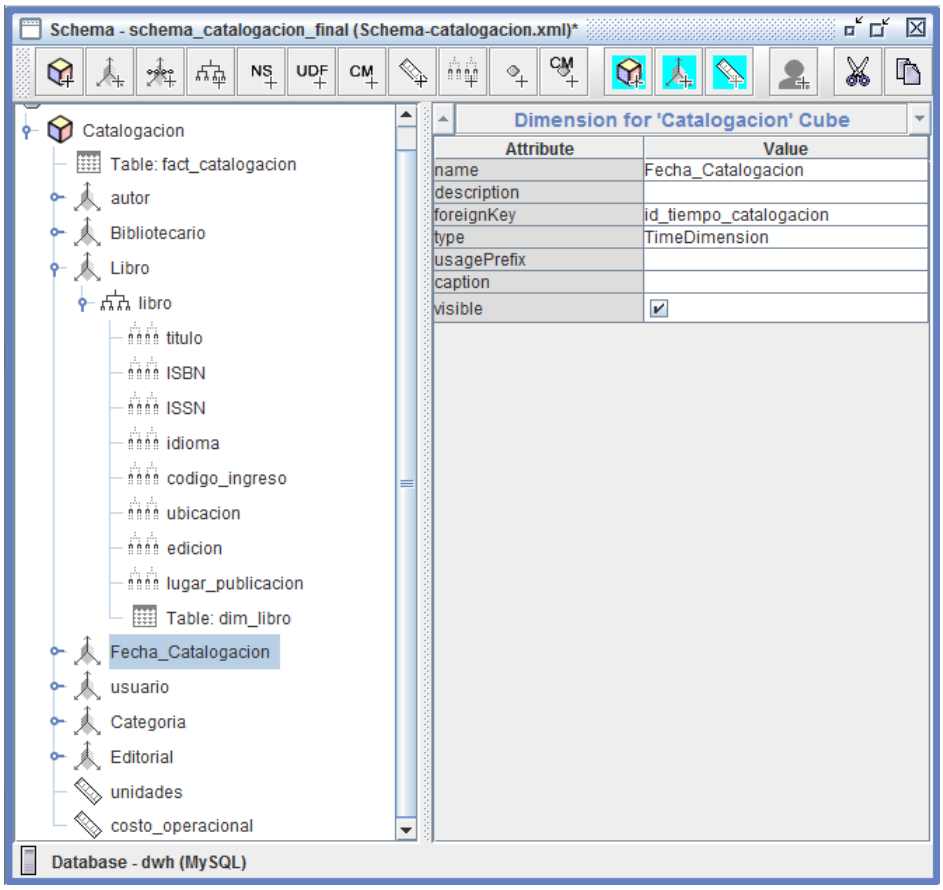


Figura 7.6: Cubo Dimensional: Catalogación

7.3. Proceso: Reserva de Material Bibliográfico

7.3.1. Identicar preguntas

1. ¿Cuantas publicaciones fueron reservadas en una determinada categoría en una unidad de tiempo?

2. ¿Cuántas publicaciones fueron reservadas por una determinada persona y su costo de operación en una unidad de tiempo?

3. ¿Cuántas publicaciones fueron reservadas de un autor determinado, de una categoría específica en una unidad de tiempo?

4. ¿Cuántas publicaciones fueron reservadas por una determinada editorial y su costo de operación en una unidad de tiempo?

7.3.2. Modelo Lógico

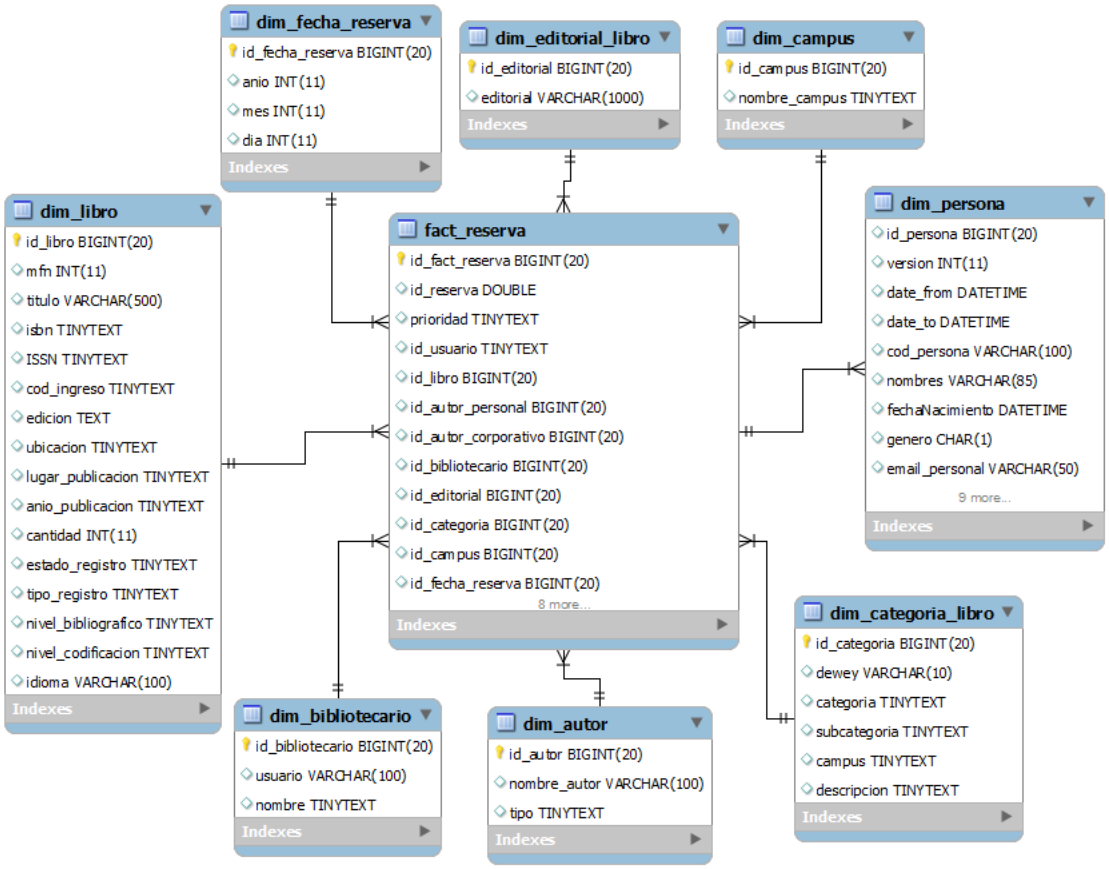
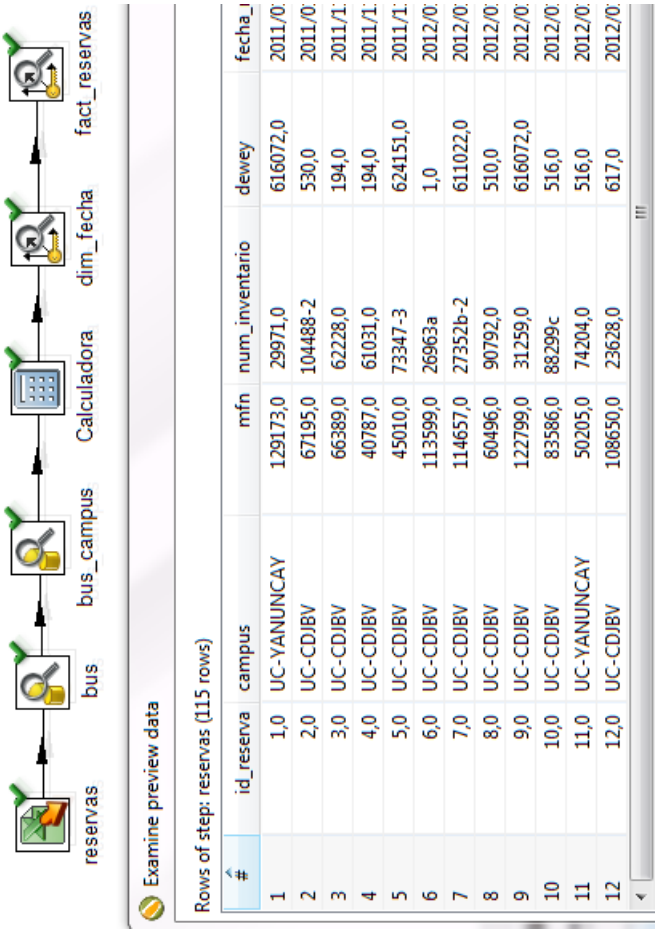


Figura 7.7: Modelo Lógico: Reservas

7.3.3. ETL



#	id_reserva	campus	mfn	num_inventario	dewey	fecha_
1	1,0	UC-YANUNCAY	129173,0	29971,0	616072,0	2011/0
2	2,0	UC-CDJBV	67195,0	104488-2	530,0	2011/0
3	3,0	UC-CDJBV	66389,0	62228,0	194,0	2011/1
4	4,0	UC-CDJBV	40787,0	61031,0	194,0	2011/1
5	5,0	UC-CDJBV	45010,0	73347-3	624151,0	2011/1
6	6,0	UC-CDJBV	113599,0	26963a	1,0	2012/0
7	7,0	UC-CDJBV	114657,0	27352b-2	611022,0	2012/0
8	8,0	UC-CDJBV	60496,0	90792,0	510,0	2012/0
9	9,0	UC-CDJBV	122799,0	31259,0	616072,0	2012/0
10	10,0	UC-CDJBV	83586,0	88299c	516,0	2012/0
11	11,0	UC-YANUNCAY	50205,0	74204,0	516,0	2012/0
12	12,0	UC-CDJBV	108650,0	23628,0	617,0	2012/0

Figura 7.8: Transformación: Reservas

7.3.4. Cubo

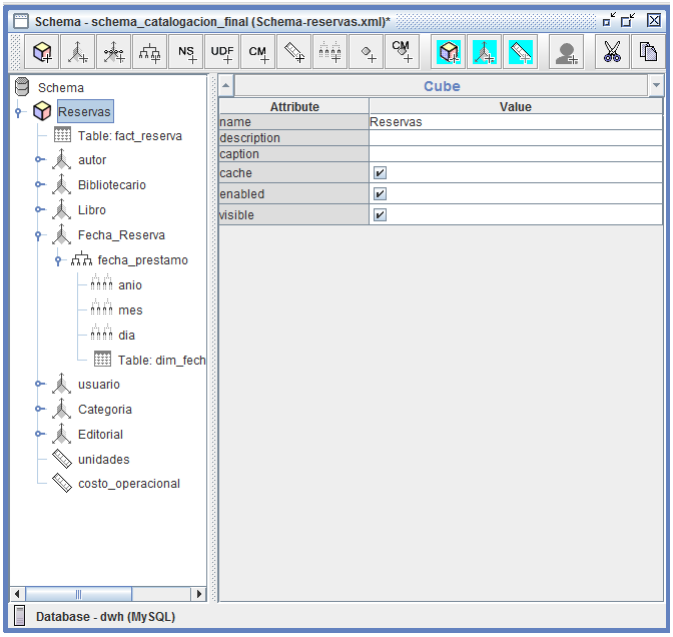


Figura 7.9: Cubo Dimensional: Reservas

7.4. Proceso: Encuestas

7.4.1. Identicar preguntas

1. ¿Cuántas encuestas se registra un determinado bibliotecario en una unidad de tiempo dada?

2. ¿Cuántas encuestas registra de una determinada terminal en una unidad de tiempo dada?

3. ¿Número de encuestas se puntuó a un servicio en una unidad de tiempo dada?

7.4.2. Modelo Lógico

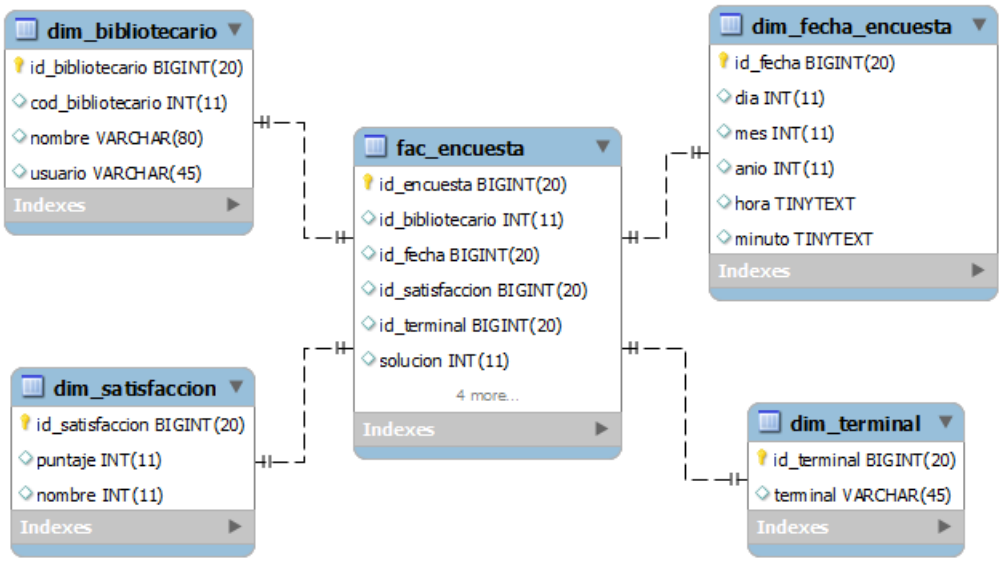


Figura 7.10: Modelo Lógico: Encuestas

7.4.3. ETL

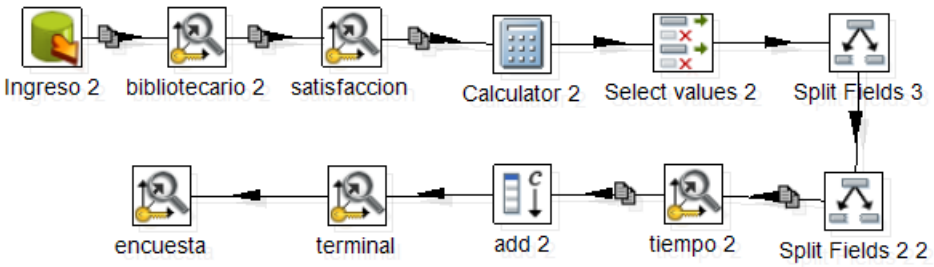


Figura 7.11: Transformación: Encuestas

7.4.4. Cubo

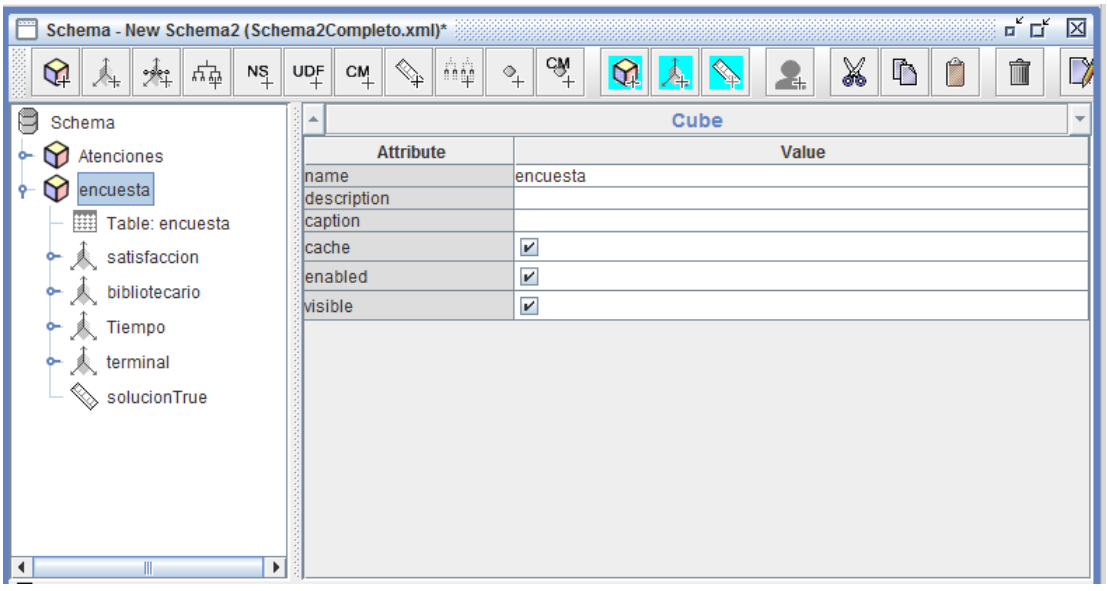


Figura 7.12: Cubo Dimensional: Encuestas

7.5. Proceso: Atención al Cliente

7.5.1. Identicar preguntas

1. ¿Cuántos soluciones registra un determinado bibliotecario en una unidad de tiempo dada?

2. ¿Cuántas inquietudes registra un determinado bibliotecario en una unidad de tiempo dada?

3. ¿Cuántas atenciones registra un determinado bibliotecario en una unidad de tiempo dada?

7.5.2. Modelo Lógico

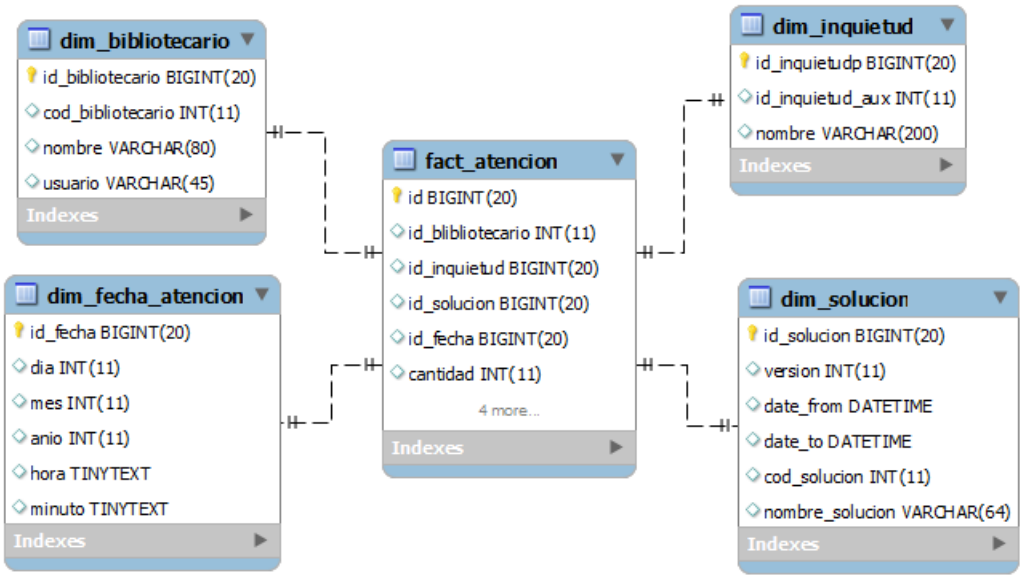


Figura 7.13: Modelo Lógico: Atención al Cliente

7.5.3. ETL

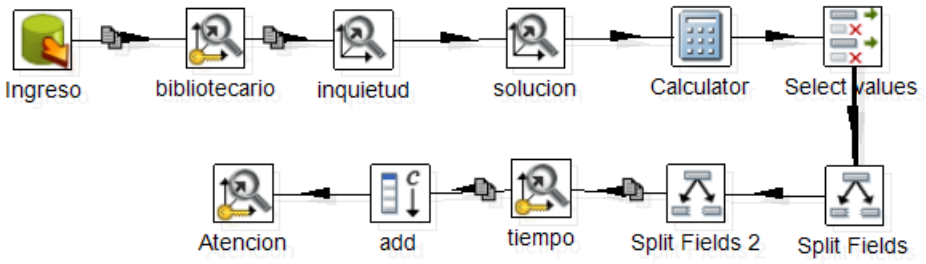


Figura 7.14: Transformación: Atención al Cliente

7.5.4. Cubo

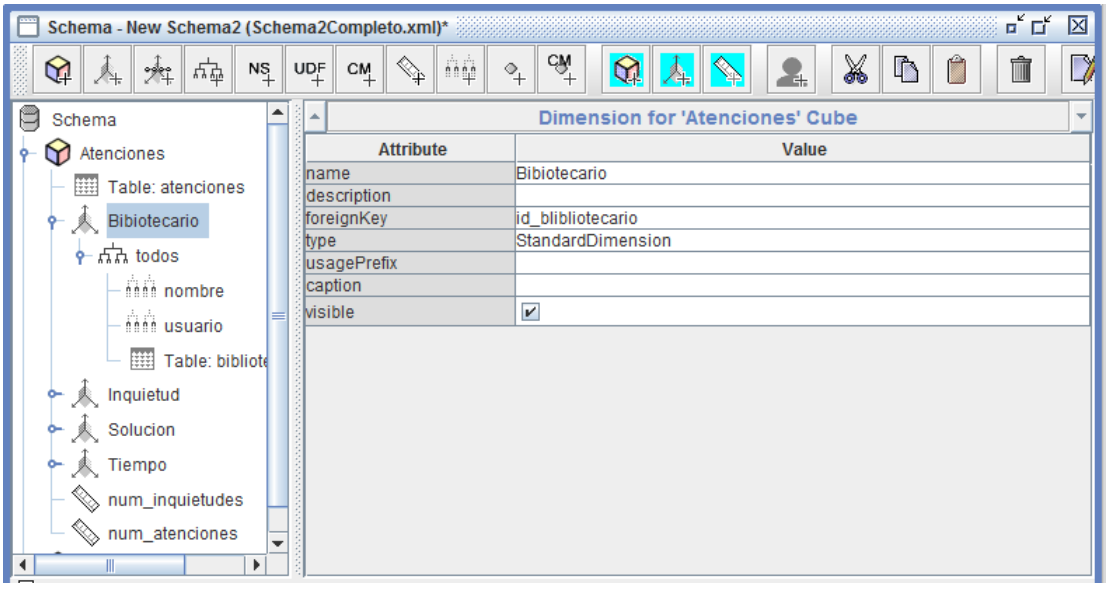


Figura 7.15: Cubo Dimensional: Atención al Cliente

7.6. Proceso: DSpace

7.6.1. Identicar preguntas

1. ¿Cuántos documentos digitales fueron catalogados por una determinada persona en una unidad de tiempo?

2. ¿Cuántos documentos digitales tiene un determinado departamento en una unidad de tiempo?

7.6.2. Modelo Lógico

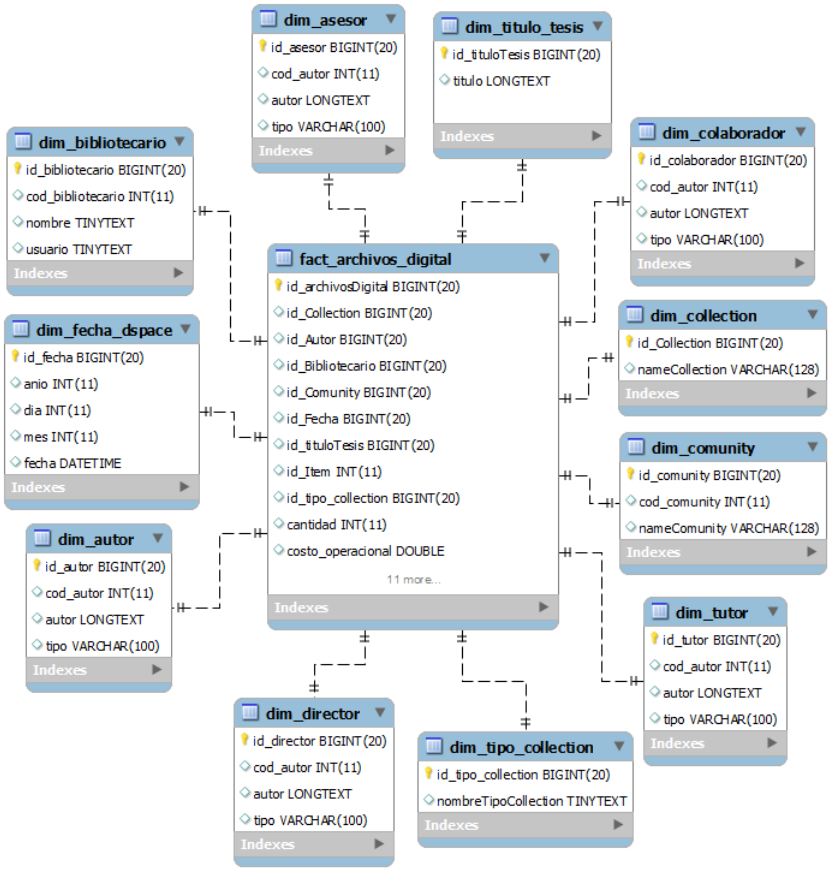


Figura 7.16: Modelo Lógico: DSpace

7.6.3. ETL

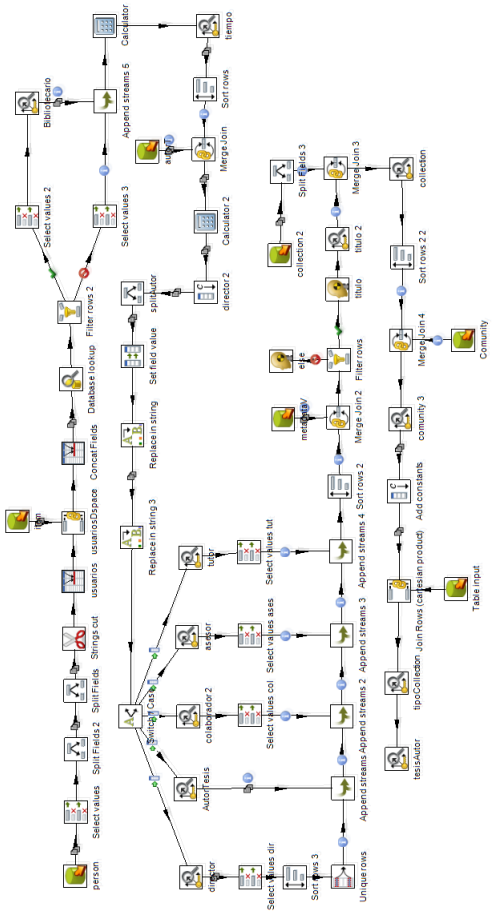


Figura 7.17: Transformación: DSpace

7.6.4. Cubo

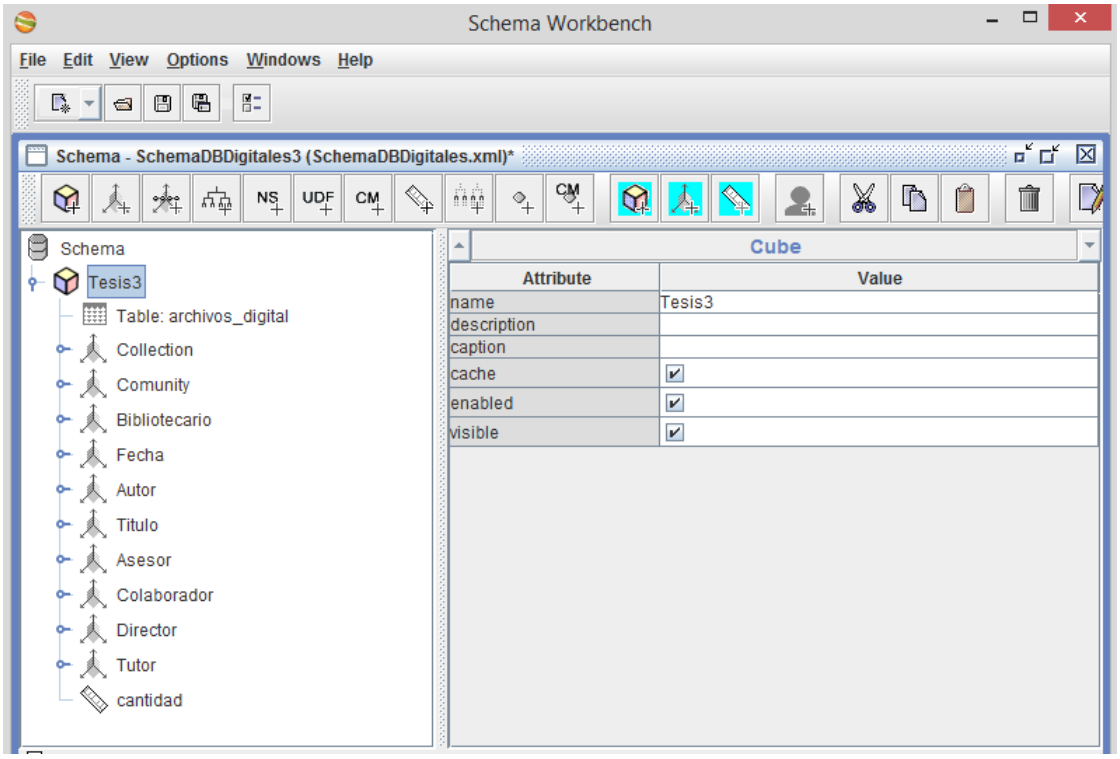


Figura 7.18: Cubo Dimensional: DSpace

7.7. Proceso: Adquisición

7.7.1. Identicar preguntas

1. ¿Cuál es el número total de adquisiciones y el costo realizadas a un proveedor específico en una unidad de tiempo dada?

2. ¿Cuál es el número total de adquisiciones realizadas de una publicación en una unidad de tiempo dada?

7.7.2. Modelo Lógico

Figura 7.19: Modelo Lógico: Adquisición

7.7.3. ETL

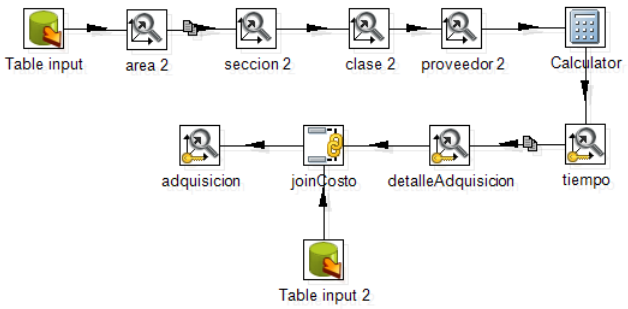


Figura 7.20: Transformación: Adquisición

7.7.4. Cubo

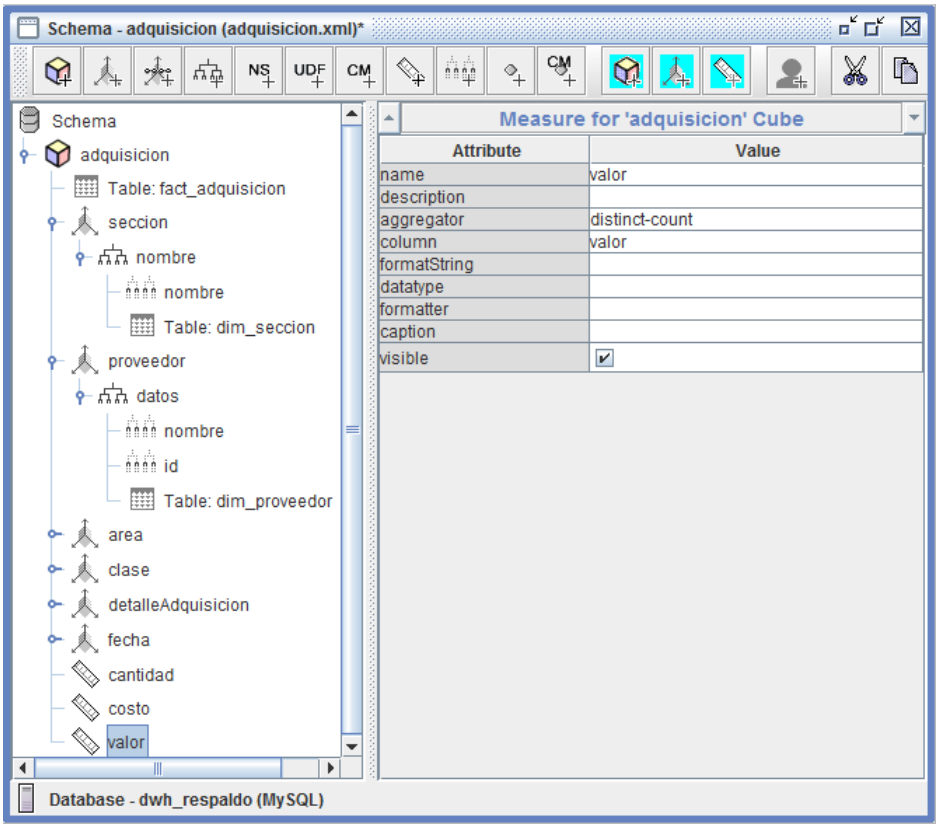


Figura 7.21: Cubo Dimensional: Adquisición

7.8. Proceso: Préstamos Interbibliotecarios

7.8.1. Modelo Lógico

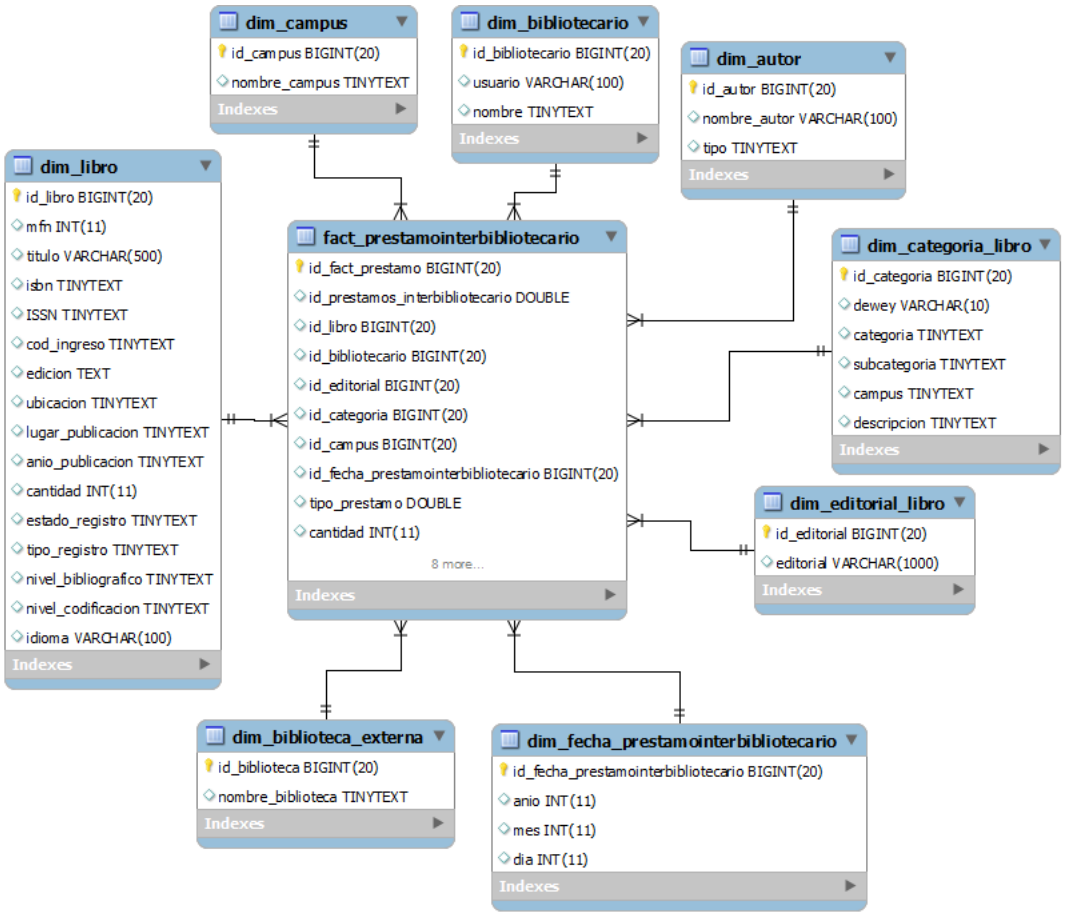


Figura 7.22: Modelo Lógico: Préstamos Interbibliotecarios

7.8.2. ETL

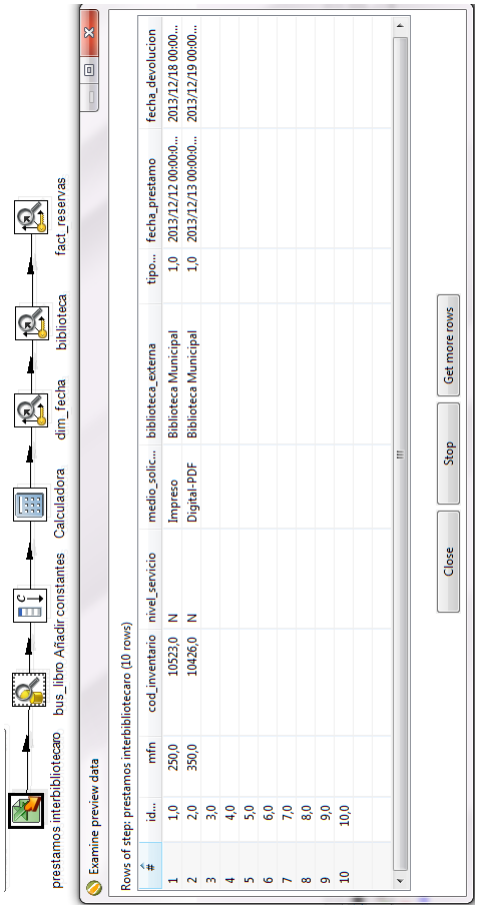


Figura 7.23: Transformación: Préstamos Interbibliotecarios

7.8.3. Cubo

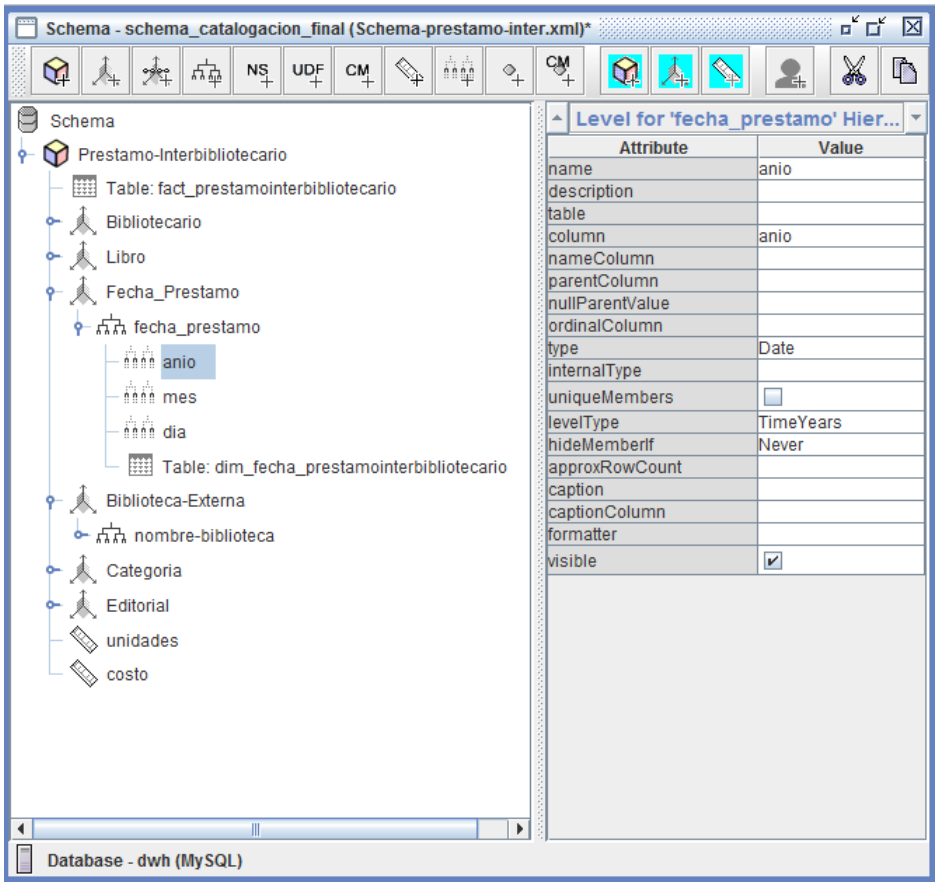


Figura 7.24: Cubo: Préstamos Interbibliotecarios

7.9. Proceso: Evaluación LOG's

7.9.1. Modelo Lógico

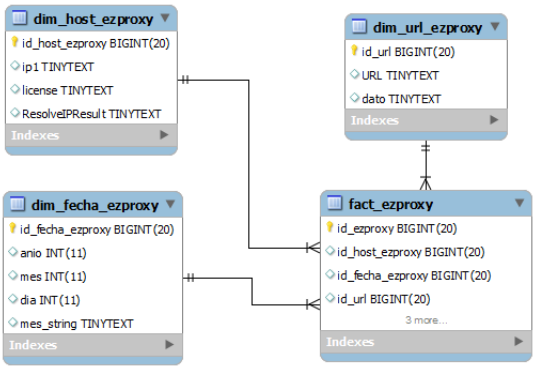


Figura 7.25: Modelo Lógico: Evaluación LOG's

7.9.2. ETL

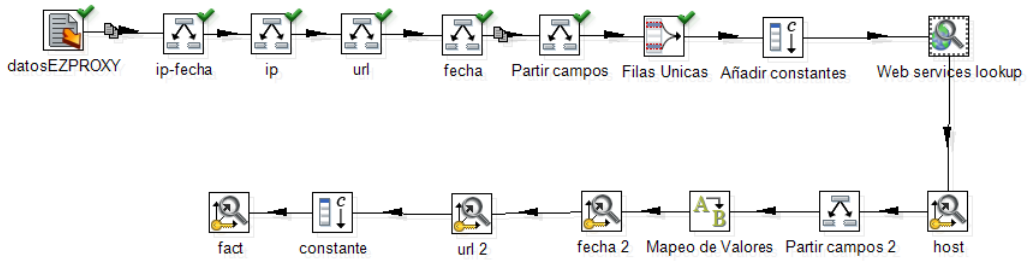


Figura 7.26: Transformación: Evaluación LOG's

7.9.3. Cubo

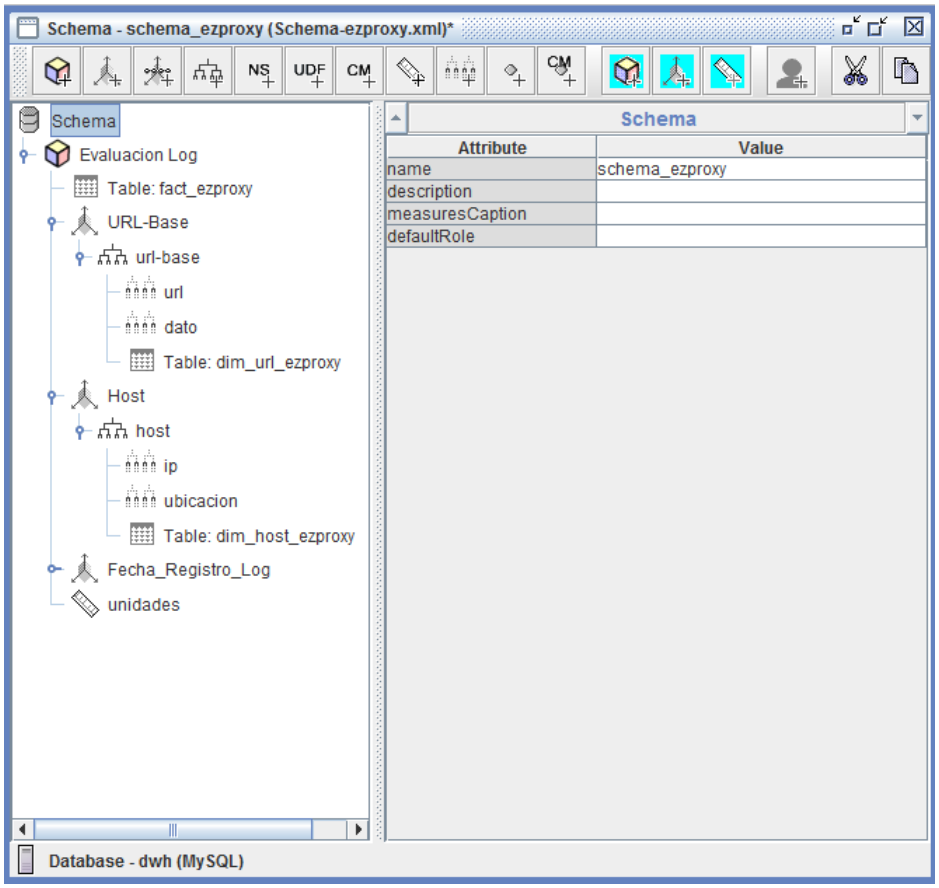


Figura 7.27: Cubo: Evaluación LOG's

7.10. Proceso: LibQual+

7.10.1. Modelo Lógico

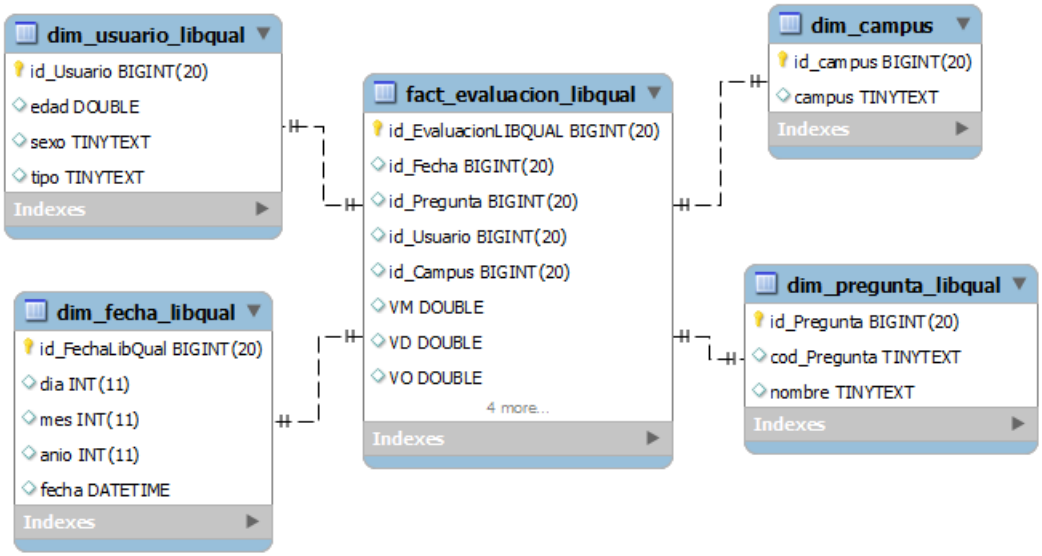


Figura 7.28: Modelo Lógico: LibQual+

7.10.2. ETL

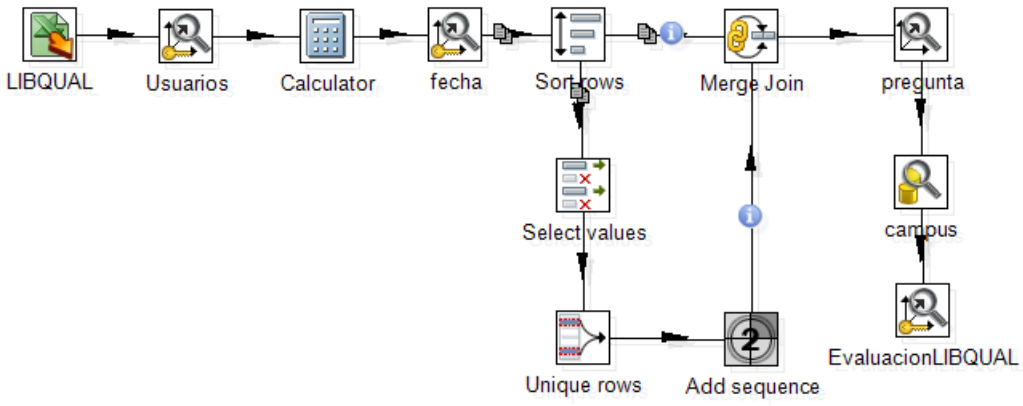


Figura 7.29: Transformación: LibQual+

7.10.3. Cubo

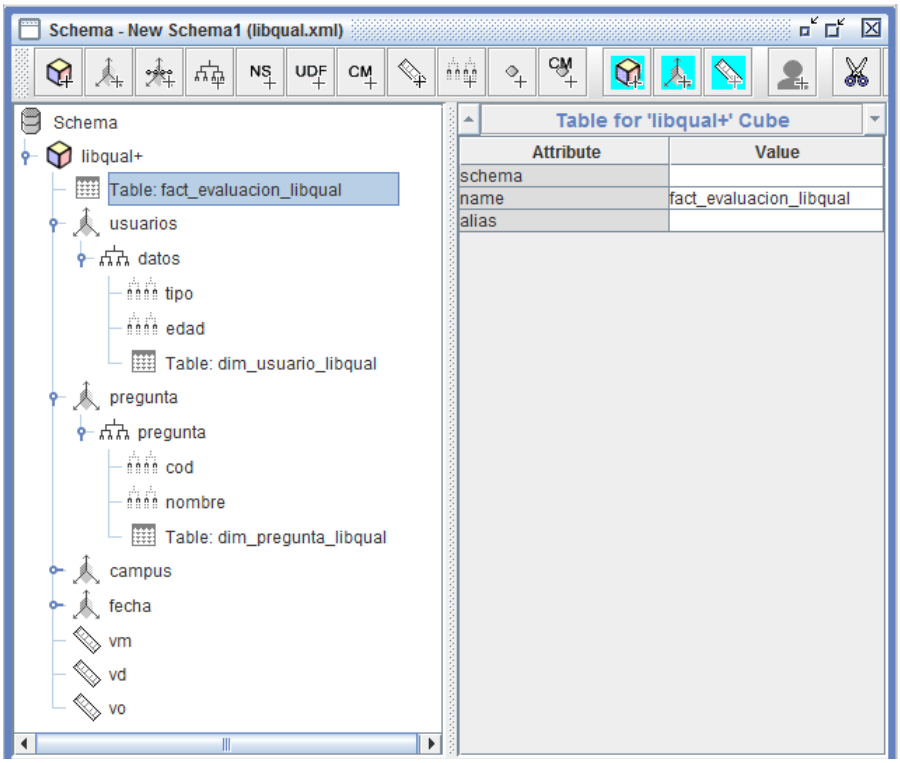


Figura 7.30: Cubo Dimensional: LibQual+

7.11. Proceso: Académico

7.11.1. Modelo Lógico

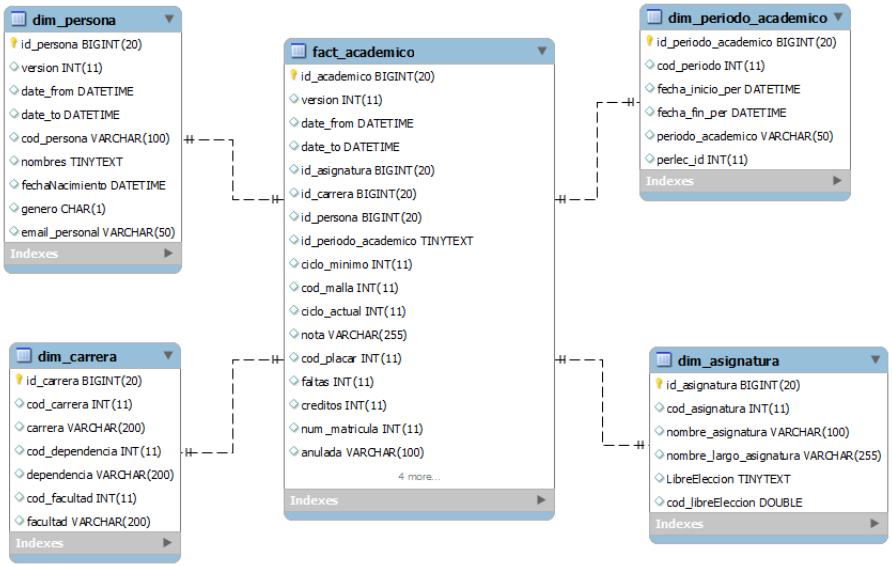


Figura 7.31: Modelo Lógico: Académico

7.11.2. ETL

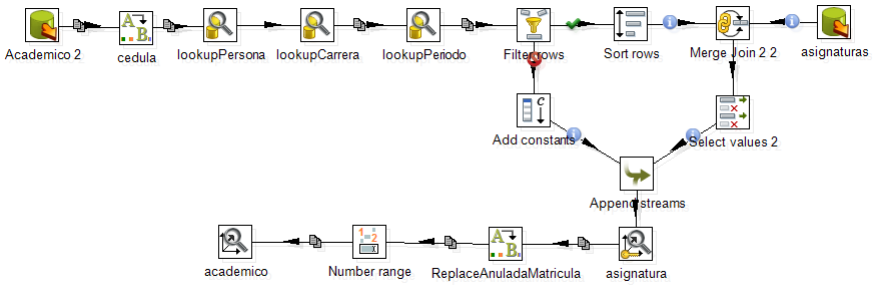


Figura 7.32: Transformación: Académico

7.12. Proceso: Socioeconómica

7.12.1. Modelo Lógico

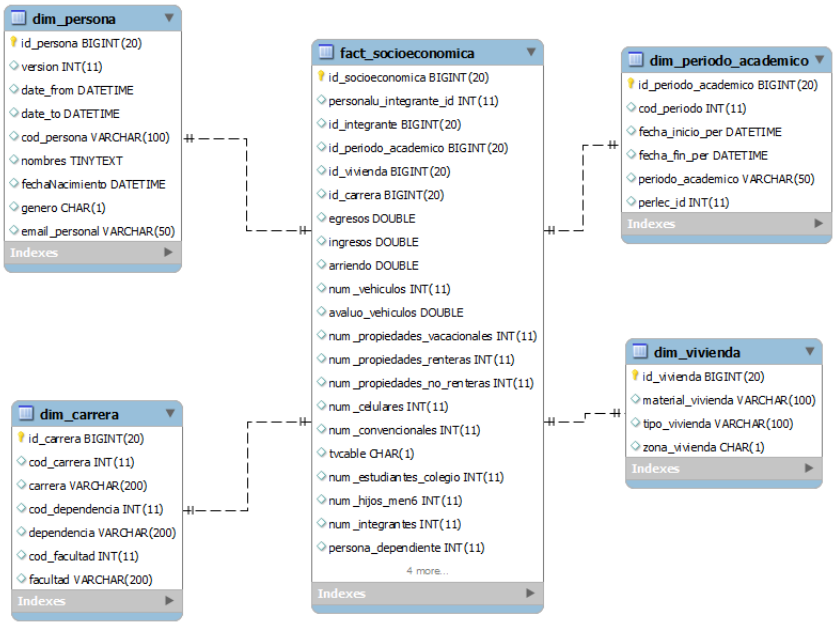


Figura 7.33: Modelo Lógico: Socioeconómica

7.12.2. ETL

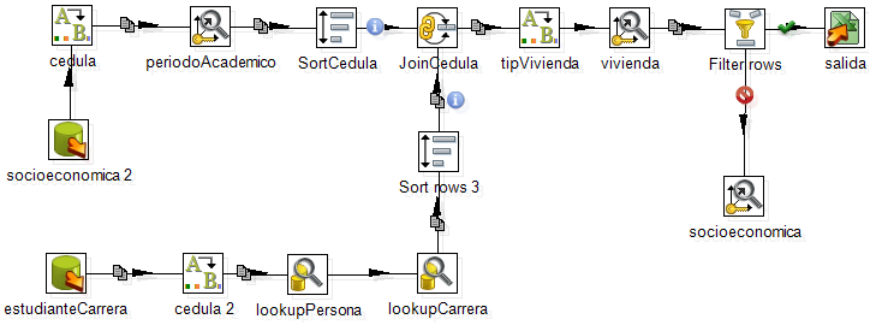


Figura 7.34: Transformación: Socioeconómica

Bibliografía

Fernando Virseda Benito y Javier Román Carrillo. Minería de datos y aplicaciones. págs. 5–7, 2010.

Ricardo Dario Bernabeu. *DATA WAREHOUSING: Investigación y Sistematización de Conceptos – HEFESTO: Metodología propia para la Construcción de un Data Warehouse. Córdoba.* 2010.

Giampaolo Del Bigio. *CDS/ISIS for Windows*, 1998.

Erika Carreón y Lic. Hugo Figueroa. *GUÍA PRÁCTICA DEL SISTEMA DE CLASIFICACIÓN DECIMAL DEWEY*, 2009.

Angel Cobo. *Diseño y programación de bases de datos.* Visión Libros, 2008.

Robert J Davenport. Etl vs elt. pág. 7, 2008. doi:http://www.dataacademy.com/files/ETL-vs-ELT-White-Paper.pdf.

Egbert de Smet y Ernesto Spinak. El abc del abcd : Manual del modulo central. 92, 2009.

Listas de Sorporte Tecnico BVS. Abcd automatizacion de bibliotecas y centros de documentacion. 55:6.

Sangeeta Namdev Dhamdhere. Abcd, an open source software for modern libraries. *Chinese Librarianship: an International Electronic Journal*, 17(32), 2011. doi: http://www.white-clouds.com/iclc/cliej/cl32dhamdhere.pdf.

Jennifer Ellis-Newman. Activity-based costing in user services of an academic library. *Library Trends 51*, 2003.

Jennifer Ellis-Newman, Haji Izan, y Peter Robinson. Costing support services in universities: An application of activity-based costing. *Journal of Institutional Research in Australasia*, 1996.

Jennifer Ellis-Newman y Peter Robinson. The cost of library services: Activity-based costing in an australian academic library. *Journal of Academic Librarianship 24*, 1998.

Ian H. Witten; Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., 2005.

William H. Inmon. *Building the Data Warehouse*. Wiley Publishing, Inc, 2005.

Gómez Gallego Juan PAblo. Modelo objeto relacional - ordbms. 2007. doi:http://es.scribd.com/doc/270513/Bases-de-datos-Objeto-relacional.

Ralph Kimball y Margy Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. John Wiley Sons, 2002.

Siguenza-Guzman Lorena, Alexandra Van den Abbeele, Joos Vandewalle, Henri Verhaaren, y Dirk Cattrysse. Using time-driven activity-based costing to support library management decisions: A case study for lending and returning processes. *Library Quarterly: Information, Community, Policy*, 84(1):1–23, 2014.

Joseph R. Matthews. Valuing information, information services, and the library: Possibilities and realities. *portal: Libraries and the Academy*, 13(1):91–112, 2013. doi:10.1353/pla.2013.0000.

Scott Nicholson. The bibliomining process: Data warehousing and data mining for library decision-making. *Information Technology and Libraries*, 22(4), 2003.

Scott Nicholson. A conceptual framework for the holistic measurement and cumulative evaluation of library services. *Proceedings of the American Society for Information Science and Technology*, 41(1):496–506, 2004.

Hernan Calle Okamura. Sistemas de gestión de bases de datos. 2012. doi:http://www.slideshare.net/HernanOkamura/sistemas-de-gestores-de-base-de-datos-13332504.

Oracle. Oracle database. 2013. doi:http://www.oracle.com/es/products/database/overview/index.html.

Carlos Ordoñez y Paul Cabrera. Desarrollo de un modulo tdabc, aplicado al centro de documentaciÓn regional "juan bautista vÁzquez". 322, 2012.

Bas Peters. *Crosswalking: Processing MARC in XML Environments with MARC4J*. Lulu.com, 2007.

PostgreSQL. *PostgreSQL 8.3.23 Documentation*, 2013. doi:http://www.postgresql.org/about/.

Gustavo R. Rivadera. La metodología de kimball para el diseño de almacenes de datos (data warehouses). *Cuadernos de la Facultad*, 5, 2010.

Manuel Rosa San Segundo. *SISTEMAS DE ORGANIZACIÓN DEL CONOCIMIENTO: La organización del conocimiento en las bibliotecas españolas*. Imprenta Nacional del boletín oficial del Estado, 1996.

Lorena Siguenza-Guzman. A holistic approach to supporting academic libraries in resource allocation processes. *VER????*, 2013a.

Lorena Siguenza-Guzman. Improving library management by using cost analysis tools: A case study for cataloguing processes. *LIBER 42nd Annual Conference*, 2013b.

Lorena Siguenza-Guzman, Ludo Holans, Alexandra Van den Abbeele, Joos Vandewalle5, Henri Verhaaren, y Dirk Cattrysse1. Towards a holistic analysis tool to support decision-making in libraries. *Proceedings of the IATUL Conferences*, 2013.

Jorge Sánchez. Principios sobre bases de datos relacionales. (1), 2004. doi:http://www.jorgesanchez.net/bd/bdrelacional.pdf.

Scott Nicholson; Jeffrey Stanton. Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries. *Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance*, págs. 247–262, 2003.

The DSpace Developer Team. *DSpace 1.8 Documentation*, 2012.

Simón Mario Tenzer. *Introducción a la Computación: Archivos, formatos y extensiones*, 2007.

Ioannis Kopanakis; Babis Theodoulidis. Visual data mining modeling techniques for the visualization of mining outcomes. *Journal of Visual Languages Computing*, 14(6):543–589, 2003.

Fayyad U, Piatetsky-Shapiro G., y Smyth P. *The KDD process for extrating useful know ledge from volumes of data Communications of the ACM*. Cambridge, 1996.

Stuart Weibel. The dublin core: A simple content description model for electronic resources. *You have free access to this content Bulletin of the American Society for Information Science and Technology*, 24(1):9–11, 2005.