FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS

DETECCIÓN DE PATRONES DE APRENDIZAJE EN UNA UNIVERSIDAD MEDIANTE UNA PLATAFORMA DE INTELIGENCIA DE NEGOCIOS

AUTORES

CRISTIAN OSWALDO CAISA VILLARROEL

EMILIO JOSÉ GARCÍA JARAMILLO

AÑO

2019

FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS

DETECCIÓN DE PATRONES DE APRENDIZAJE EN UNA UNIVERSIDAD MEDIANTE UNA PLATAFORMA DE INTELIGENCIA DE NEGOCIOS

Trabajo de Titulación presentado en conformidad a los requisitos establecidos para optar por el título de Ingenieros en Electrónica y Redes de Información.

Profesor Guía

Ms. William Eduardo Villegas Chiliquinga

Autores

Cristian Oswaldo Caisa Villarroel

Emilio José García Jaramillo

Año

2019

DECLARACIÓN PROFESOR GUÍA.

"Declaro haber dirigido el trabajo, detección de patrones de aprendizaje en una universidad mediante una plataforma de inteligencia de negocios, a través de reuniones periódicas con los estudiantes Cristian Oswaldo Caisa Villarroel y Emilio José García Jaramillo, en el semestre 201920, orientando sus conocimientos y competencias para un eficiente desarrollo del tema escogido y dando cumplimiento a todas las disposiciones vigentes que regulan los Trabajos de Titulación".

_____

William Eduardo Villegas Chiliquinga
Magister en Redes de Comunicaciones
C.I. 1715338263

DECLARACIÓN PROFESOR CORRECTOR.

"Declaro haber revisado este trabajo, detección de patrones de aprendizaje en una universidad mediante una plataforma de inteligencia de negocios, de los estudiantes Cristian Oswaldo Caisa Villarroel y Emilio José García Jaramillo, en el semestre 201920, dando cumplimiento a todas las disposiciones vigentes que regulan los Trabajos de Titulación".

_____

Iván Patricio Ortiz Garcés
Magister en Redes de Comunicaciones
C.I. 0602356776

DECLARACIÓN DE AUTORÍA DEL ESTUDIANTE.

“Nosotros, Cristian Oswaldo Caisa Villarroel y Emilio José García Jaramillo, declaramos que el presente trabajo de titulación es de nuestra autoría, todas las fuentes utilizadas para el desarrollo han sido debidamente citadas y se ha respetado la normativa que protege a los autores”

Cristian Oswaldo Caisa Villarroel	Emilio José García Jaramillo
C.I. 1724167612	C.I. 1723475115

AGRADECIMIENTOS

Agradezco a mis padres por ser un apoyo incondicional, y por saberme guiar con su sabiduría y buenos valores. Agradezco también a mis compañeros y amigos con los que se han compartido buenos y malos momentos durante la carrera. Agradezco a mis profesores quienes aparte de transmitir su conocimiento, también han sabido inculcar valores para convertirnos en buenos profesionales.

Cristian.

AGRADECIMIENTOS

Agradezco a mis padres por haberme brindado su apoyo constante en este trabajo, por enseñarme a que siempre hay algo que aprender. Agradezco a mis hermanos, amigos y personas muy cercanas que me han apoyado durante este proceso decisivo, por brindarme todo su apoyo, por enseñarme que las risas son necesarias y por un gran amor incondicional.

Emilio.

DEDICATORIA

Este trabajo va dedicado principalmente a mis padres por siempre apoyarme en cada obstáculo que se me ha presentado. La finalización de este trabajo y etapa universitaria es fruto del esfuerzo y la dedicación que han puesto en mí durante este tiempo.

Cristian.

DEDICATORIA

Este trabajo se lo dedico a mis padres por apoyarme durante todo este proceso de educación, por haberme permitido elegir esta carrera, por ayudarme con cada problema presentado y por guiarme como persona. El fin de esta etapa es fruto de sus grandes esfuerzos que han forjado, les dedico este gran paso a ellos.

Emilio.

RESUMEN

Este proyecto de titulación plantea implementar una plataforma de Business Intelligence (BI), que sea Open Source, para facilitar el análisis de la información almacenada tanto en la base de datos Moodle de la Universidad de las Américas, como en diferentes fuentes de información. El objetivo es identificar y dar seguimiento a los estudiantes que muestren problemas en el proceso de aprendizaje. Además, se busca ayudar a la toma de decisiones y generar conocimiento con respecto a los problemas detectados en los reportes que son obtenidos gracias al software Pentaho.

De esta manera se busca integrar a las diferentes áreas que conforman la Universidad para unir fuerzas y realizar todas las correcciones respectivas para que la calidad de educación mejore cada vez más.

ABSTRACT

This titling project proposes to implement a Business Intelligence (BI) platform, which is Open Source, to facilitate the analysis of the information stored in both the Moodle database of the University of the Americas and in different sources of information. The main goal is to identify and be able to follow up students who show problems in the learning process. Furthermore, it seeks to help the decision-making process and generate knowledge from the problems detected in the reports that are obtained thanks to the Pentaho software.

In this way we seek to integrate the different areas that make up the university to join forces and make all the corrections for the quality of education to improve more and more.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. Capítulo I. Introducción.

Esta implementación e investigación tiene como primera instancia realizar minería de datos con una herramienta de BI, conocida como Pentaho Data Integration (PDI). Esta herramienta sirve para realizar la extracción de información y volverla a cargar dentro de un Data Warehouse (DW), con distintos tipos de herramientas de diseño. Para poder realizar un procedimiento de BI de manera óptima, la herramienta de Pentaho incluye un proceso que se denomina como Extract, Transform and Load (ETL).

En el primer proceso denominado Extracción, se debe tomar en cuenta que se requiere de una fuente de datos en este caso una base de datos de la Universidad. Después de poseer ya la fuente, se procede a la puesta en escena, y se empieza por realizar la minería de datos, la cual recopila toda la información relevante que encuentre. Hay que mencionar que al tratar con bases de datos académicas los datos a extraer pueden ser nombre, apellidos, notas, tiempos, asistencias, entre otros.

En el proceso de transformación, se toma la información obtenida anteriormente y atraviesa por una serie de procesos realizados con la herramienta PDI. Estos procesos requieren de un diagrama en el cual consten trabajos y transformaciones *Jobs and Transformations*, dentro de PDI, el cual es intuitivo debido a que se realiza por medio de gráficos interactivos con la opción de arrastrar y soltar.

Además, en la transformación se puede observar distintas herramientas como la conexión con la base de datos que es muy importante ya que la conexión suele tener problemas; al obtener la conexión con la base de datos se pueden realizar distintas transformaciones, con cada uno de los campos a cambiar en las tablas de la base de datos. Toda la información que pasa por un proceso de transformación puede tener un formato de salida distinto los cuales serán mencionados posteriormente.

En el proceso de carga, se tiene como referencia la carga de la información hacia un DW donde todos los datos ya están almacenados de una mejor manera y su acceso va a ser óptimo. PDI ofrece distintas opciones para la salida de información con distintos formatos como, archivos de texto (.txt), archivos Excel (.xls), entre otros. En este caso se toma la salida de información como Structured Query Language (SQL), para la posterior carga al DW. Con toda la información consolidada se puede realizar un análisis de notas y detectar patrones de distintos tipos de actividades estudiantiles, con el fin de obtener informes para detectar problemas en el aprendizaje y hallar una solución al problema.

Conforme avanza el tiempo, aumenta la necesidad de utilizar la tecnología en la actualidad. La cantidad de información que se puede almacenar de grandes empresas es masiva. El volumen de datos en el mundo de los negocios está aumentando entre 35% - 50% cada año. En promedio una empresa grande llega a procesar 60 Terabytes de datos anuales, que representa mil veces más que la década anterior (Beath, Ross, Short, & Becerra, 2012).

En los años 80, gran parte de las empresas se concentraban más en la consolidación de datos en un DW para incrementar la eficiencia y la exactitud de los datos, pero no se tomó en cuenta el uso que se le daría, sino que comenzaron a utilizar una metodología de gestión en cascada con una línea de tiempo basada en años. Desafortunadamente la gran inversión que significó la implementación de esta tecnología no ofreció nada satisfactorio en sus primeras etapas, distintos equipos fueron descartados de la idea y los proyectos de DW fueron cerrados antes de brindar algún servicio a los usuarios finales.

En los años 90 un grupo de proveedores de BI como Cognos (IBM), Business Objects, Hyperion y Microstrategy entró en el mercado con interfaces gráficas mucho más fáciles de usar que permitieron una rápida respuesta del desarrollo de informes y análisis que permitieron a los usuarios comerciales dividir y dar datos (slice and dice) con o sin un almacén de datos completamente desarrollado. Al mismo tiempo, los proveedores de DBMS comenzaron a agregar funcionalidades de tipo BI a sus bases de datos con tecnologías como cubos de

datos y On-Line Analytical Processing (OLAP) (Thomsen, 2002) lo que ocasionó que la industria comience a ver los beneficios de invertir en dicha área.

La gran cantidad de datos que se comenzaron a generar necesitaba ser almacenada y es aquí donde empieza el surgimiento de las bases de datos, pero de igual forma hay que saber administrar información masiva de datos de manera correcta. En la actualidad, BI y DW están firmemente posicionadas en la parte de almacenamiento de datos, pero todavía existía la necesidad de desarrollar una gestión más exacta y eficiente de todos estos datos para poder manejar las nuevas tecnologías y los tipos de datos que se siguen presentando.

En cuanto al almacenamiento de datos, los proveedores de DBMS como Teradata y Netezza, están establecidos con diferentes arquitecturas que llevaban consigo un mejor sistema de análisis de datos. Los conceptos de modelado de datos, como los almacenes de datos operativos, los almacenes de datos lógicos, con los conceptos de *datamart*, encontraron su lugar en el modelo. Estos conceptos, junto con los avances en ETL y Middleware, ahora permiten una mayor eficiencia en el movimiento y procesamiento de datos.

1.1 Alcance

Para el actual trabajo de titulación se plantea instalar el software Pentaho en el centro de datos de la Universidad, en el cual se comenzará con la fase de extracción de la información de la base de datos de una plataforma educativa (Moodle), una vez extraídos los datos se procederá a establecer un formato ya que el objetivo de las herramientas ETL es extraer información de distintas fuentes entre los cuales se pueden encontrar bases de datos, archivos planos, hojas de cálculo, entre otros. Para posteriormente analizar la información obtenida y así poder observar distintos indicadores como: porcentajes de faltas, calificaciones bajas o la influencia de la situación económica en el desempeño de los estudiantes.

1.2 Justificación

La necesidad surge porque se busca mejorar el rendimiento académico en los estudiantes de la Universidad, y para ello es indispensable saber los problemas que impiden que se pueda lograr con dicho objetivo. Al procesar la información del Moodle se podrá identificar variables que indiquen dónde radican los problemas, entre los cuales se podría mencionar problemas económicos o familiares, observar los progresos o semestres en donde se obtienen las mejores/peores calificaciones.

Toda esta información tendría un gran valor para el departamento encargado porque ayudaría a la toma de decisiones para mejorar la calidad educativa.

1.3 Objetivos

1.3.1 Objetivo general

- Implementar una plataforma de inteligencia de negocios para el análisis de datos educativos con el uso de herramientas Open Source.

1.3.2 Objetivos específicos

- Implementar una arquitectura de BI con el uso de Pentaho Data Integration.
- Aplicar algoritmos de minería de datos para identificar las diferentes variables que aporten al aprendizaje de estudiantes universitarios.
- Presentar informes y tableros de control para la toma de decisiones.

2. Capítulo II. Marco Teórico.

En este capítulo se recopilará las definiciones conceptuales de lo que implica un análisis de BI donde se conocerá cada una de las herramientas a utilizar. Además, se menciona las características de cada una de estas.

2.1 Inteligencia de Negocios

BI está definido por sistemas que combinan los siguientes puntos:

- Recopilación de datos
- Almacenamiento de datos
- Manejo del conocimiento

Y de cada sistema se realiza un análisis de la información crítica y competitiva para presentarla al personal que se encarga de la toma de decisiones que va de la mano con la calidad de la información de entrada para el proceso de decisiones.

El objetivo de BI es analizar grandes cantidades de datos (estructurados o no estructurados) que brindan información en el momento, ubicación y forma adecuada. (Negash & Gray, 2008)

2.1.1 Aplicaciones

La fortaleza de BI se encuentra en el conocimiento que proporciona para la toma de decisiones, y los vendedores de herramientas de BI han sacado provecho de ello ya que cada una se distingue por ofrecer diferentes formas de generar los reportes, cada vez de manera más fácil, para que la empresa pueda sacar provecho y ventaja sobre sus competidores.

Dentro de las aplicaciones se considera:

- Reportes. Se generan reportes con un formato establecido para ser distribuidos entre los usuarios pertinentes para la toma de decisiones.
- Cubos. Proporcionan información analítica para la empresa.
- Minería de datos. La herramienta ayuda a desarrollar el modelamiento predictivo.
- Ad Hoc Query. La herramienta permite observar la base de datos para extraer información de fácil comprensión.

2.2 Almacenes de datos

Un almacén de datos permite recolectar datos de distintas fuentes externas para ser centralizados dentro de una base de datos multidimensional, y que de esta manera se los pueda procesar, para luego ser convertidos en herramientas que ayuden a la competitividad de la empresa en el momento exacto.

2.2.1 Características

Dentro de las características de los DW se puede encontrar:

- Orientado a temas. El DW debe estar directamente relacionado al giro de negocio de la empresa para que de esta manera el análisis de las fuentes de datos pueda brindar una ventaja competitiva real y eficaz. Es decir, la información que se va a clasificar debe ser de real interés para la organización ya que de esto depende el modelamiento e implementación de la información que va a contener. (Sinnexus, 2017)
- Integrado. Es integrado porque aquí se concentra toda la información de distintas fuentes de datos y la integra como uno solo. Dentro de esta característica se encuentra el proceso de ETL. La información generada por distintos departamentos de la organización será integrada dentro de una sola instancia para ser cargados al DW. La integración también hace referencia a que los distintos datos van a ser incluidos dentro de un mismo estándar para evitar problemas o confusiones entre los nombres de las tablas, campos, claves, entre otros. La importancia de esta característica está en que toda la información dentro del DW siga un modelo aceptable y confiable para que el personal que use dicha información no tenga que preocuparse por la solidez o veracidad de esta. (Sinnexus, 2017)
- No volátil. Un DW almacena la información de la empresa a lo largo del tiempo para analizar sus datos, por lo que, resultaría inútil que toda esta información se pierda de un momento a otro. (Sinnexus, 2017)
- Estable. No debe ser afectado por cambios en otros sistemas, por ejemplo, sistemas OLTP o sistemas operacionales. (Sinnexus, 2017)

2.2.2 Ventajas

Dentro de las ventajas de tener un DW implementado se encuentran:

- Toda la información generada por las aplicaciones de los departamentos de la empresa ahora tiene una orientación a la toma de decisiones que, además, está dentro de una plataforma confiable e integrada.
- Aumenta la competitividad porque se puede tomar acciones inmediatas ante cambios en el mercado.
- Genera reportes de fuentes confiables y que son correctos, oportunos y de fácil acceso.
- Incremento en la eficiencia de toma de decisiones estratégicas de los empleados ya que la información es de mejor calidad y está disponible en línea para mejorar la accesibilidad. (Kimball & Caserta, 2004)

2.2.3 Desventajas

El gran capital que se debe depositar en esta nueva infraestructura es una de las principales ventajas, además, los usuarios deben ser capacitados y originalmente se van a oponer al cambio hasta que sus resultados sean reales a su perspectiva.

2.3 Minería de datos

La minería de datos es una de las etapas de un proceso mucho más grande que se lo conoce como el *Descubrimiento de conocimiento en bases de datos*, *Knowledge Discovery from Data* (KDD) y, generalmente, se lo define como el proceso de descubrir patrones o conocimiento de interés a partir de grandes cantidades de datos. Dichos datos pueden provenir de bases de datos, almacenes de datos, repositorios web o simplemente datos generados por un sistema. (Han, Pei, & Kamber, 2011)

El proceso de la minería de datos se lo puede observar en la figura 1.

Un proceso de KDD consiste en los siguientes pasos:

- Recopilación. Consiste en extraer los datos de diversas fuentes.
- Selección, limpieza y transformación. En esta parte se comienza a filtrar los datos que no son necesarios o que son incorrectos para luego transformarlos y normalizarlos.
- Minería de datos. Se comienza con técnicas para obtener modelos de predicción a partir de la detección de patrones, tendencias, comportamientos de los datos, entre otros.
- Interpretación y evaluación. Aquí se comienza a analizar los modelos que se han generado de la minería de datos para lograr entender el contenido de las fuentes de datos.

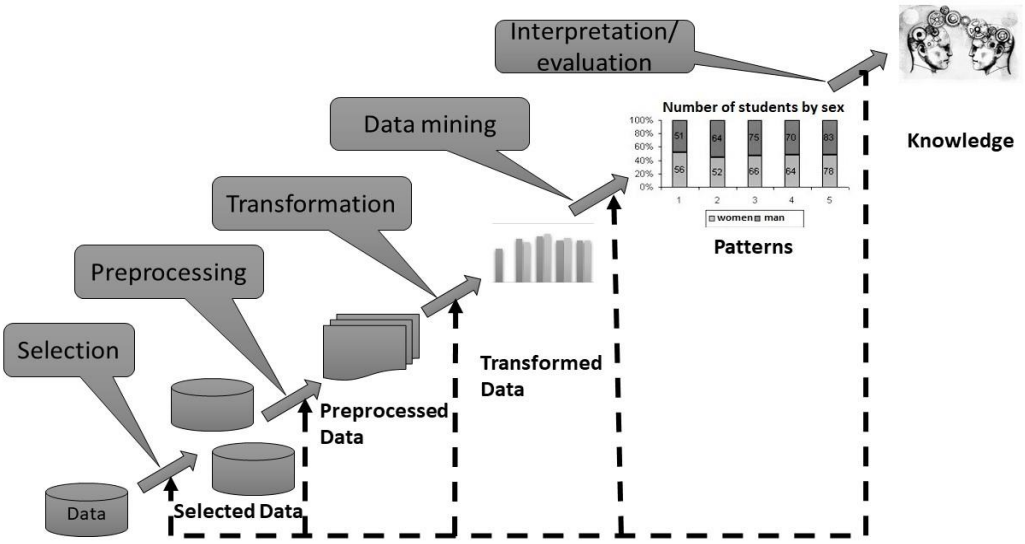


Figura 1. Proceso KDD.

Tomado de (Villegas & Luján, 2017)

2.3.1 Tipos de datos que pueden ser minados

Desde un punto de vista general, a cualquier tipo de dato se le puede aplicar la minería mientras sean relevantes entre los cuales destacan bases de datos, DW y datos transaccionales que son los más importantes para el desarrollo de este proyecto de titulación. (Han, Pei, & Kamber, 2011)

2.3.2 Información de bases de datos

Consisten en colecciones de datos interrelacionados, que tienen estructuras y almacenamiento definido por un software conocido como database management system (DBMS). Este software permite administrar y acceder a los datos, así como también manejar la concurrencia de datos distribuidos o compartidos, además, asegura la consistencia y seguridad de la información de la base de datos.

Al minar bases de datos relacionales se puede avanzar con mayor rapidez en la búsqueda de patrones de datos que pueda predecir, por ejemplo, el riesgo de otorgar créditos a nuevos clientes de un banco basados en sus ingresos mensuales, edad o información crediticia. Es por esta razón que las bases de datos relacionales son las más usadas y con mejores repositorios de información para la minería de datos.

2.3.3 Información de almacenes de datos

Es un repositorio de información que tiene distintas fuentes, pero con un almacenamiento unificado. Generalmente tienen una estructura multidimensional donde cada dimensión es un atributo y cada celda tiene el valor de alguna operación. Las herramientas para minería de datos ayudan a este tipo de estructuras para tener un análisis mejor, y se lo conoce como minería de datos multidimensional o minería de datos multidimensional exploratoria.

Con el uso de OLAP se puede mejorar la minería porque permite ampliar la exploración de múltiples dimensiones con diferentes niveles de granularidad y

de esta manera se tiene un mejor potencial para descubrir patrones aún más interesantes.

2.3.4 Información transaccional

Las bases de datos transaccionales almacenan todo tipo de información que está relacionada con las compras que hacen usuarios en páginas web, estas compras son identificadas con un ID que a su vez relaciona la compra con los ítems adquiridos y otro tipo de información.

Con la minería de datos transaccionales se puede, por ejemplo, minar los ítems que se venden juntos como una computadora con una impresora y ofrecer descuentos u ofertas que aumenten las ventas.

2.4 Extracción, transformación y carga

ETL hace referencia al proceso de extraer, transformar y cargar la información. Este proceso consiste en, primero, organizar la información de las fuentes de datos para después transformarlas en un solo formato y finalmente cargarlas a un almacén de datos, tal como muestra la Figura 2.

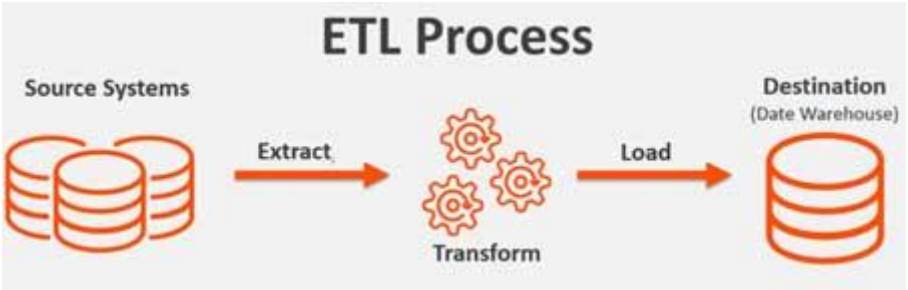


Figura 2. Proceso ETL.

Tomado de (Databricks, 2019)

2.4.1 Extracción

Es la primera fase de un proceso ETL donde se obtiene la información de diferentes fuentes de datos heterogéneas y se realiza tareas tales como:

- Extracción de información de fuentes como bases de datos, archivos planos, hojas de cálculo o mainframes.
- Breve análisis de datos extraídos para ubicarlos en un formato establecido que se utilizará en la siguiente fase.

Dentro de esta etapa se debe tener precaución de no afectar el rendimiento de las fuentes de datos.

2.4.2 Transformación

Es la segunda fase en la cual los datos previamente extraídos son colocados dentro del formato establecido, y también, se aplican reglas que las define el propietario.

Al extraer información de distintas fuentes, una etapa de transformación es imprescindible para el proceso porque muchas de las veces las fuentes externas pueden tener diferencias en formatos, idiomas, unidades de medida o nomenclatura en las bases de datos lo que haría imposible comparar y analizar la información. (PowerData, 2013)

2.4.3 Carga

En la fase final toda la información transformada se la carga en el DW definido y se obtiene el resultado para que el sistema ETL fue diseñado.

Cabe recalcar que en el proceso de extracción se definen las fuentes de datos que son necesarias para que el resultado final sea el deseado y todo esto es debidamente documentado.

2.4.4 Ventajas

- Gracias a que el diseño comienza desde el resultado deseado se puede definir previamente las fuentes de datos necesarias para lograr dicho

objetivo, de esta manera se logra reducir considerablemente recursos y tiempo de ejecución.

- Debido a que la solución está muy bien definida, el DW contiene únicamente la información relevante.
- Un proceso ETL puede ejecutar operaciones más complejas en diagramas de flujo simples.

2.4.5 Desventajas

- En la fase de transformación todo el procesamiento cae sobre la herramienta y servidor ETL, lo que causa que el tiempo requerido para terminar la operación sea mayor, y también, el costo de hardware puede subir.
- La herramienta ETL muchas veces también verifica la calidad de los datos fila por fila, lo que causa que el resto de los procesos se encuentren en un cuello de botella.
- Existe un desperdicio de recursos en la red debido a que los datos viajan de las fuentes al servidor ETL y luego del servidor al DW.
- Debido a que el diseño se enfoca en el resultado deseado, se extrae únicamente la información necesaria; pero, esto se puede ver como una desventaja al momento que en el futuro se necesite información extra para distintos requerimientos y, por lo tanto, todo el proceso ETL se tendrá que repetir. (Ranjan, 2011)

2.5 Pentaho Data Integration (PDI)

Es un aplicativo que brinda un servicio de BI con herramientas ETL. Estas herramientas tienen habilidades la cuales facilitan el proceso de la extracción, limpieza y carga de datos usando un formato único, con el fin de que sean accesibles para los usuarios finales. (Hitachi, 2018)

2.5.1 Kettle

Kettle (acrónimo recursivo: kettle extraction, transformation, transport, and load environment), esta palabra hace referencia al proceso ETL. En la versión de Pentaho Community Edition existen componentes con el nombre de Spoon, Pan y Kitchen, las cuales hacen referencia a una metáfora de lo que ofrece un ETL. (Hitachi, 2018)

2.5.2 Requerimientos del sistema

Entre los requerimientos de todo el sistema de Pentaho 8.1 se tiene:

- Requerimientos de servidor Pentaho (Hardware y Software de 64 bits):
    - CentOS 5 o posterior
    - Redhat Enterprise 5 o posterior
    - Microsoft Windows Server 2008 R2 o posterior
    - SUSE SLES 11
    - Ubuntu Server 12.04 LTS o posterior
    - RAM de 8 GB con 4 GB dedicados a los servidores de Pentaho.
    - 20 GB de almacenamiento libre después de la instalación
    - Tomcat 8.0 y 8.5
    - JBoss EAP 7.x
    - Oracle Java 8
- Requerimiento para Herramientas de Pentaho (Hardware y Software de 64 bits):
    - Microsoft Windows 7, 8 y 10
    - Ubuntu Desktop 14.04 LTS y 16.04 LTS
    - RAM: 2 GB para la mayoría de las herramientas de diseño y PDI necesita 2 GB dedicados
    - 2 GB de almacenamiento libre después de la instalación

Requerimiento para herramientas en línea (Hitachi Vantara, 2017)

- Apple Safari 10 y 11 solo en OS X

- Google Chrome 64 y 65
- Microsoft Edge 40.15063 y 41.16299.248.0
- Microsoft Internet Explorer 11
- Mozilla Firefox 58 y 59

2.5.3 Diseñador de Reportes

Es una herramienta de Pentaho que tiene como exclusividad ser el único que crea reportes muy detallado respecto a la información analizada. Report Designer forma parte de la distribución de Business Analytics de Pentaho. Todos los reportes creados por esta herramienta son desarrollados con datos preparados adecuadamente desde cualquier fuente de datos.

El diseñador de reportes es una de las herramientas más convenientes para crear informes con software de Pentaho. Además, Pentaho posee una interfaz web del servidor instalado, en la cual se puede observar de manera detallada la proyección de información interactiva que tiene como objetivo ser más intuitiva para el lector. También se puede integrar una herramienta perteneciente a Pentaho llamada Reporting Engine, en donde se integra el diseño de informes. (Hitachi, 2018)

2.5.4 Diseñador de agregación

Esta herramienta facilita la creación y la presentación de las tablas seleccionadas que mejoran el rendimiento de Pentaho Analysis en sus cubos OLAP. Pentaho Analysis es un motor OLAP relacional puro, con el fin de funcionar únicamente con los datos almacenados en su propia base de datos relacional en lugar de brindar su propio modelo multidimensional. Esto facilita la implementación y administración de datos, pero implica ciertas repercusiones cuando se trabaja con conjuntos de datos muy grandes. (Hitachi Vantara, 2018)

Para mejorar el rendimiento Pentaho Analysis permite la agregación de tablas. Estas tablas coexisten con la table base, la cual está preconfigurada basándose en una tabla modelo. Este tipo de práctica mejora el ambiente de trabajo para

que cumpla ciertas consultas con una tabla pequeña, en lugar de agregar una cantidad larga de información de hechos individuales desde la tabla modelo. (Hitachi Vantara, 2018)

El diseñador de agregación proporciona una interfaz de fácil manejo hacia el usuario en la cual se permite la creación de tablas agregadas de acuerdo con los niveles que se especifique. Sin embargo, según las decisiones tomadas en el punto anterior generan el Data Definition Language (DDL) para crear las tablas agregadas, Data Manipulation Language (DML) para poder realizar evaluaciones de estas. Si no se encuentra familiarizado con la creación de tablas agregadas, el diseñador de agregación posee un instructor inteligente el cual evalúa la estructura y cardinalidad de sus cubos OLAP y sugiere unas tablas agregadas iniciales para su creación y mejorar el rendimiento. (Hitachi Vantara, 2018)

2.5.5 Editor de Metadatos

Es una herramienta de Pentaho, como su nombre lo indica su funcionamiento es la edición de metadatos. En esta herramienta se pueden construir dominios y modelos de metadatos, para su facilidad esta herramienta también cuenta con muestras si desea probarlo antes de importar sus propios datos.

Cualquier persona con pocos conocimientos sobre el manejo de la información puede usar la herramienta, pero es recomendable que un administrador de base de datos (DBA) la use. El administrador deberá tener distintos conocimientos en las siguientes tareas:

- Importación de tablas
- Creación de relaciones entre tablas
- Asignar agregaciones
- Agregar categorías
- Asignar seguridad

El administrador debe tomar en cuenta la información que posee su base de datos para poder brindar los tipos de datos que desean los usuarios de negocios.

Este tipo de procedimientos permite asignar un modelo de usuario empresarial (lógico) a una base de datos relacional compleja. Lo que permite a los usuarios crear informes de Pentaho sin la necesidad de un administrador de base de datos (DBA). (Hitachi Vantara, 2018)

2.5.6 Esquema de banco de trabajo

Pentaho es una herramienta que trabaja con distintos tipos de esquemas y a su vez con un sinfín de grandes consultas con grandes volúmenes de datos. Un esquema de Mondrian es, en toda su esencia, un archivo XML que impulsa esta consulta. Tomando en cuenta lo anterior un esquema de Mondrian define su estructura como una base de datos multidimensional, este tipo de herramientas se pueden usar en Pentaho Schema Workbench (Hitachi Vantara, 2018).

Por ejemplo, en un escenario muy básico, se creará un esquema de Mondrian con un cubo los cuales tendrán una única tabla de hechos y distintas dimensiones, cada una con su propio orden jerárquico que se caracteriza por tener una clasificación de niveles (Pentaho, 2011). Los esquemas más difíciles pueden terminar con una cantidad mayor de cubos virtuales y, en lugar de hacer una consulta hacia la única tabla de hechos en el centro de un esquema, podrían realizar las consultas orientándose a las vistas o tablas direccionales en su lugar. (Hitachi Vantara, 2018)

Todos los archivos XML de Mondrian se agrupan en una sola lista rápida y en una pequeña lista completa para cada elemento. Pentaho Schema WorkBench se utilizará con el fin de crear el esquema de Mondrian en forma de gráficos para que sea más entendible, se trata de una demostración. También la herramienta cuenta con un asistente de fuente de datos el cual puede realizar un diseño más complejo.

2.5.7 Java Database Connectivity Driver

JDBC es el estándar de la industria que sirve para entablar una comunicación entre el lenguaje de programación de Java y un amplio rango de bases de datos

SQL y otras bases de datos tabulares. También realiza el envío de acuerdo a SQL y el proceso de resultados. (Oracle, 2015) En cualquier sitio web de empresas de bases de datos se puede encontrar el controlador JDBC siempre para la descarga gratuita, en la mayoría de los casos también es conocido como conector J. Este controlador es indispensable adjuntarlo a Pentaho debido a que se realiza una conexión con una base de datos que requiere de una comunicación con el programa y este lo hace mediante el conector J.

2.5.8 Tomcat

Apache Tomcat es un servicio de código abierto de Java Servlet desarrollado por Apache Software Foundation (ASF). Este programa lo que ofrece es, en simples palabras, un ambiente de servidor HTTP en el cual Java puede ser ejecutado. (Apache, 2010)

2.6 MySQL Database

My Structured Query Language (MySQL), es una de las herramientas más populares de bases de datos de código abierto que entrega una fuerte confianza en cuanto a rentabilidad, escalabilidad, alto rendimiento, procesos de transacción y aplicaciones de bases de datos embebidas. La base de datos MySQL brinda las siguientes características:

- Alto rendimiento y escalabilidad, con el fin de conocer la demanda exponencial de información y usuarios.
- Clústeres de replicación autorreparable, para proveer escalabilidad, rendimiento y disponibilidad.
- Cambios de esquema en línea, para conocer los cambios del negocio.
- Esquema de rendimiento, para el manejo correcto de usuarios, niveles de rendimiento de aplicaciones y consumo de recursos.
- Tablas relacionales y archivos JSON, implementados en una misma base de datos.

- Plataforma independiente, brindando flexibilidad en distintos sistemas operativos.
- Interoperación de big data, usando MySQL como almacén de datos para Hadoop y Cassandra. (Oracle , 2018)

2.7 VMware ESXi

VMware ESXi es un hypervisor con una arquitectura bare-metal, se considera una herramienta robusta debido a que su instalación ocurre directamente en el servidor físico. Al ser instalada en un servidor físico se puede hacer la creación de máquinas virtuales dentro del mismo servidor y además se puede acceder y controlar de manera directa los recursos. Es líder en la industria por su eficiente arquitectura y por brindar confiabilidad, rendimiento y soporte. Como características posee:

- Seguridad mejorada, brinda la posibilidad de asignar permisos de administrador a usuarios con menos privilegios. Se elimina la característica de tener una cuenta maestra que maneje a las demás.
- Registros y auditorias, toda acción que realice un usuario se almacena.
- vMotion, realiza el traspaso de una máquina virtual completa desde un servidor hacia otro, sin perder la disponibilidad de la información.
- Las máquinas virtuales admiten hasta 128 CPU virtuales.
- Las máquinas virtuales admiten hasta 4 TB de RAM.
- El tamaño máximo de archivos (.vmdk) es de 62 TB.
- Tiene integración con Active Directory.
- Posee una interfaz en línea con la asignación de usuario y el monitoreo del servidor y cada una de las máquinas virtuales instaladas. (VMware, 2017)

3. Capítulo III. Diseño e implementación del modelo.

En este capítulo se detallan distintos aspectos técnicos como la comparación y selección de herramientas, instalación de programas, selección de fuentes de

datos, procesos de BI, diseño de data warehouse y procesos para la minería de datos.

3.1 Comparación de las herramientas de BI

En la actualidad existe una gran variedad de herramientas de BI, entre ellas se pueden destacar algunas herramientas desarrolladas por marcas conocidas; por ejemplo, Cognos de IBM, MicroStrategy, Oracle Business Intelligence (OBI), Business Object de SAP, entre otras. Al comparar una herramienta con otra se tiene en cuenta el rendimiento, escalabilidad, disponibilidad, integridad de la información, precios, compatibilidad con fuentes de datos, entre otras características (Finances Online, 2019). Como se menciona en el documento, la herramienta de BI que se va a utilizar se denomina Pentaho Data Integration que pertenece a la compañía de Pentaho. Se va a dar una breve explicación de los pros y contras de esta herramienta, tomando en cuenta cada una de sus características para así dar una comparación argumentada.

3.1.1 Pentaho

Se toma en cuenta que Pentaho es una compañía que se dedica a desarrollar herramientas de BI, por lo tanto, es lógico que posee distintos tipos de herramientas para todos sus consumidores. Pentaho es una herramienta Open Source y ofrece un servicio sin costo, sin embargo, también existe un servicio de pago. PDI (Community Edition), es la herramienta de BI sin costo alguno y PDI (Enterprise Edition), es la herramienta de BI con un costo mensual o anual.

3.1.2 PDI Community Edition (CE) y PDI Enterprise Edition (EE)

Primero se realiza un análisis entre las herramientas de la misma compañía para después empezar con la comparación con herramientas externas que son consideradas las más usadas (Finances Online, 2019); PDI CE tiene como característica los reportes empresariales los cuales brindan los siguientes servicios a sus usuarios:

En la parte de funcionalidad se tiene:

- Reportes empresariales para los desarrolladores de informes.
- Buen diseñador de gráficos.
- Reportes empresariales perfectos.
- Basado en Wizard (Procedimientos por pasos).
- Diseñador de gráficos para la integración de datos.
- Conectividad con bases de datos relacionales, analíticas y fuentes no SQL.
- Procesos ETL.
- Análisis de predicciones.
- Diseñador de gráficos para Big Data.

Para el soporte y mantenimiento, PDI CE posee un único servicio el cual es soporte con foros de la comunidad en-línea. En cuanto al licenciamiento, PDI CE tiene una ventaja sobre los demás debido a que su licencia no tiene costo alguno. A diferencia de PDI EE, que posee los servicios antes mencionados con la agregación de otros tipos debido a que es la versión de pago.

Entre los servicios agregados de PDI CE se pueden nombrar los siguientes:

- Reportes interactivos para usuarios de negocios.
- Reportes Ad Hoc (Reportes con un fin).
- Integración segura de datos.
- Procesos ETL.
- Conectividad con aplicaciones empresariales como Google Analytics, Google Docs y Servicio de mensajes de Java.
- Instaladores automatizados, administración centralizada, monitoreo de rendimiento, solución y diagnóstico, expiración de contenido automatizado, respaldo y recuperación, reportes de auditoria e inicio de sesión único.
- Análisis de predicciones
- Análisis instantáneo de Big Data.

Para soporte y mantenimiento se tiene:

- Software certificado y de calidad garantizada.
- Mantenimiento de actualizaciones.
- Portal de soporte, soporte por teléfono, servicio premium de Service Level Agreements (SLAs).
- Tres contactos de soporte nombrados.
- Foros de la comunidad, foros de los clientes

Entre los servicios profesionales se tiene:

- Soluciones y servicios de Pentaho y socios certificados.
- Tres consultorías gratuitas.
- Entrenamiento y clases en-línea.
- 12 créditos de entrenamiento.

Finalmente, licenciamiento:

- Existe una suscripción anual o una suscripción mensual, la parte de los costos únicamente pueden ser obtenidos por parte de la compañía. (Pragmatic, s.f.)

3.1.3 PDI y Cognos Analytics

Luego de conocer las características de PDI aparece Cognos (perteneciente a IBM) que es un autoservicio, lo que quiere decir, que el usuario puede escoger de manera subjetiva cada uno de los servicios que esta herramienta posea. Cognos también es una herramienta que está basada en un ambiente web por lo que todo su monitoreo va a ser a través de una página web. Cognos brinda un sin número de características estándar como, reportes, análisis, monitoreo, eventos y métricas.

Cognos posee distintas herramientas las cuales proveen un funcionamiento adicional. Cognos Framework Manager, Cognos Cube Designer y Cognos

Transformer, estas herramientas agilitan el procesamiento de grandes cantidades de datos y sus reportes. Los reportes dan al usuario la habilidad no solo de visualizar las percepciones, sino también compartirlo entre colegas. (SelectHub, s.f.)

De acuerdo con Finances Online, PDI tiene una calificación de 8.1 sobre 10, en comparación con Cognos que tiene una calificación de 9.1. Las satisfacciones del usuario son equilibradas para ambas herramientas. Sin embargo, en comparación con PDI, se impartirá mencionando las características adicionales de Cognos, las cuales son las siguientes:

Características de Cognos faltantes en PDI:

- Distintos tipos de visualizaciones y estilos en los reportes.
- Control de la protección de la información de reportes.
- Datos protegidos con distintos niveles de permisos.
- Arrastrar y soltar en la herramienta móvil.
- Contenido disponible sin la necesidad de internet
- Basado en ambiente web.
- Combinar fuentes de datos.

En cuanto al licenciamiento, se puede mencionar que Cognos de IBM tiene distintos planes con el fin de cumplir con las necesidades del usuario. Sus precios son elevados según “SelectHub” con una calificación de cinco puntos sobre cinco puntos (SelectHub, s.f.). Y según Finances Online, para un grupo de trabajo su precio está en 1.990,00 dólares mensuales, su versión estándar alrededor de 10.100,00 dólares mensuales y, finalmente, su versión Enterprise la cual tiene un valor de 19.950,00 dólares mensuales. Cada plan se menciona con distintos ajustes de acuerdo con lo que el usuario necesite. (SelectHub, s.f.)

3.1.4 PDI y Business Objects

Business Objects (BO) es una herramienta de BI que pertenece a la compañía SAP, está disponible con numerosos paquetes y diferentes herramientas para

compañías de distintos tamaños. La plataforma de análisis reduce los costos de TI y carga de trabajo, también puede descubrir y brindar percepciones para mejorar la toma de decisiones. Business Objects se conecta con una variedad de aplicación como, SAP business warehouse y SAP HANA para tener la función de análisis en tiempo real.

Características de Business Objects faltantes en PDI:

- Aplicaciones móviles tanto en Android como iOS.
- Arrastrar y soltar tanto para desarrolladores como para usuarios finales.
- Auditorias.
- Manejo del ciclo de vida.
- Análisis visual, reportes y monitoreo, todo en tiempo real.
- Entrega de contenido eficiente y de alto volumen.
- Integración de visualizaciones con Microsoft PowerPoint.
- Información entregada con Microsoft Excel.

En cuanto a licenciamiento. Empieza desde una prueba gratuita por 30 días, de haber cumplido los 30 días, el precio comienza a ser definido por la compañía con diferentes tipos de cuotas mensuales o anuales, las cuales se van a basar de acuerdo con el plan que desea el usuario. Existen dos versiones con un precio ya establecido; la versión Edge comienza con 14.000,00 dólares al año y la versión Enterprise Premium comienza con 55.000. dólares al año. Según “SelectHub” SAP Business Objects tiene una calificación de tres puntos sobre cinco puntos en cuanto al costo. (SelectHub, s.f.)

3.1.5 PDI y Sisense

Sisense es una solución para BI debido a que brinda herramientas para la preparación y análisis de datos complejos. Sisense administra todo el ciclo de vida del análisis de datos, desde la extracción, procesamientos, minería, visualización hasta donde los datos se unen. Provee capacidades ETL y un

ambiente de administración visual para el manejo de modelos complejos de datos.

Características de Sisense faltantes en PDI:

- Capacidad de mover la información de 50 a 100 veces más rápido.
- Es escalable.
- Memoria optimizada en bases de datos basadas en columnas.
- Permite el manejo de Terabytes de información y docenas de consultas (Queries) simultáneas.
- Divide a las consultas en bloques y cada bloque es procesado individualmente.
- Costos adicionales por mantenimiento.
- Bases de datos con arquitectura en chip.

En cuanto al licenciamiento, Sisense está orientado para pequeñas, medianas y grandes empresas. Su acceso está disponible en múltiples plataformas Windows, Linux y Mac. Al igual que todas, Sisense posee distintos planes y una versión gratuita, sin embargo, la suscripción para cinco usuarios tiene un costo de 21.000 dólares al año. “SelectHub” le da una calificación en cuanto a su precio de dos puntos sobre cinco puntos. (SelectHub, s.f.)

3.1.6 PDI y MicroStrategy

MicroStrategy es una compañía pionera en la parte de BI, según “SelectHub” es líder mundial de software de análisis. Esta herramienta ofrece soluciones de gran alcance y servicios expertos que ayudan a organizaciones en la correcta toma de decisiones.

Características de MicroStrategy faltantes en PDI:

- Ofrece un gran servicio de BI en la parte de aplicativos móviles.
- La aplicación provee un servicio de análisis sin conexión, autenticación multi factores, bioseguridad, GPS e integración con la cámara.

- Servicios Web.
- Conexión con alrededor de 100 bases de datos.
- Aplicaciones web con conectores ODBC.
- Amigable con el usuario, ayuda a crear nubes de datos sin la necesidad de conocimientos técnicos de programación o SQL.

En cuanto al licenciamiento de MicroStrategy. Existen distintos tipos de licencias de acuerdo con la necesidad del negocio, entre los planes se puede mencionar basado en la web con un precio de 600,00 dólares por usuario al año o 300 mil dólares por procesador, basado en móvil con un precio igual al anterior, servidor con un precio 1200,00 dólares por usuario o 600 mil dólares por procesador y, por último, en la versión On-Premise su precio es de 5.000 dólares por usuario al año. (SelectHub, s.f.)

3.1.7 PDI y Power BI

Power BI pertenece a la compañía de Microsoft al igual que Microsoft SL Server Reporting Services. La herramienta Power BI se divide al igual que todas en sub-herramientas, las cuales son Power Query se usa para los procesos ETL, Power Pivot es principalmente usado para el modelo y el análisis de datos, Power View and Map herramienta usada para visualizar la información, Power BI Desktop brinda una interfaz gráfica para el escritorio y Power BI Services que provee el servicio de BI basado en la nube. (Mamani, 2018)

Las características de Power BI faltantes en PDI son:

- Visualizaciones de mapas empoderados por Bing Maps.
- Actualización de los datos.
- Servicios de Cloud como Cortana, provee resultado de las consultas verbales.
- Arquitectura en chip, ofrece una balanceada simplicidad y rendimiento.
- DAX (Data Analysis Expressions), sirve para crear columnas con cálculos y medidas, similar a Microsoft Excel.

- Función Q&A (Question and Answer) es una capacidad y beneficio top en el logro de autoservicio.
- Power BI posee actualizaciones constantes que brindar innovación en la herramienta.

En cuanto al licenciamiento, por parte de “SelectHub” posee una calificación de dos puntos sobre cinco puntos. Está orientado para todo tipo de tamaño de empresas. En cuanto a costos se mencionan los siguientes, se empieza por una versión gratuita la cual posee menos características, seguido de la versión Pro la cual tiene un costo de 9,99 dólares al mes por usuario. Por último, la versión Premium la cual tiene un costo de 4.995,00 dólares al mes por nodo; 24.975 dólares al mes para 5000 usuarios con tres nodos. (SelectHub, s.f.)

Al tener en cuenta toda la información sobre las herramientas de BI mencionadas en los puntos anteriores, se llega a la conclusión de que la herramienta PDI de Pentaho, que es de código abierto, ofrece una menor cantidad de servicios a los que las distintas compañías pioneras ofrecen. Sin embargo, a pesar de que Pentaho no posea todas las características se puede decir que sus características ofrecen la cantidad de funciones necesarias para cumplir con los objetivos del presente trabajo de titulación.

Al enfocarse en la parte académica, se observa que el requerimiento financiero es limitado, por lo tanto, hay que optar por conseguir una herramienta sin costo y que brinde la mayor cantidad de funciones. PDI es la herramienta que cumple con las necesidades debido a sus excepcionales funcionalidades y sin costo alguno.

Se escoge PDI porque brinda grandes beneficios sin fines de lucro. En cuanto a beneficios se tiene lo siguiente:

- Es de código abierto, es decir, existe una versión gratuita.
- Ofrece el servicio en tiempo real.
- Análisis de predicciones para la toma de decisiones.

- Característica de arrastrar y soltar en el diseño.
- Amplia compatibilidad con fuentes de datos.
- Interfaz intuitiva y herramientas interactivas.
- Proceso de ETL.
- Acceso e integración de datos desde Hadoop a Excel.
- Minería de datos.
- Encuentra anomalías y patrones.
- Integración de BA con Google Maps.
- Reportes Ad Hoc.
- Proceso OLAP.

Con las características mencionadas se llega a tener una idea clara, es decir, al ver que las funcionalidades que ofrece PDI en comparación con las otras herramientas son similares, y sin costo alguno, se llega a la decisión de tomar a PDI como la herramienta para la implementación del trabajo de titulación. (Finances Online, 2019)

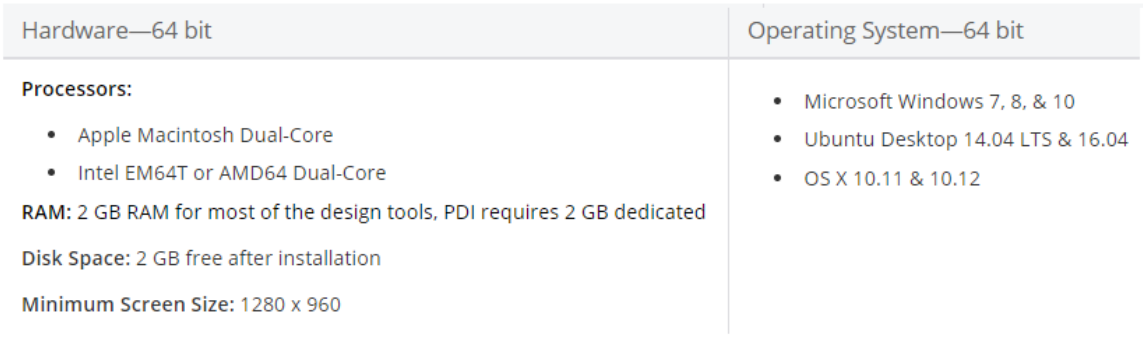
3.2 Instalación de herramientas

Después de un análisis, se llega a realizar una selección de herramientas de acuerdo con los objetivos del trabajo de titulación. Se toma en cuenta cada una de las características y se procede con la selección y posterior instalación la cual está detallada en los puntos siguientes.

3.2.1 Instalación de Pentaho

Para que se pueda realizar la instalación del PDI de forma exitosa es necesario cumplir con los requerimientos tanto de hardware, como de software. Los requisitos se muestran en la figura 3.

Dentro de los requisitos de hardware se tiene:



Hardware—64 bit	Operating System—64 bit
Processors: • Apple Macintosh Dual-Core • Intel EM64T or AMD64 Dual-Core **RAM:** 2 GB RAM for most of the design tools, PDI requires 2 GB dedicated **Disk Space:** 2 GB free after installation **Minimum Screen Size:** 1280 x 960	• Microsoft Windows 7, 8, & 10 • Ubuntu Desktop 14.04 LTS & 16.04 • OS X 10.11 & 10.12

Figura 3. Requisitos de hardware de PDI.

Tomado de (Hitachi, s. f.)

Por otro lado, los requisitos de software serán instalados y detallados a continuación.

3.2.2 Instalación de Java

Pentaho requiere de la instalación de Java Runtime Environment (JRE) desde la versión 7 o superior para que los programas de Java se puedan ejecutar, para este proyecto de titulación se instalará la versión 8u211.

Lo que se necesita es el archivo ejecutable (.exe) que se encuentra disponible en la página web de Oracle, y se procede a descargar la versión de Windows. (Oracle, s. f.)

Una vez descargado el archivo hay que ejecutarlo para que aparezca su instalador y comenzar el proceso de instalación.

En próximo paso en la instalación, en caso de ser necesario, podría ser cambiar la carpeta de destino para posteriormente avanzar hasta que el proceso de instalación finalice y cerrar la ventana.

3.2.3 Instalación de Tomcat

Para su instalación es necesario descargar el instalador disponible en la página web de Apache, de donde se escoge el ejecutable llamado Windows Service Installer. (Apache Tomcat, s. f.)

Una vez que se abre el instalador comienza el proceso de instalación para navegar a través de las ventanas hasta llegar a la configuración básica donde se configura lo siguiente, ver figura 4:

- Puertos de los servicios
- Nombre del servicio de Windows
- Cuentas y roles

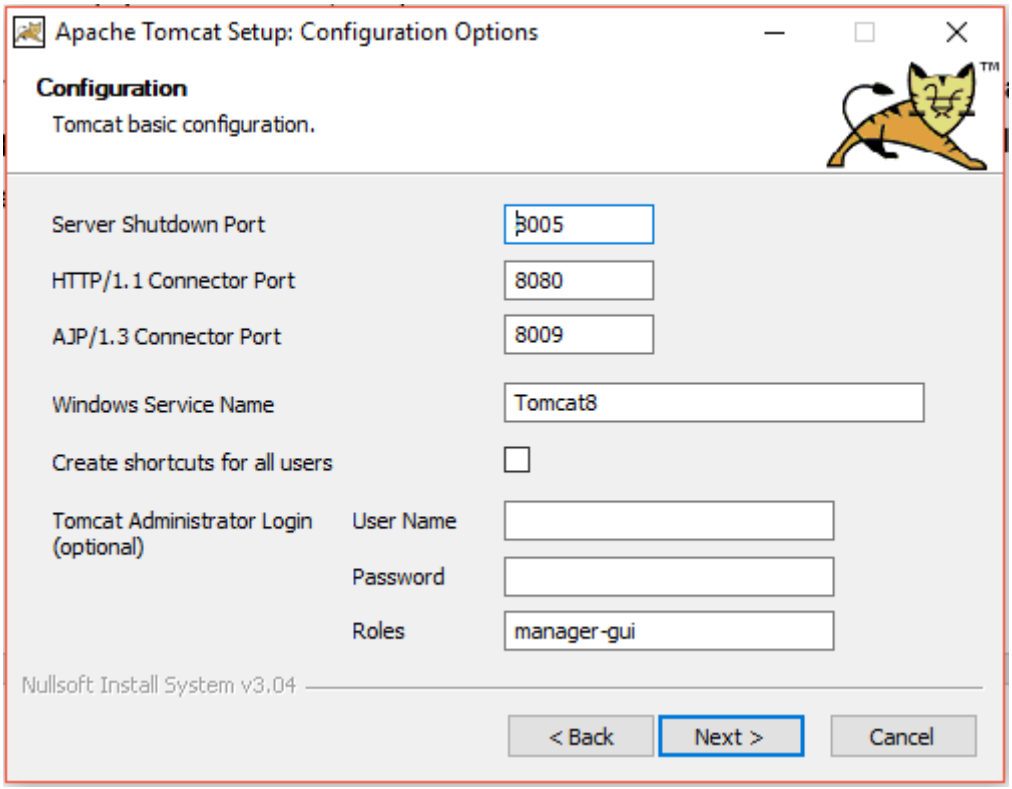


Figura 4. Configuración básica de Tomcat.

Posterior a la configuración básica se puede cambiar la carpeta de destino de Tomcat y finalizar la instalación.

3.2.4 Instalación de PDI

El instalador de PDI se encuentra disponible en su página web Hitachi Vantara y desde ahí se lo descarga para posteriormente comenzar su instalación.

Una vez que se tiene el instalador descargado, hay que ejecutarlo para empezar el proceso de instalación. En las siguientes ventanas únicamente se da clic en Next hasta llegar a la ventana donde se escoge el lugar o la compartición donde se quiere instalar PDI. En la próxima ventana se crea un usuario para la base de datos Postgres que Pentaho la utiliza para almacenar reportes, usuarios y otra información del sistema. Hay que crear una contraseña como se muestra en la figura 5.

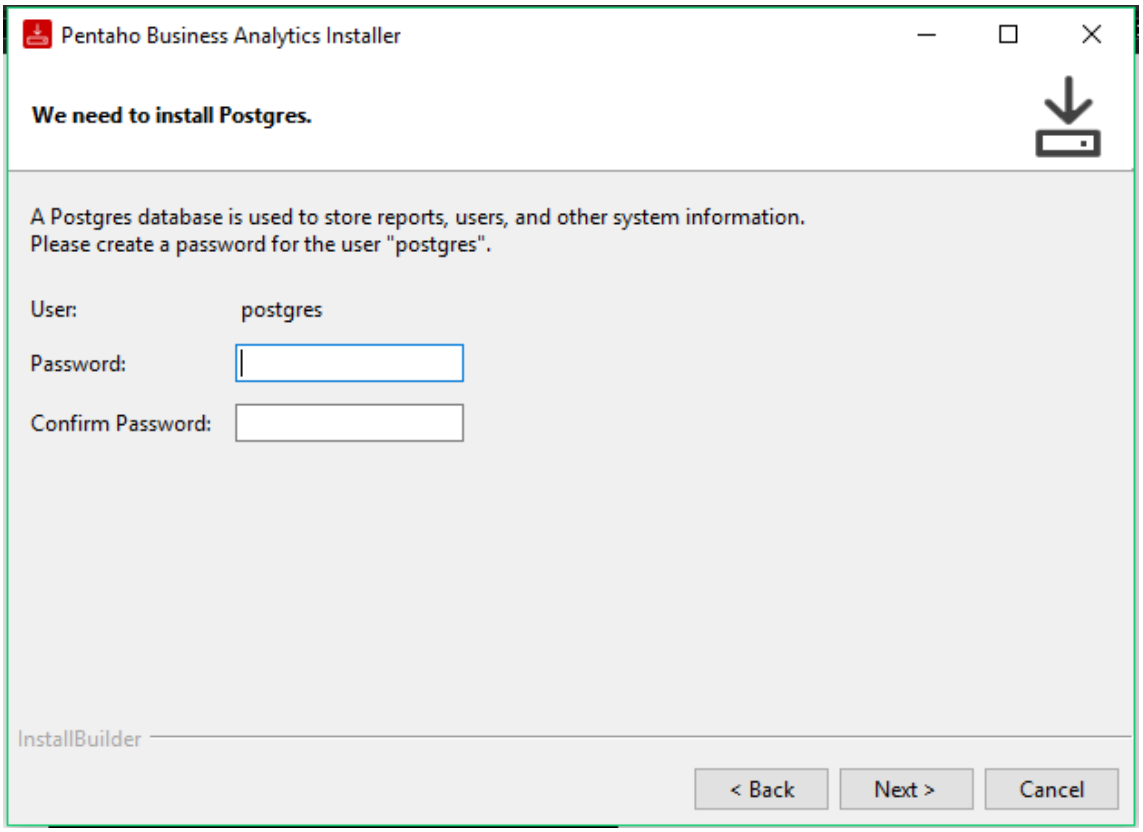


Figura 5. Creación del usuario Postgres.

Finalmente hay que esperar que la instalación finalice para luego iniciar el software que ahora se encuentra listo para entrar en operación.

3.2.5 Instalación de PhpMyAdmin

La instalación del servicio PhpMyAdmin se la realiza en la máquina virtual que tiene como sistema operativo Centos 7. En la parte de la instalación se descargan diferentes paquetes tales como:

- Httpd. Servidor apache.
- Mariadb-server. Motor de base de datos.
- Phpmyadmin. DBMS.
- Epel-release. Paquetes de alta calidad para la línea empresarial de Linux, especialmente Fedora (Fedora Wiki, s. f.).

Una vez que se instalan todos los paquetes es necesario configurar el servicio PhpMyAdmin para que se pueda acceder a él de manera remota, el archivo que se debe modificar se encuentra en la ruta /etc/httpd/conf.d/phpMyAdmin.conf y su modificación es mostrada en la figura 6.

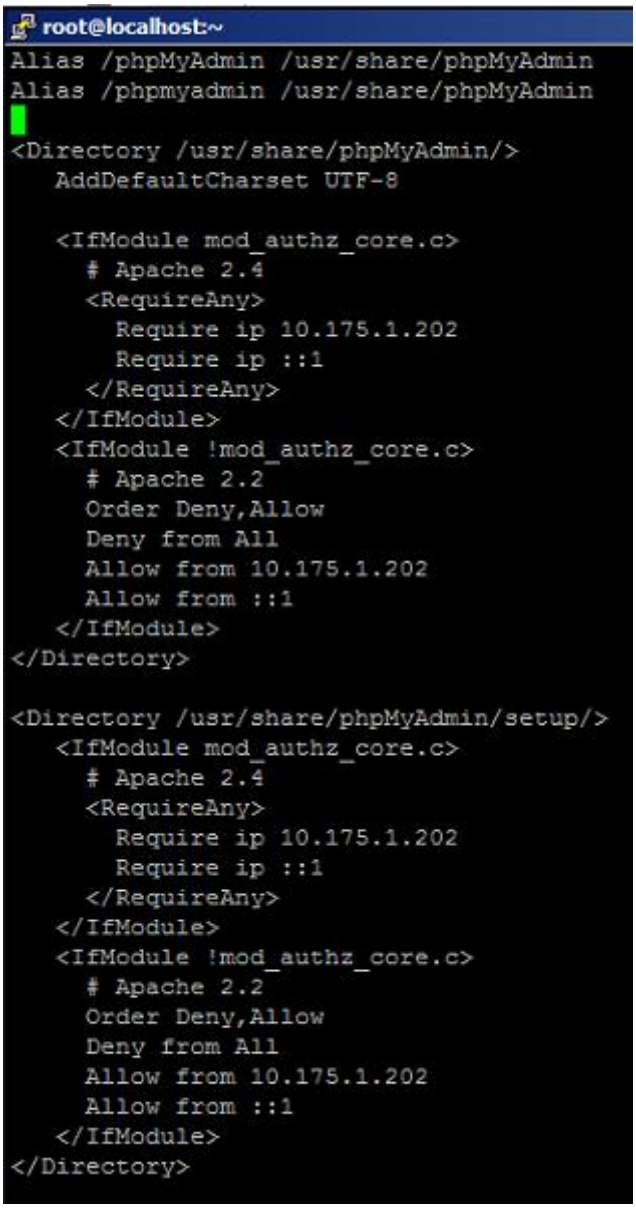


Figura 6. Configuración de PhpMyAdmin para conexiones remotas.

Originalmente en este archivo las direcciones IP que vienen por defecto son las de su tarjeta de red, es decir, la dirección IP 127.0.0.1; dicha dirección IP es reemplazada por la dirección IP del servidor desde el cual se va a acceder al servicio de PhpMyAdmin que, en este caso, es la dirección 10.175.1.202 (ver la figura 6) y se cierra el archivo de configuración guardando los cambios realizados. Finalmente, se reinicia el servicio httpd para que los cambios realizados surtan efecto, y posteriormente se inicia el servicio de mariadb.

Una vez terminada la instalación, es necesario crear los usuarios de la base de datos para poder acceder a la misma, la cual se la realiza dentro del servicio de MySQL.

Para finalizar, se procede a comprobar que el servicio PhpMyAdmin esté levantado y se pueda acceder al mismo. Para comprobar que el servicio esté corriendo, en la ventana de un navegador se digita la dirección IP del servidor que está alojando el servicio seguido del nombre de este; y para comprobar que se pueda acceder se escribe el usuario y la contraseña creadas previamente que, en caso de que el acceso sea exitoso, se mostrará la pantalla principal del servicio.

3.2.6 Instalación de Weka

Para instalar el software Weka es necesario descargar su instalador desde la página sourceforget.net, y una vez que finalice su descarga ejecutar el archivo. La interfaz de instalación es bastante amigable e intuitiva ya que los requisitos que pide para su instalación ya se encuentran en la máquina virtual como lo es el JRE.

Finalmente, luego de pasar por varias ventanas que no requieren configuración, por el contrario, son informativas se llega a la finalización de la instalación para que la herramienta sea utilizada en el próximo capítulo.

3.3 Selección de las fuentes de datos

Las estructuras de datos más importantes que serán usadas en un sistema ETL son las siguientes:

3.3.1 Archivos planos

En algunos de los casos, los datos no se encuentran almacenados en un DBMS, sino que se manipulan archivos de texto. Un archivo es plano cuando los datos son almacenados en columnas y filas dentro de un archivo en el sistema

simulando a una tabla de una base de datos, también se lo conoce como archivo secuencial.

Si el sistema operativo que se usa es UNIX o Windows, los datos de los archivos planos siguen el estándar American Standard Code for Information Interchange (ASCII), y gracias a ello pueden ser manipulados por herramientas ETL o lenguajes de scripts como si correspondieran a tablas dentro de una base de datos.

Entre las ventajas de usar archivos planos en lugar de DBMS se encuentran que tareas como clasificación, eliminación, actualizaciones, combinaciones y migraciones se pueden realizar de manera mucho más rápida en entornos externos a DBMS ya que los metadatos no siempre son necesarios en el almacenamiento de la información. Pero, se debe tener en cuenta que, si se trata con archivos planos y scripts, es recomendable separar las tablas de metadatos de las transformaciones del proceso ETL.

Decidir si usar archivos planos en lugar de tablas de bases de datos para ejecutar un proceso ETL puede ser difícil por depender de distintas variables, sin embargo, se mencionarán casos en los que se podría usar archivos planos:

- Almacenamiento de datos para recuperación y protección. Cuando se comienza el proceso de extracción es recomendable ingresar al sistema, seleccionar la información de la fuente y ubicarla en un archivo plano; esto para que en caso de que falle el proceso de extracción no se tenga que ingresar continuamente al sistema para reiniciar el proceso, sino que se escoge la información del archivo plano.
- Organización de datos. Es más eficiente organizar la información en el sistema en lugar de seleccionar de una tabla y utilizar una sentencia como order by para extraer. Una de las características de un proceso ETL es integrar datos aislados y combinarlos de manera eficiente, de esta manera el procesamiento será menor para el ETL.

- Filtrado. Existen casos en los que la información de una fuente de datos no tiene índices entonces lo que generalmente se hace es forzar a la fuente para asignar índices a las filas y después filtrar una sentencia WHERE para filtrar la información. La opción más eficiente es extraer los datos a un archivo plano y utilizar el comando grep (en un sistema UNIX) para filtrar las filas que cumplan con los requisitos, de esta manera, se logra un filtrado mucho más eficiente y fuera de la base de datos.
- Actualización de texto. Para actualizar campos dentro de una base de datos se tiene una alternativa, la cual consiste en sacar la información a un archivo plano y con el comando tr hacer el reemplazo de strings de manera rápida y eficiente, sin necesidad de entrar en actualizaciones o funciones de una base de datos.
- Referencias. En casos de que la base de datos utilice una tabla específica para servir como base a otras tablas (en bases de datos normalizadas), lo más eficiente es extraer dicha tabla de referencias y ubicarla por una vez en el área de almacenamiento y a partir de ahí la herramienta ETL se encarga de almacenarla en memoria mientras el proceso tenga vida. (Kimball & Caserta, 2004)

3.3.2 Datos XML

Los conjuntos de datos XML dentro de un sistema ETL sirven como formato para entrada y salida de datos del sistema, además, podrían ser usados también como almacenamiento de datos persistentes tanto en el ETL como en el DW.

El lenguaje XML sirve para el intercambio de datos, parte de un archivo plano pero la diferencia está en que contiene metadatos y no información para dar formato a los datos. Por otro lado, la diferencia entre XML y HTML está en que HTML contiene información para dar formato a los datos, pero no tiene metadatos. Entender la diferencia entre XML y HTML es sumamente importante para ver la manera en que se podría afectar el DW.

Los metadatos XML son etiquetas que identifican cada ítem de un documento XML, pero sin que sea ambiguo.

XML tiene la capacidad de declarar estructuras jerárquicas en forma compleja que no se agregan directamente a un estándar de dos dimensiones dentro de tablas relacionales. Cuando el DW recibe un conjunto de datos XML. Habrá un proceso complejo de extracción para enviar los datos de forma permanente a un DW relacional. Lo ideal sería que las bases de datos relacionales pueden soportar estructuras jerárquicas XML, lo que implica que la sintaxis y semántica de las bases de datos relacionales SQL se extienda, lo que no ha pasado hasta ahora.

Una de las ventajas es que XML es muy efectivo para mover datos entre sistemas que son incompatibles, esto debido a que se provee suficiente información para poder crear las tablas dentro de una base de datos relacional y después se pueda esparcir esta table con los datos apropiados, es por esto por lo que XML es un lenguaje universal para compartir datos. (Kimball & Caserta, 2004)

3.3.3 Tablas relacionales

Otra de las opciones que se tiene para el almacenamiento de la información es utilizar un DBMS relacional, especialmente cuando no se tiene una herramienta ETL dedicada. Entre las ventajas de utilizar un DBMS se tiene:

- Aparentar metadatos. El problema de los archivos planos es que no tienen metadatos, y para solucionar este problema, se utiliza un DBMS que contiene metadatos técnicos y automáticos dentro de sus tablas.
- Técnicas relacionales. En un ambiente relacional los datos y sus relaciones entre las tablas son más fáciles de entender.
- Repositorios abiertos. Los datos pueden ser accedidos fácilmente si se encuentran en un DBMS por cualquier herramienta SQL, esto resulta ser importante en el control de calidad.

- Soporte DBA. En los ambientes de trabajo, el grupo de DBA se encarga únicamente de la información dentro del DBMS, pero la información que se encuentra afuera no se la toma en cuenta por lo que el equipo encargado del desarrollo del ETL debe enfocarse también en temas como respaldos, recuperación o seguridad cuando el almacenamiento no estará en un DBMS.
- Interfaz SQL. Muchas de las veces se necesita manipular la información y para esto se utiliza el lenguaje SQL, que es el más conocido en el ambiente de TI. La mayoría de las bases de datos ya vienen incorporadas con funcionalidades para que la manipulación se pueda realizar a través de sentencias SQL, razón por la cual el almacenamiento se lo hace en una base de datos. (Kimball & Caserta, 2004)

3.3.4 Tablas DBMS independientes

Si se decide almacenar los datos de prueba en un DBMS, se puede decidir entre distintas arquitecturas para el esquema, e incluso mezclas ya que un ambiente de datos de prueba necesita insertar y extraer datos de manera eficiente.

El uso de tablas independientes hace que el desarrollo y la implementación de la base de datos sea mucho más simple. Las tablas independientes no tienen relaciones con ninguna otra tabla de la base de datos y, es por esto, por lo que son fuertes candidatos para almacenar información fuera de una base de datos tipo relacional. Por ejemplo, en DW pequeños basta que los datos sean ubicados en tablas independientes en lugar de una base de datos relacional ya que es solo un ambiente de pruebas.

El enfoque de las tablas independientes no va hacia reducir el espacio necesario o a mejorar el desempeño de la base de datos, por el contrario, basta con que tenga una definición lógica y un indexado apropiado en todas las tablas independientes, por lo que los procesos ETL son los únicos que utilizan este tipo de tablas. (Kimball & Caserta, 2004)

3.3.5 Fuentes de datos no relacionales

El principal problema de los desarrolladores de herramientas ETL es que deben integrar diferentes fuentes de datos heterogéneas porque el DW expande su visión para incluir más y más datos de diferentes áreas, que muchas de las veces no tienen relación alguna.

La primera solución para el problema fue insertar toda la información dentro de un DBMS, pero al pasar el tiempo se observó el potencial de las herramientas ETL para manipular datos heterogéneos y con esto se pudo evitar almacenar toda la información en una sola base de datos. En la figura 7 se puede ver cómo una plataforma ETL puede integrar diferentes fuentes de datos heterogéneas desde sus almacenamientos nativos para luego migrarlos al DW, el ETL está asociado tanto a la base de datos de prueba como a los archivos externos en caso de que sea necesaria una manipulación.

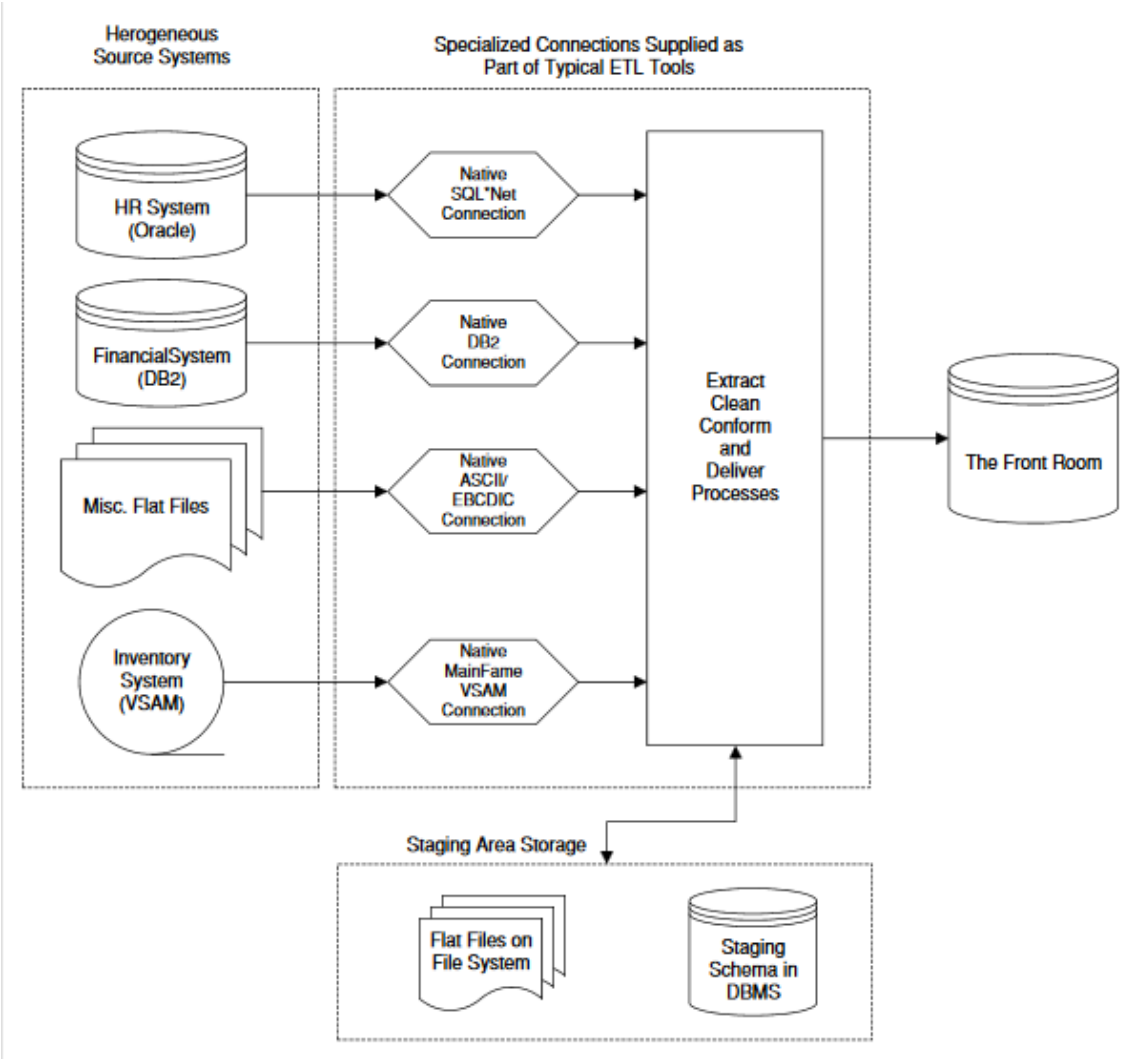


Figura 7. Integración de datos heterogéneos en un proceso ETL.

Tomado de (Kimball & Caserta, 2004)

Una vez que se han incluido todas las fuentes de datos dentro de una base de datos es necesario hacer una revisión de la integridad de los datos, lo que requiere que una parte personalizada del proceso ETL establezca reglas que sean similares a la naturaleza de una base de datos relacional. Posteriormente, ciertas tablas se convertirán en padres y otras en hijos, las tablas que sean padres no podrán ser eliminadas a menos que todos sus hijos desaparezcan, o sino los hijos vendrían a ser huérfanos; dentro de este ambiente quiere decir que una clave primaria no puede ser eliminada si existen claves foráneas que hacen

referencia a otras tablas, en caso de existir huérfanos se puede decir que existe fallas en la integridad de la base de datos.

Las fuentes de datos no relacionales no aseguran la integridad referencial, y un sistema no relacional es una colección de tablas independientes. Al transcurrir el tiempo se ha comprobado que las bases de datos que no han forzado la integridad se han visto comprometidas en temas de seguridad, y también es garantizado que tendrán problemas de calidad en la aplicación final. (Kimball & Caserta, 2004)

3.4 Proceso ETL

Proceso ETL, es conocido por la extracción, transformación y carga de la información. Tiene como función principal extraer la información de una o varias bases de datos, limpiar y optimizar la información; y finalmente, volverla a cargar dentro de un DW. En la figura 8 se encuentra ilustrado cómo funciona el proceso de un ETL.

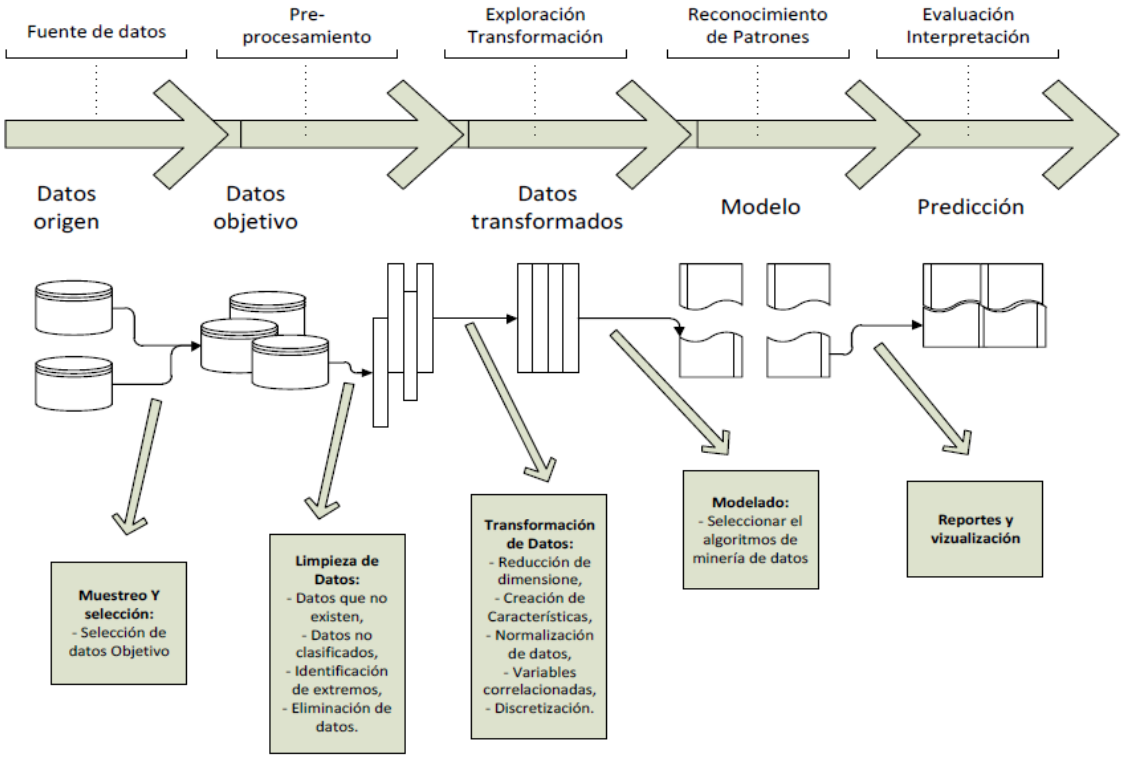


Figura 8. Proceso ETL.

Tomado de (Hidalgo, 2016)

3.4.1 Extracción

Se empieza por la extracción de información y se toma en cuenta el origen de esta. Al decir origen, se hace referencia a la selección de la fuente de información que se encuentra en el apartado de este capítulo. Las fuentes de información pueden ser una o varias y cada una de ellas tiene distintas características, esto requiere ser administrado con un orden para poder tener una extracción efectiva del proceso ETL. El proceso necesita integrar efectivamente los sistemas que tienen diferentes plataformas, como un sistema que administre diferentes tipos de bases de datos, sistemas operativos o protocolos de comunicación.

En la extracción de información con diferentes fuentes de datos, el proceso ETL debe ser consciente de usar los controladores correctos en las fuentes de base de datos, entender la estructura de la información entrante y saber manejar las fuentes con diferentes naturalezas como si se tratara una fuente principal. El

proceso de extracción tiene dos fases; extracción inicial y extracción de datos modificados. En la extracción inicial se obtiene la información desde diferentes fuentes operacionales para ser cargadas dentro del DW. Este proceso se lo realiza una sola vez después de construir el DW para una posterior prueba con grandes cantidades de información que provienen de las fuentes de datos. Para la extracción de datos modificados se entiende que es la captura de datos modificados, en donde los procesos ETL actualizan al DW con los datos modificados y añadidos en las fuentes desde la última extracción. El proceso es periódico, es decir, se repite mientras ocurran actualizaciones o depende de las necesidades del negocio, sin embargo, solo captura los datos modificados desde la última extracción por medio de distintas técnicas como columnas de auditoria, registro de base de datos, fecha del sistema, entre otros. (Kimball & Caserta, 2004)

3.4.2 Transformación

El segundo paso del proceso ETL es la transformación de datos. La etapa de transformación tiende a realizar la limpieza y conformación de la información proveniente para ganar precisión sobre la información la cual es correcta, completa, consistente y no ambigua. Este proceso tiene como características el limpiar, transformar y cargar la información. Define la granularidad, es decir, se establece un nivel jerárquico dentro de las tablas en las cuales se puede entender como la granularidad mínima al nivel más bajo en las tablas de hechos, tablas de dimensiones, esquemas de DW, hechos derivados, lento cambio de dimensiones y tablas sin hechos. Todas las reglas establecidas por las transformaciones y los esquemas entregables se describen en el repositorio de metadatos. (Shaker, Adeltawab, & Ali, 2011)

3.4.3 Carga

Cargar la información en el DW es el último paso del proceso ETL. En este paso, la información que se extrae y transforma es almacenada dentro del DW con el fin de que los usuarios finales y las aplicaciones puedan acceder. El paso de

carga incluye ambos casos cargar las tablas de dimensiones y las tablas de hechos. (Shaker, Adeltawab, & Ali, 2011)

3.5 Diseño de data warehouse

Un DW es una base de datos la cual contiene información para los usuarios, sin embargo, hay que mencionar que esta información está optimizada de acuerdo con el cumplimiento del proceso ETL. Esta base de datos ordena y almacena la información importante para su futuro uso en un proceso de análisis con un determinado tiempo.

El desarrollo de un DW facilita el acceso a la información útil, es decir, gracias a todos los procesos que lleva dentro responde de mejor forma a las consultas en comparación con una base de datos tradicional. Se caracteriza por ser no volátil, es decir, que toda su información cargada no puede ser modificada, ni eliminada, y se puede mencionar de igual manera que un DW es diseñado de acuerdo con lo que el negocio necesite. Sin embargo, se menciona que para la construcción de un DW se requiere de una serie de iteraciones y hay que tomar en cuenta cuatro decisiones claves para el modelado de una iteración (Hidalgo, PROYECTO DE DETECCIÓN DE PATRONES, 2016).

Las cuatro decisiones claves son:

- Escoger un proceso de negocio.
- Declarar la granularidad
- Determinar las dimensiones
- Determinar los hechos

Las aplicaciones de DW, bases de datos multidimensionales y procesos OLAP brindan a distintas empresas, que a su vez almacenan grandes cantidades de datos, un proceso de toma de decisiones. Es ampliamente aceptado que todos los sistemas mencionados anteriormente son basados en modelos multidimensionales.

El modelado multidimensional estructura la información en base a hechos y dimensiones. Un hecho tiene medidas de un proceso de negocio como ventas, entregas, entre otros; mientras que una dimensión representa el contexto para el análisis de un hecho, que puede contener el producto, cliente o tiempo (Sergio, Juan, & Il, 2005). Los beneficios que un modelo multidimensional puede ofrecer son:

- La manera de analizar la información se aproxima a la de los analizadores de datos, lo que facilita a los usuarios un mejor entendimiento de la información.
- Sirve como una estructura para la predicción de las intenciones de los usuarios finales.

3.5.1 Modelo multidimensional

En el modelo multidimensional se puede mencionar que su información está basada en hechos y dimensiones. Donde un hecho es considerado un artículo de interés y es descrita a través de un grupo de atributos denominados en inglés como “measures” o “fact attributes”, los cuales son contenidos en celdas en el cubo de datos. Este grupo de medidas está basado en un grupo de dimensiones que determinan la granularidad adoptada para representar los hechos. Por otro lado, las dimensiones proveen el contexto en el cual los hechos son analizados. Además, las dimensiones son inclusive caracterizadas por los atributos, los cuales son a veces denominados en inglés “dimension attributes”. Se toma como referencia el siguiente ejemplo:

Se trata de una compañía que se relaciona con la venta de vehículos (carros y vans), por distintas provincias. El DW tiene tres *datamarts*, tales como venta de vehículos, parte de ventas y trabajos de servicio. Están separadas debido a que son usadas por los usuarios finales. Sin embargo, estos *datamarts* comparten algunas dimensiones en común como la concesión o el tiempo, a pesar de estos posean su propia dimensión, como un vendedor o servicio:

- Ventas de vehículos, considera la venta de vehículos.
- Ventas de partes, representa las ventas de las partes de vehículos como llantas de emergencia o focos.
- Servicio Técnico, considera los cambios de aceite o cambios de líquido de frenos.

Cada uno de estos modelos tienen sus correspondientes hechos los cuales contienen medidas específicas para ser analizadas. Además, ellos consideran las siguientes dimensiones para analizar medidas: concesión, tiempo, cliente, vendedor y automóvil para el *datamart* de ventas de vehículos; concesión, tiempo, servicio, mecánica y partes para el *datamart* de venta de partes; y concesión, tiempo, servicio y mecánica para el *datamart* de servicio técnico.

Como se aprecia en la figura 9 en la parte izquierda, se observa un cubo de datos comúnmente usado para representar un modelo multidimensional. Este cubo se basa en el datamart de venta de autos para analizar las medidas a lo largo de las dimensiones del automóvil, cliente y tiempo. Hay que tomar en cuenta las relaciones de las tablas; se tiene entendido que entre un hecho y una dimensión la relación es de muchos a uno. Sin embargo, usualmente estas relaciones son de muchos a muchos. Se entiende que “un cliente compra un auto en un determinado tiempo”. (Sergio, Juan, & Il, 2005)

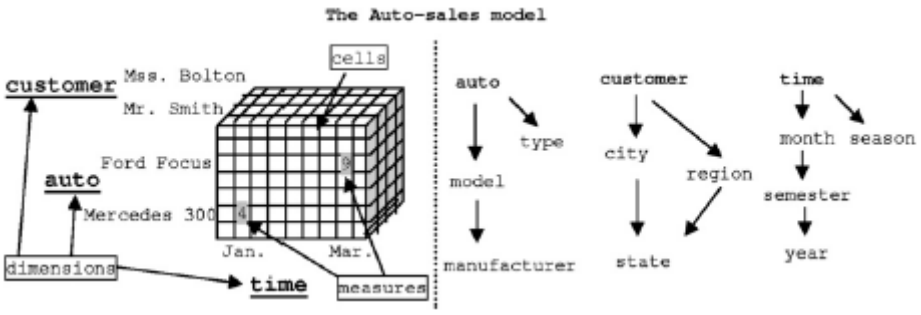


Figura 9. Modelamiento multidimensional.

Tomado de (Sergio, Juan, & Il, 2005).

3.5.2 Tipos de esquemas (Estrella y Copo de Nieve)

La figura 10 demuestra que un esquema tipo “estrella” posee una tabla de hechos de gran tamaño en el centro, con múltiples tablas de dimensiones que rodean la tabla de hechos. Existe una relación de uno a muchos entre la tabla de dimensiones y la tabla de hechos. La tabla de hecho representa en algunas ocasiones eventos y transacciones de negocios, o un resumen de transacciones/eventos. Debido a que las aplicaciones en el mundo real no conforman directamente una estructura en estrella, es posible que algunas tablas de dimensiones no estén en la tercera forma normal, cuando se habla de normalización. (Martyn, 2004)

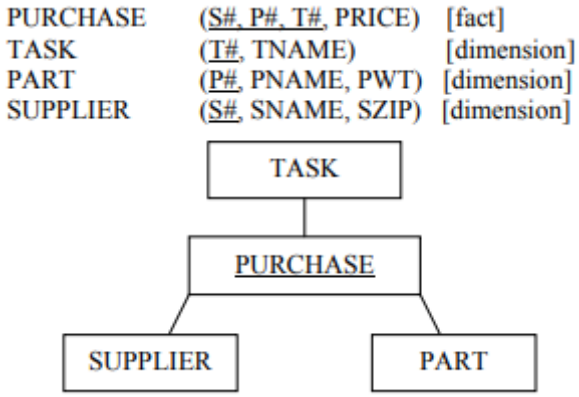


Figura 10. Esquema estrella.

Tomado de (Martyn, 2004).

En la figura 11 se ilustra el esquema “copo de nieve” que puede ser interpretado como una modificación del esquema en “estrella” antes mencionado. La figura 11 demuestra que en centro del esquema existe una estrella. La extensión principal es la presencia de tabla de dimensiones de “nivel externo”; resaltadas con contorno grueso. Existe una ruta entre las tablas externas con la tabla de hecho central relacionadas con una relación de uno a muchos, la cual representa una jerarquía dimensional. Se reduce la presencia de tablas de dimensión a nivel externo, pero necesariamente no se eliminan el número de tablas normalizadas. (Martyn, 2004)

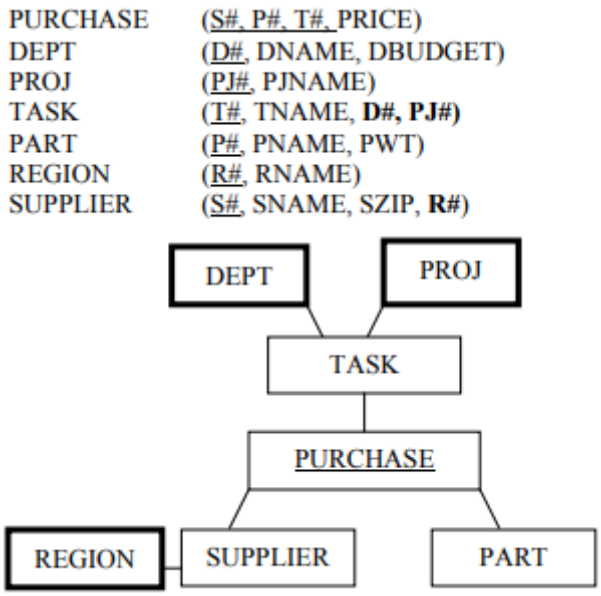


Figura 11. Esquema copo de nieve.

Tomado de (Martyn, 2004).

3.6 Generación de cubos OLAP

Un cubo OLAP puede ser definido como un conjunto de coordenadas que determinan un valor de medida. Las coordenadas también son llamadas dimensiones y pueden tener una estructura de forma jerárquica, lo que hace posible que el análisis de datos pueda ser realizado en diferentes niveles, lo que quiere decir que ítems de niveles inferiores pueden ascender a niveles superiores dentro de la estructura jerárquica y así el usuario analiza los datos a diferentes niveles.

Las jerarquías de las dimensiones tienen diferentes propiedades relacionadas a la sumarización. La sumarización está basada en tres condiciones:

- Separación de los grupos de atributos
- Integridad de la agrupación
- El tipo de atributo y la función de agregación

La base de datos OLAP es usualmente almacenada en una base de datos relacional usando el conocido esquema tipo estrella que consiste en una tabla de hecho y varias tablas dimensionales. Cada valor de medida es almacenado en una tabla de hecho con claves de dimensión mientras que las tablas dimensionales contienen datos de la dimensión. En este esquema cada dimensión tiene una tabla dimensional no normalizada, en caso de ser normalizada, se convierte en un esquema llamado copo de nieve. (Niinimaki & Niemi, 2009)

Un cubo es otro de los nombres que se le da al modelamiento dimensional donde cada cubo representa una tabla de hechos y varias tablas dimensionales. Este modelo es extremadamente útil para la generación de reportes y análisis de los datos alojados en una tabla de hechos. La generación de cubos virtuales aparece debido a la necesidad de relacionar, combinar o analizar información de diferentes cubos. Los elementos XLM que componen un cubo virtual se explican a continuación:

- CubeUsages. Especifica los cubos que van a ser importados al cubo virtual y contiene a elementos <CubeUsage>.
- CubeUsage. Especifica al cubo base que será importado al cubo virtual. También se puede definir una medida (VirtualCubeMeasure) para importaciones similares sin usar el CubeUsage.
- El atributo CubeName define el nombre que será asignado al cubo base.
- El atributo ignoreUnrelatedDimensions determina si las medidas del cubo base tendrán miembros sin uniones elevados a niveles superiores, su valor es falso por defecto.
- VirtualCubeDimension. Importa una dimensión de los cubos constituyentes. En caso de que el atributo CubeName no sea establecido, significa que se importa una dimensión compartida, y en caso de que dicha dimensión sea usada más de una vez en el mismo cubo, no habrá manera de determinar el uso de la dimensión compartida que se quiso importar.

- VirtualCubeMeasure. Importa una medida de uno de los cubos constituyentes y es importado con el mismo nombre. Para renombrar al cubo o implementar alguna fórmula se utiliza el elemento CalculatedMember.

Los cubos virtuales se usan en casos donde las tablas de hecho tienen diferentes granularidades, o diferentes dimensiones y los resultados deben ser presentados a personas que no lograrían entender el lenguaje técnico o que no comprenden la estructura de las tablas.

Las dimensiones compartidas se sincronizan automáticamente y las medidas de cada cubo se relacionan a este contexto. Las dimensiones que pertenecen a un solo cubo se las conoce como no conformadas. (Hitachi Vantara, 2016)

En el caso de que se quiera generar un esquema relacional con un sistema OLAP. Primero el sistema muestra la lista de posibles dimensiones al usuario, quien escoge algunas de ellas para ser incluidas en el análisis OLAP. Una vez que se tienen las dimensiones, se comienza a generar el esquema específico para la tabla de hecho y las dimensiones seleccionadas, generalmente tablas que tienen una asociación 1:1 para un esquema más simple. (Inoue, Amagasa, & Kitagawa, 2013)

3.7 Generación de dashboard

3.7.1 Pentaho Dashboard Designer

Crear un dashboard con la herramienta Dashboard Designer es tan simple como escoger una plantilla, tema o contenido que se quiere mostrar y puede contener cuadros, tablas de datos, URLs.

Dashboard Designer tiene diferentes filtros para un mejor control, que permite a los usuarios cambiar detalles escogiendo diferentes valores de una lista y controlar el contenido de cada panel, mejor conocido como enlaces de contenido, ver la figura 12.

En la figura 12 se puede ver diferentes ítems que serán detallados a continuación:

- Vista Opened. Muestra botones para acceso rápido en la parte superior para crear y guardar reportes y dashboards.
- Panel de avisos. Muestra la forma para agregar más filtros a las partes individuales del dashboard.
- Panel de búsqueda de carpetas y archivos. Ubica los archivos y los agrega al dashboard.
- Dashboard canvas. Muestra una vista de manera dinámica del dashboard con el que se está trabajando, además, se va actualizando a medida que se sigue aumentando contenido de los paneles previamente mencionados.
- Panel de objetos. Mejora la vista del dashboard con paneles de objetos ya que permite agregar una plantilla o cambiar nombres de cada objeto en el dashboard.

Figura 12. Ejemplo de dashboard con Dashboard Designer.

Tomado de (Hitachi Vantara, 2017)

3.7.1.1 Crear un dashboard

El primer paso es iniciar sesión en la consola de usuario, a continuación, seguir los pasos mencionados:

- Desde la página principal, clic en Crear Nuevo y luego seleccionar Dashboard.
- Al final de la página, clic en la ventana de Propiedades e ingresar un nombre para el dashboard. El nombre ingresado va a aparecer en la esquina superior izquierda del dashboard y sirve para identificarlo si se quiere editar después.
- Clic en Plantillas para escoger una, por defecto aparecerá un fondo en blanco.
- Clic en Tema, el tema seleccionado será el que se aplique al dashboard.

3.7.1.2 Agregar un reporte desde Report Designer

Los pasos para agregar un reporte serán detallados a continuación

- Seleccionar un panel de Dashboard Designer.
- Clic en Insertar y escoger un archivo.
- Seleccionar el archivo deseado.
- Clic en Seleccionar para ubicar el reporte dentro del panel correspondiente. El nombre del reporte se muestra debajo del contenido, en caso de que existan parámetros obligatorios en blanco, el reporte se mostrará vacío. (Hitachi Vantara, 2017)

3.7.2 Pentaho Report Designer

Es otra de las herramientas sofisticadas, pero puede ser utilizada de forma independiente, o como parte de una distribución más grande de Pentaho. Con esta herramienta se puede crear dashboards más detallados, con calidad de

impresión preparados de cualquier fuente de datos. Report Designer es una de las varias formas que ofrece Pentaho para la creación de reportes.

Para ejecutar esta herramienta basta con digitar Report Designer en el menú de inicio, y una vez que se esté ejecutando crea un directorio “.pentaho” en el directorio principal y lo llena de información con subdirectorios y archivos.

3.7.2.1 Crear un reporte

Continuar con los siguientes pasos:

1. Conectarse a una fuente de datos o a los archivos que se van a utilizar.
2. Extraer los datos con una consulta.
3. Organizar los datos en el área de trabajo.
4. Aplicar un formato a los ítems del informe.
5. Agregar elementos gráficos.
6. Crear campos con fórmulas o cálculos usando datos extraídos con la consulta.
7. Publicar el reporte, ya sea en el servidor de Pentaho o localmente como archivo PDF.

En la mayoría de los casos, el reporte va a constar de los datos extraídos de la base de datos con la consulta que se crea previamente. Una vez que se tiene un conjunto de datos, se puede especificar detalles y luego moverlos a una plantilla de reportes o diseño. (Hitachi Vantara, 2018)

4. Capítulo IV. Implementación a un caso específico.

Para el caso específico se utilizará la información de tres cursos de la Universidad, que serán cargados al Moodle instalado y posteriormente se va a buscar indicadores, con la ayuda de la herramienta Pentaho Data Integration, que permitan mejorar la calidad educativa. Para tener una idea más clara de los elementos que van a intervenir en todo el caso específico observar la figura 13.

4.1 Infraestructura

Como muestra la figura 13, para la implementación del caso específico se utilizará tres servidores virtuales que están alojados dentro del centro de datos de la Universidad de las Américas. Dos de ellos tienen como sistema operativo CentOS y servirán para alojar las bases de datos tanto del Moodle, como del DW; por otro lado, la tercera máquina virtual tiene como sistema operativo Windows Server 2008 R2, que es donde está instalado PDI que es la herramienta que ayudará en el proceso ETL.

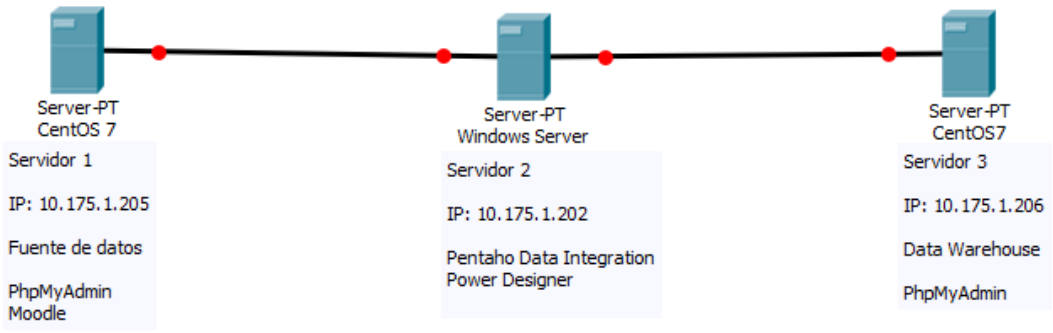


Figura 13. Arquitectura del caso específico.

4.1.1 Servidor 1

Una vez que toda la información de los tres cursos sea cargada al Moodle también aparecerá en una serie de tablas en la base de datos llamada Moodle (figura 14), la misma que servirá como fuente de datos para iniciar con la etapa de extracción en el proceso ETL en el servidor 2.

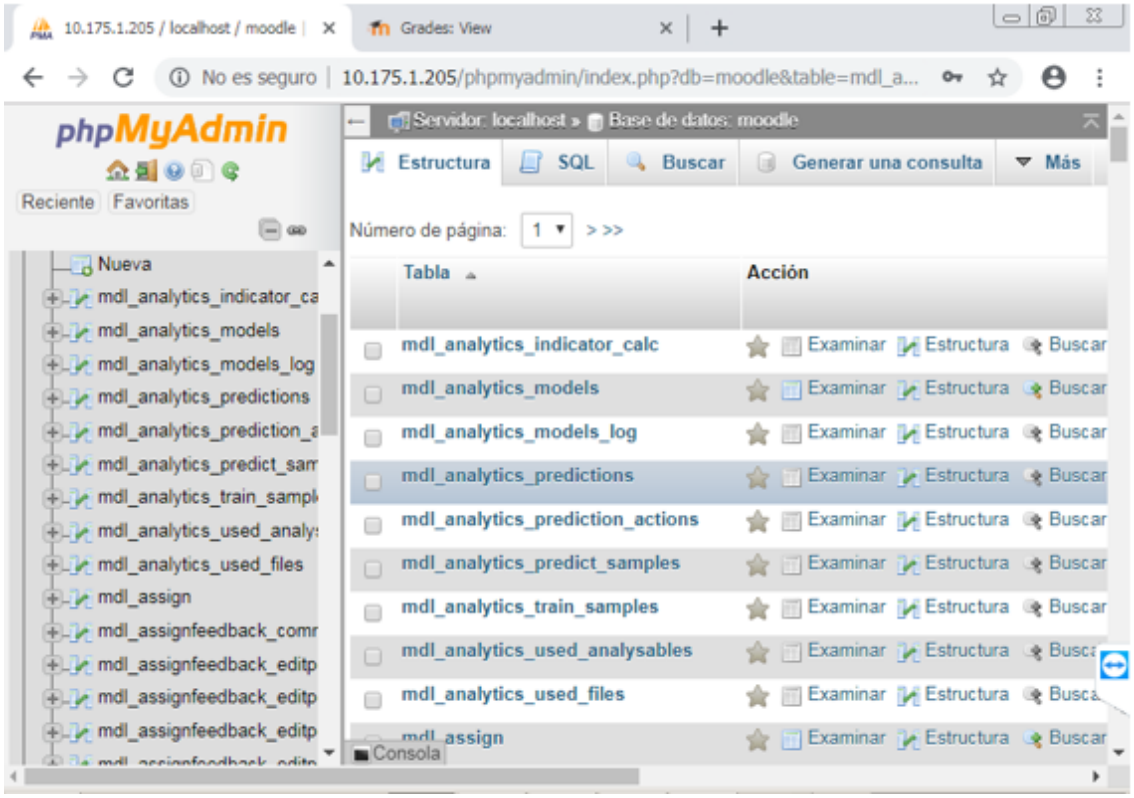


Figura 14. Base de datos generada por Moodle.

4.1.2 Servidor 2

En este servidor se encuentra instalada el software PDI, y es desde aquí donde se comenzará el proceso ETL. También con la herramienta DBDesigner se va a obtener el esquema de la base de datos previamente mencionada para tener una mejor comprensión.

4.1.3 Servidor 3

Una vez que la etapa de transformación esté terminada y los datos sean filtrados, comienza la etapa final del proceso ETL que es la carga de la información a un DW, la base de datos para el DW está lista para ser alojada en el servidor 3.

4.2 Conexión PDI a fuente de datos

Para establecer una conexión con la base de datos es necesario descargar el driver correspondiente a la misma. En este caso el driver correspondiente a la base de datos MySQL es llamado ConnectorJ y se lo puede encontrar en su página oficial. (Hitachi Vantara, 2018)

Una vez descargado el Driver es necesario moverlo al directorio “C:/ pentaho/jdbc-distribution” y luego abrir la terminal para redistribuir el driver en todas las ubicaciones necesarias para un correcto funcionamiento del software. En la terminal hay que ubicarse en el directorio antes mencionado y ejecutar el archivo .bat llamado distribute files.bat seguido del nombre del driver descargado, como muestra la figura 15. (Hitachi Vantara, 2016)

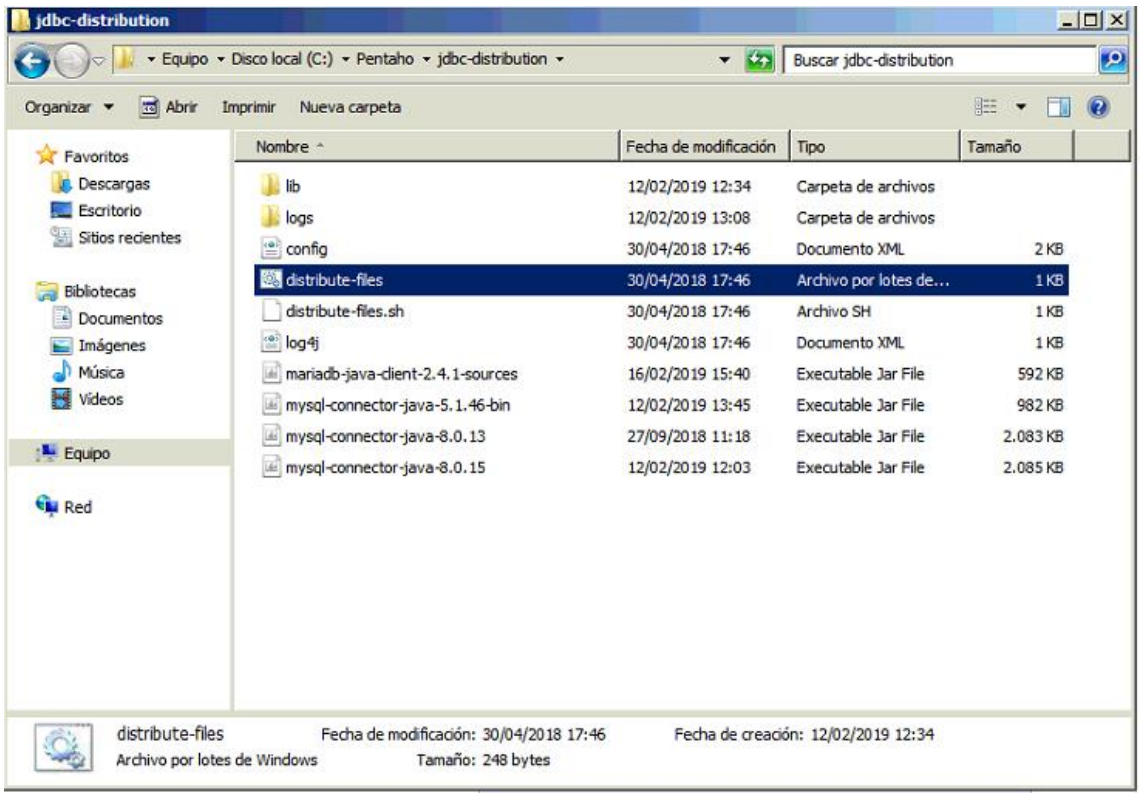


Figura 15. Distribución del driver ConnectorJ.

Finalmente, dentro del PDI se abre la ventana para crear la nueva conexión. Posteriormente se ingresa información como el Host Name, nombre de la base

de datos, número de puerto, usuario y contraseña para probar y establecer la conexión, véase la figura 16.

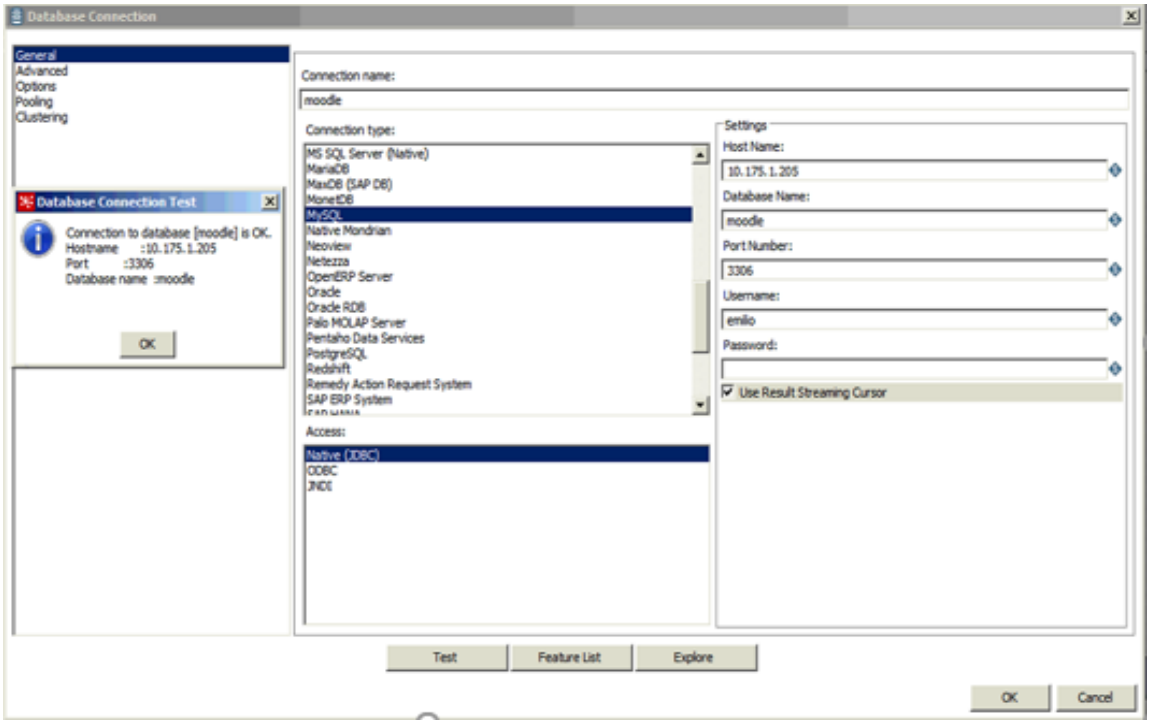


Figura 16. Prueba de la conexión con la base de datos exitosa.

Ahora se puede ver toda la información y las tablas alojadas en la base de datos establecida en la conexión.

4.3 Proceso ETL

Una vez que se tiene las conexiones con los diferentes servicios funcionando apropiadamente se puede comenzar con el proceso ETL de la fuente de datos.

En la figura 17 se muestra las diferentes etapas del proceso ETL. Para la etapa de extracción se han escogido seis diferentes tablas de la base de datos del Moodle las cuales van a servir como una de las fuentes de datos, además, también se utiliza otra fuente de datos que es un archivo plano el cual contiene las fechas que van a servir para obtener reportes en base a diferentes fechas; en la etapa de transformación se realizan diferentes tareas como selección de valores, utilización de filtros, separación de información, entre otros; y,

finalmente, se cargan los datos con el objeto Insert/Update que verifica también que los datos cargados no hayan sido alterados durante el proceso.

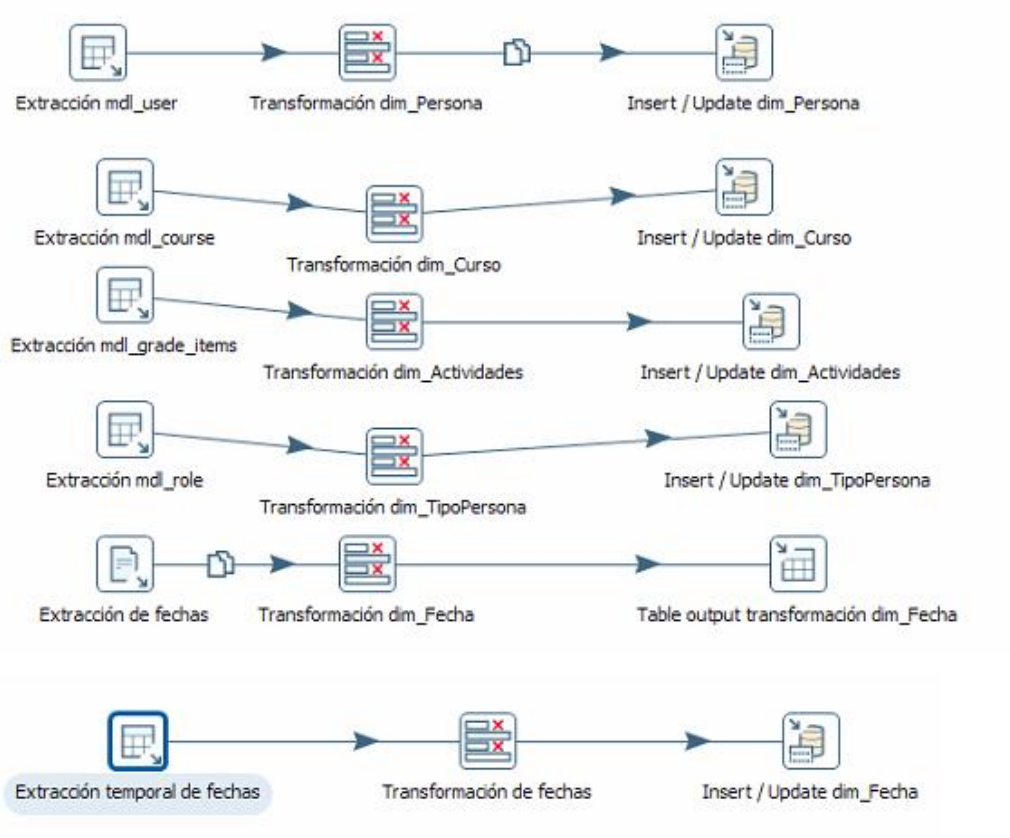


Figura 17. Proceso ETL del caso específico.

La base de datos del Moodle cuenta con alrededor de 400 tablas, de las cuales se han escogido seis tablas, y de cada tabla se han seleccionado los campos con la información más relevante para la construcción del DW que son las siguientes:

- Mdl_user. Contiene información de los usuarios.
- Mdl_grade_item. Contiene información de las actividades de cada curso.
- Mdl_role. Contiene información de los roles existentes.
- Mdl_grade_grades. Contiene información de las actividades y las notas de cada una.
- Mdl_course. Contiene información de los cursos.

- Mdl_role_assignments. Enlaza información entre roles y usuarios.

4.3.1 Extracción de la fuente de datos

4.3.1.1 mdl_user

En la tabla 1 se puede observar los campos que van a ser utilizados de la tabla mdl_user y una breve descripción de cada uno de ellos.

Tabla 1. Campos utilizados de la tabla mdl user.

Mdl_user	
Nombre del campo	**Descripción**
Id	Identificador de la tabla
Username	Nombre de usuario
Password	Contraseña del usuario
Firstname	Nombre del usuario
Lastname	Apellido del usuario
Email	Correo electrónico del usuario
auth	Tipo de autenticación

4.3.1.2 mdl_grade_grades

En la tabla 2 se muestran los campos que se van a utilizar en el proceso, y una breve descripción de estos. Cabe recalcar que userid se conecta con la tabla mdl_user, y el campo itemid se conecta con la tabla mdl_grade_item.

Tabla 2. Campos utilizados de la tabla mdl_grade_grades.

Mdl_grade_grades	
Nombre del campo	**Descripción**
Id	Identificador de la tabla
Userid	Identificador del usuario
Itemid	Identificador de la actividad
finalgrade	Nota final de la actividad

Timecreated	Fecha y hora del inicio de la actividad
Timemodified	Fecha y hora del fin de la actividad

4.3.1.3 mdl_grade_item

En la tabla 3 se muestran los campos que se utilizan de la tabla mdl_grade_item y una breve descripción de cada uno. El campo courseid se conecta con la tabla mdl_course.

Tabla 3. Campos utilizados de la tabla mdl_grade_item.

Mdl_grade_item	
Nombre del campo	**Descripción**
Id	Identificador de la tabla
Courseid	Identificador del curso
Itemname	Nombre de la actividad
Itemmodule	Tipo de actividad
Grademax	Nota máxima de la actividad
grademin	Nota mínima de la actividad

4.3.1.4 mdl_course

En la tabla 4 se muestran los campos que han sido seleccionados de la tabla mdl_course y una breve descripción de cada uno.

Tabla 4. Campos utilizados de la tabla mdl_course.

Mdl_course	
Nombre del campo	**Descripción**
Id	Identificador de la tabla
Fullname	Nombre completo del curso
shortname	Abreviatura del curso

4.3.1.5 mdl_role

En la tabla 5 se muestran los campos que han sido seleccionados de la tabla mdl_role y una breve descripción de estos.

Tabla 5. Campos utilizados de la tabla mdl_role.

Mdl_role	
Nombre del campo	**Descripción**
Id	Identificador de la tabla
Shortname	Nombre corto del rol
Archetype	Tipo de usuario

4.3.1.6 mdl_role_assignments

En la tabla 6 se muestran los campos utilizados y una breve descripción de estos. En esta tabla se enlazan las tablas de rol y de usuario por medio de sus identificadores.

Tabla 6. Campos utilizados de la tabla mdl_role_assignments.

Mdl_role_assignments	
Nombre del campo	**Descripción**
Id	Identificador de la tabla
Roleid	Identificador del rol
userid	Identificador del usuario

4.3.1.7 Generación y extracción de las fechas

Las fechas son muy importantes dentro de un proceso ETL y una plataforma de inteligencia de negocios, en este caso, sirve para poder obtener un histórico de las diferentes consultas que se puedan realizar. Por ejemplo, promedio de los estudiantes del primer semestre del año 2010. Para generar las fechas se utilizó un archivo plano (.txt) que contiene las fechas desde el año 1990 hasta 2100.

4.3.2 Transformación de los datos

En esta etapa se realiza la transformación de la información previamente extraída. Cada tabla recibe una transformación muy similar la cual puede consistir en selección de valores, filtrado de datos, separación de campos, entre otros.

4.3.2.1 Transformación de las tablas del Moodle

Para las tablas que han sido extraídas desde la base de datos del Moodle no se ha hecho más que cambiar los nombres antiguos de los campos por unos nuevos que faciliten la compresión.

Esto se ha logrado con un objeto Select Values como muestra la figura 18.

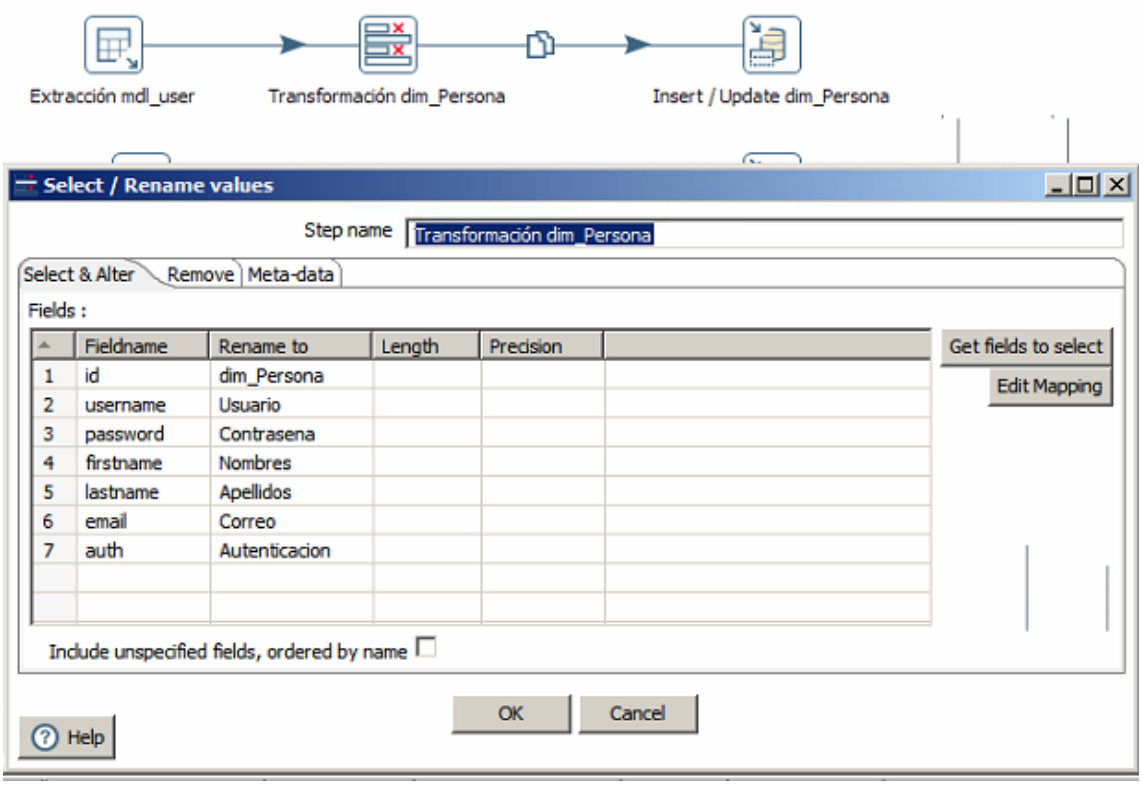


Figura 18. Transformación de la tabla mdl_user.

El proceso ha sido el mismo para el resto de las tablas donde se tienen diferentes campos y se cambia su nombre por uno que sea más fácil de comprender.

Hay que tomar en cuenta que en esta etapa no se han limpiado los valores nulos porque podrían contener información importante para otras tablas que no se han tomado en cuenta dentro de este caso específico.

4.3.2.2 Transformación de la fuente de datos de fecha

Una vez que se han extraído los datos de la fuente de datos, se utiliza el mismo objeto de Select values, pero esta vez para darle un tipo de datos, que será string, y una longitud máxima para evitar que se ingresen datos basura, ver figura 19; finalmente se suben estos datos a una tabla temporal en la base de datos.

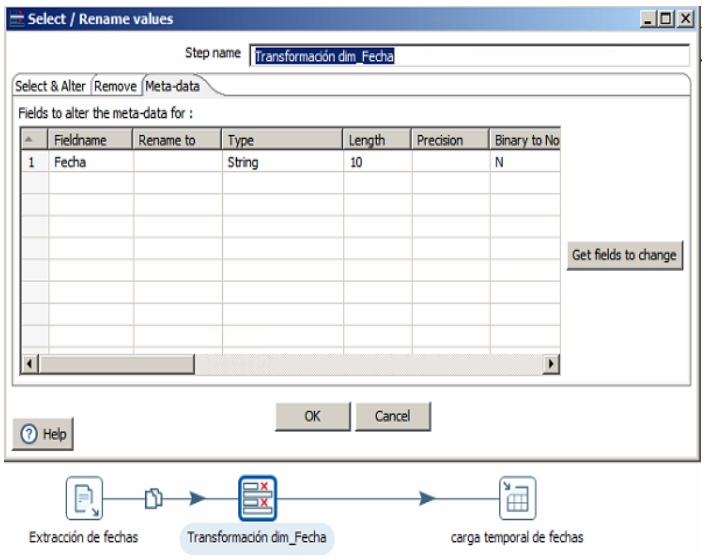


Figura 19. Transformación de la fecha.

Una vez que la tabla está en la base de datos, se la extrae para separar y unir los campos con la ayuda de sentencias y funciones SQL como year, date, replace, right y left, como muestra la figura 20.

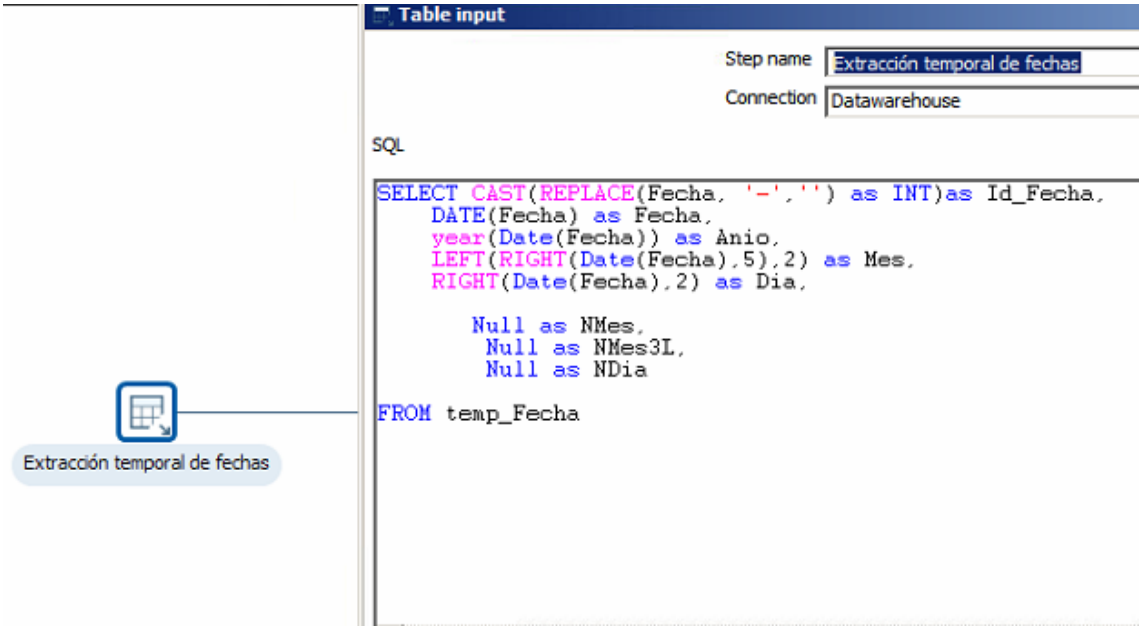


Figura 20. Extracción y trasformación de la tabla temporal de fechas.

Finalmente, con los campos ya listos, se realiza la transformación final que consiste en asignar un tipo de dato y longitud a cada uno de los campos. Ver figura 21.

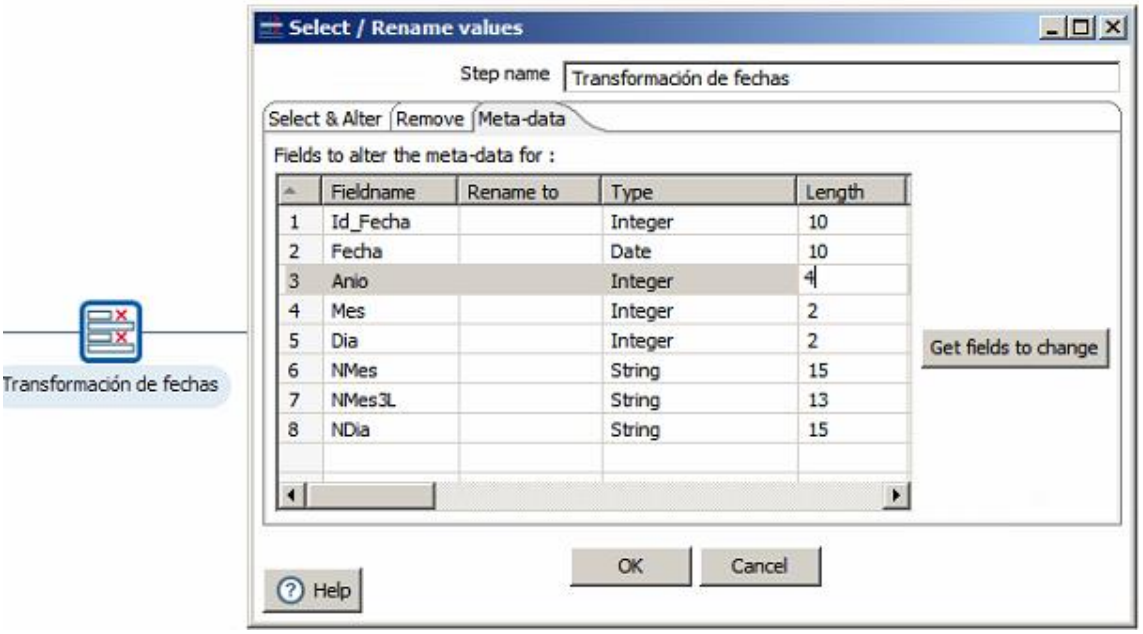


Figura 21. Transformación final de las fechas.

Cabe mencionar que el campo Id_fecha es la concatenación de los caracteres y suprimidos los guiones.

4.3.3 Carga de los datos

En el proceso de carga de los datos, consiste en recolectar los datos ya transformados y con ayuda de un objeto Insert/Update carga los datos a su respectiva tabla de dimensiones previamente creada en el DW. El objeto Insert/Update sirve para verificar que los datos extraídos y transformados sean los mismos que se cargan en el DW. Cuando se comienza a ejecutar el proceso ETL y llega a la etapa de carga, se inicia el Insert que carga todos los datos a la tabla de dimensiones respectiva, y luego, el Update revisa que cada uno de los registros sea consistente y actualice los que tengan errores. El campo de ID se excluye del update, ver figura 22.

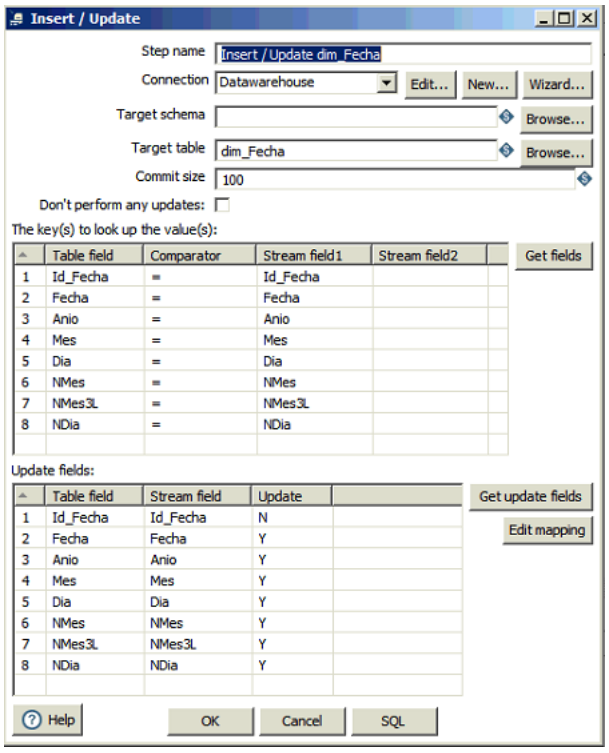


Figura 22. Insert/Update para carga de los datos al DW.

4.4 Diseño del DW

Cada una de las tablas de los subcapítulos anteriores se convirtieron en tablas de dimensiones tras ser cargadas al DW, y la tabla de hechos se forma de la unión de todas las tablas de dimensiones, dando resultado a un diseño tipo estrella, como muestra la figura 23.

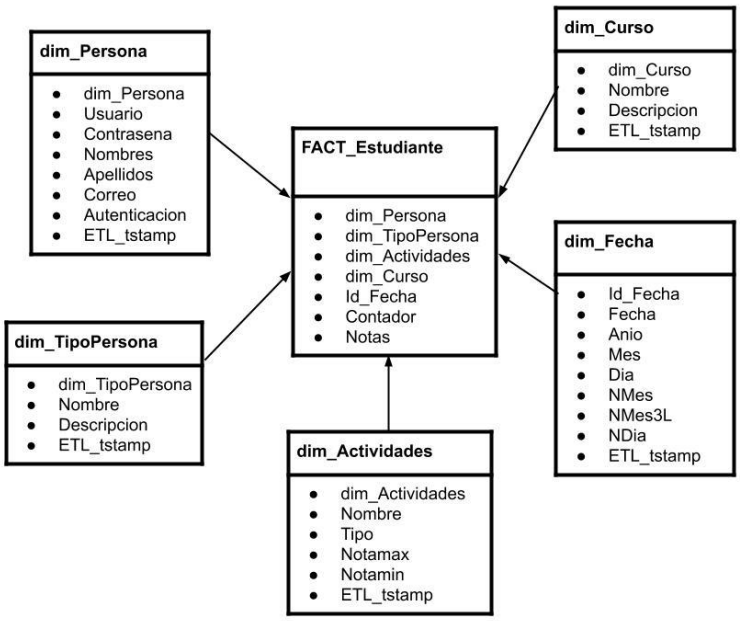


Figura 23. Tablas de dimensiones y tabla de hechos.

4.4.1 Tablas de dimensiones

Las tablas de dimensiones se crean previamente en el servidor 3 basándose en los campos y tipo de datos que se obtendrían a partir del proceso ETL, cada uno de los campos de las tablas del Moodle fueron renombrados en la etapa de transformación y su nuevo nombre aparece en la figura 23 con su tabla de dimensión correspondiente.

Existe un campo llamado ETL_tstamp en cada una de las tablas de dimensiones, este campo sirve para almacenar la hora y fecha de los cambios que se realicen en la tabla por lo que su importancia se la puede apreciar más en temas de auditoría.

Cada uno de los identificadores de las tablas de dimensiones se relaciona con la tabla de hechos (FAC_Estudiante) para dar paso a la creación de esta.

4.4.2 Tabla de hechos

La tabla de hechos contiene los identificadores de todas las tablas de dimensiones y es a la que se le hace las consultas para generar y obtener los reportes que ayudarán a la toma de decisiones.

En la tabla de hechos, aparte de los identificadores se tiene un campo llamado contador, el mismo sirve para realizar los conteos de los estudiantes de los que se quiera sacar información por medio del reporte.

La figura 24 muestra los pasos para la creación de la tabla de hechos, y son:

- Se crean dos consultas diferentes para la construcción de la tabla, una de la base de datos del Moodle de donde se sacan los identificadores de Persona, TipoPersona, Actividades, curso, fecha de inicio y fin de las actividades y notas; por otro lado, la segunda conexión se la realiza con la base de datos del DW de donde se extrae el identificador y la fecha.
- Ordenamiento de los campos correspondientes a sus consultas, con el objeto sort rows.
- En este paso se utiliza el objeto Merge Join para unir a las dos consultas por medio de los campos INICIO y Fecha. Lo que se busca con esto es que la fecha de inicio de la actividad se empareje con la fecha de la tabla de dimensiones de fecha y de esta manera obtener un histórico en cualquier instante de tiempo. Ver figura 25.
- Una vez que los campos de las dos consultas han sido unidos, es momento de seleccionar sus valores para que se los pueda cargar a la tabla de hechos previamente creada en la base de datos del DW. En la figura 26 se puede apreciar los campos obtenidos con el Merge Join.
- El objeto unique rows sirve para verificar que la información no se repita.

- Finalmente, en el paso 6 se realiza la carga de los datos a la tabla FACT_Estudiante alojada en el DW.

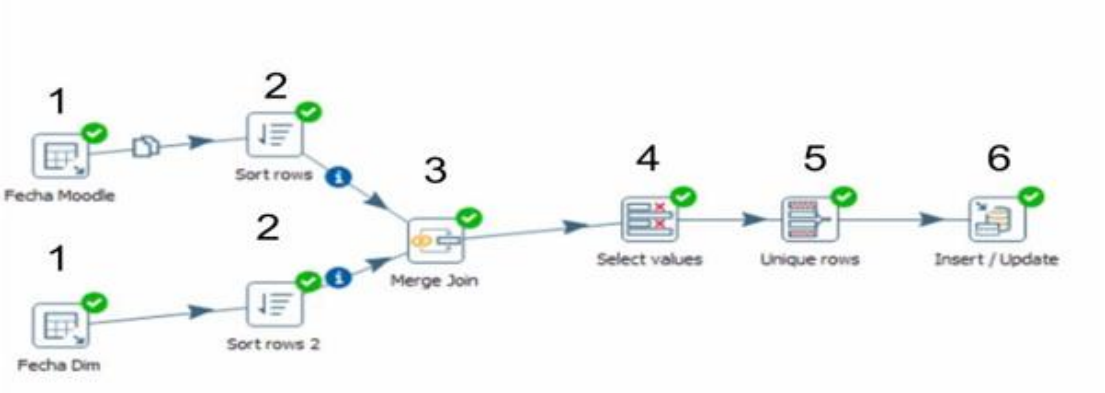


Figura 24. Proceso de creación de la tabla de hechos.

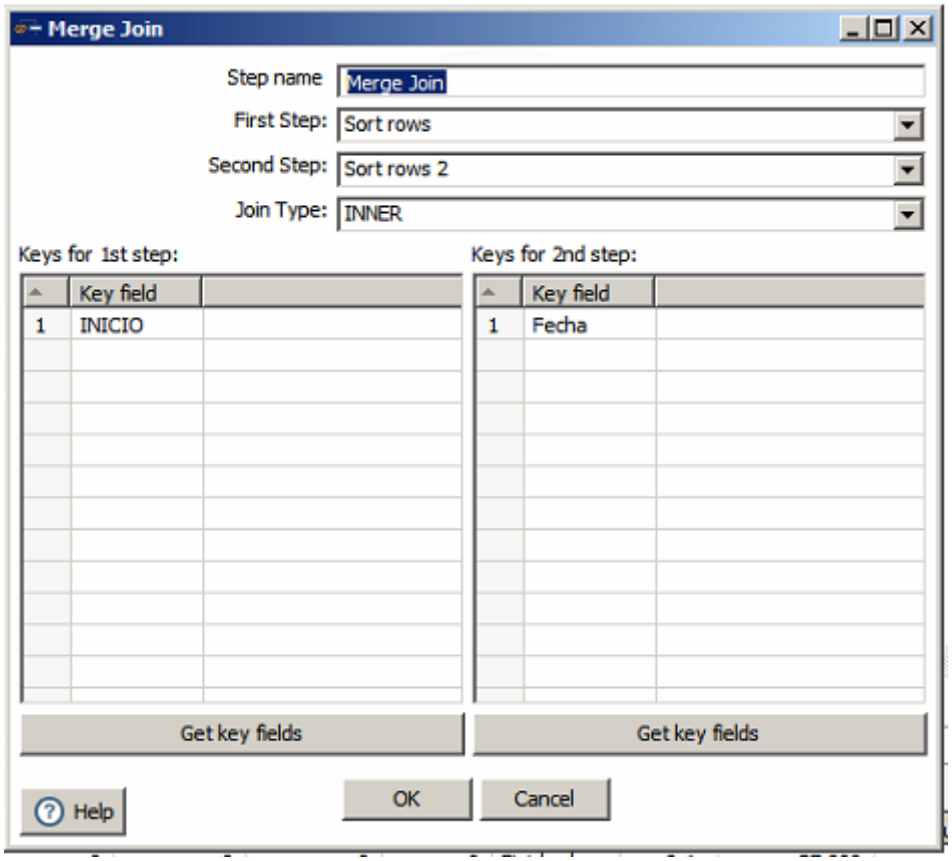


Figura 25. Unión de las dos consultas por medio del Merge Join.

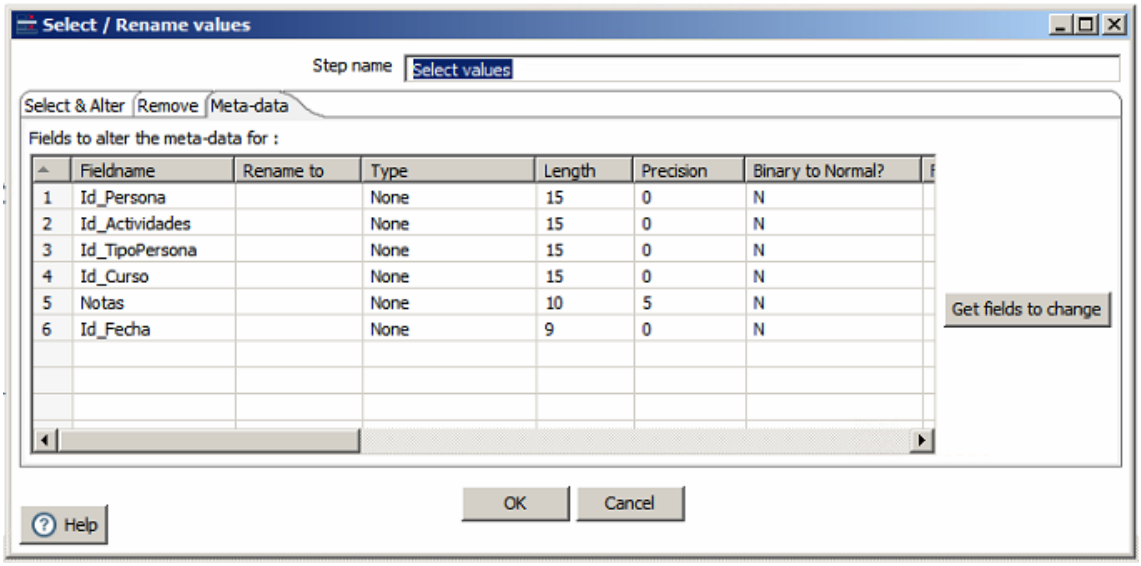


Figura 26. Campos obtenidos de la unión de consultas.

Un tema importante que se necesita profundizar es que la información de la fecha y hora tanto de creación, como de fin de las actividades se guarda en la base de datos de Moodle en un formato llamado Unix Time Stamp lo que hace que la información aparezca en forma de una cadena de números que dificultan la comprensión del diseño de la base de datos.

Con la ayuda del software DBDesigner también se ha logrado obtener el modelo final del DW, el cual se puede observar en la figura 27.

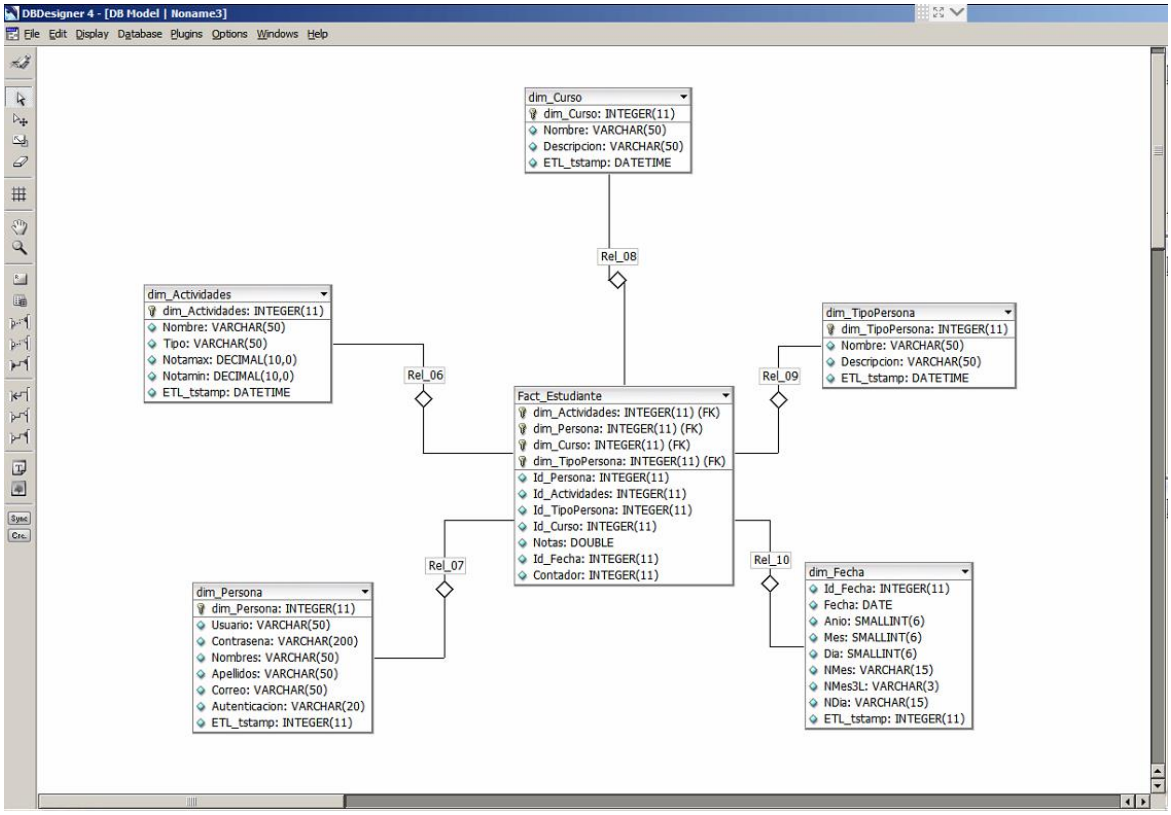


Figura 27. Modelo de la base de datos DW por DBDesigner.

4.5 Cubo OLAP

Para la creación del cubo se utiliza la herramienta Schema Workbench que viene incluida con la instalación del PDI.

Como muestra la figura 28, el primer paso es realizar la creación de cada una de las dimensiones, a las cuales se les asigna un nombre, un nivel de jerarquía y una tabla. El nombre que llevan es el mismo que tienen en la base de datos del DW, el nivel de jerarquía es el mismo para todas las dimensiones y queda en un valor por defecto, y, finalmente se le asigna la tabla correspondiente que está alojada en la base de datos del DW.

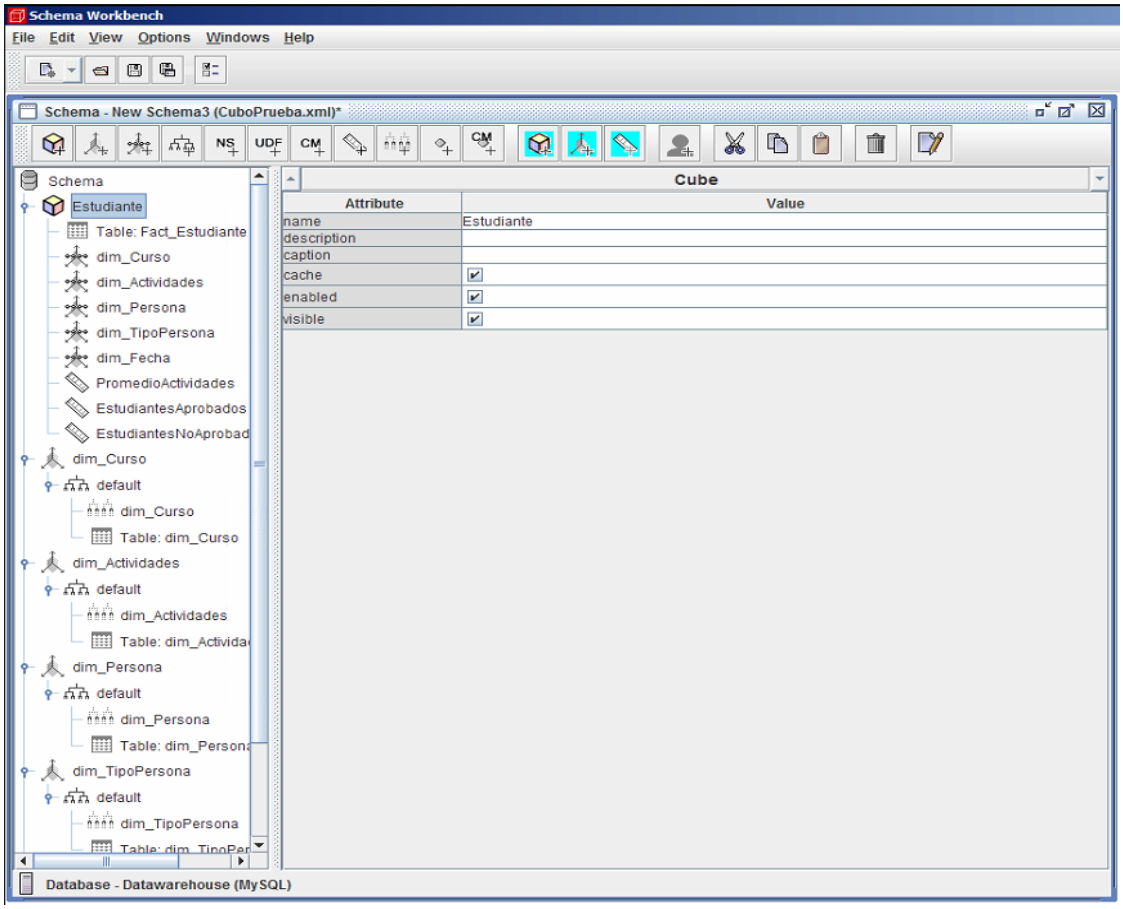


Figura 28. Estructura del cubo creado.

En el nivel de jerarquía se escoge también la columna que va a identificar a la dimensión y también el nombre de esta, las dos características mencionadas se escogen de los campos que existen en la tabla relacionada a la dimensión. Este proceso se repite con cada una de las dimensiones

El siguiente paso es crear el cubo en sí y agregar la tabla de hechos para que sus identificadores se puedan relacionar con cada una de las dimensiones ya creadas. Ahora se procede a agregar las dimensiones, las cuales tendrán un nombre que es el mismo nombre de la base de datos del DW, una clave foránea que es el identificador de la tabla de hechos y la fuente que es la tabla de dimensiones del DW. Estos pasos se repiten para todas las dimensiones.

Finalmente se agregan medidas al cubo, las cuales para este caso serán promedio por actividades de cada estudiante, número de estudiantes aprobados

y número de estudiantes reprobados. Así es como se llega al modelo mostrado en la figura 28.

4.6 Minería de datos

La minería de datos ha sido realizada con la herramienta Weka y como primera instancia se realiza la conexión de la base de datos con el fin de obtener las tablas y su información. De acuerdo con la figura 29, la conexión se establece con la base de datos de MySQL con el host “10.175.1.206” por el puerto “3306” a la base de datos llamada “Data warehouse”; sin embargo, se menciona que para establecer la conexión con la base de datos es necesario del conector J de MySQL. Para poder realizar la conexión se implementa una consulta SQL, la misma que se debe realizar con todos los campos que se desean analizar y posteriormente se la pega en la sección que dice “Query” como se muestra en la figura 29.

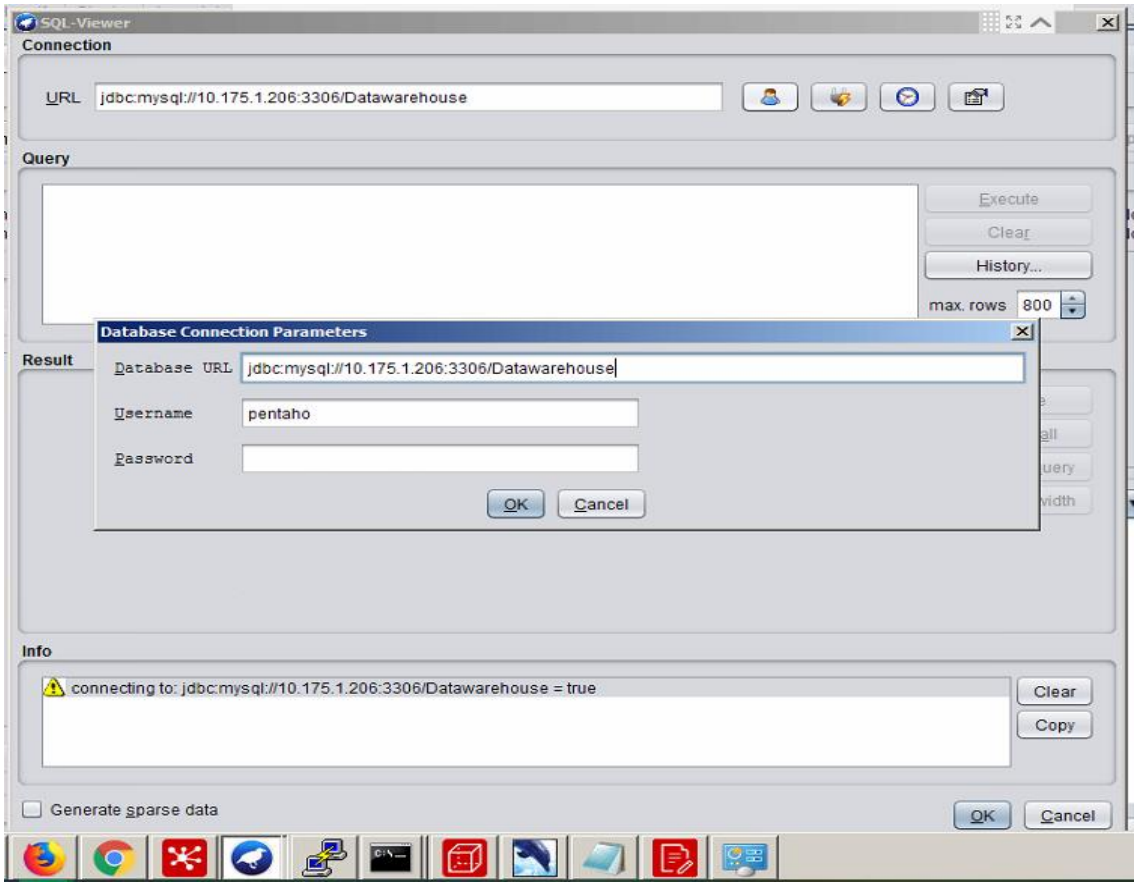


Figura 29. Conexión Weka.

Una vez se realiza la consulta SQL, en la figura 30 se observa el gráfico con la respectiva comparación de atributos. Se puede realizar distintas alteraciones a los campos con el fin de tener un gráfico diferente. Esta herramienta permite relacionar dos cosas de acuerdo con un conteo; sin embargo, cuando se selecciona una variable numérica, ésta puede presentar variaciones.

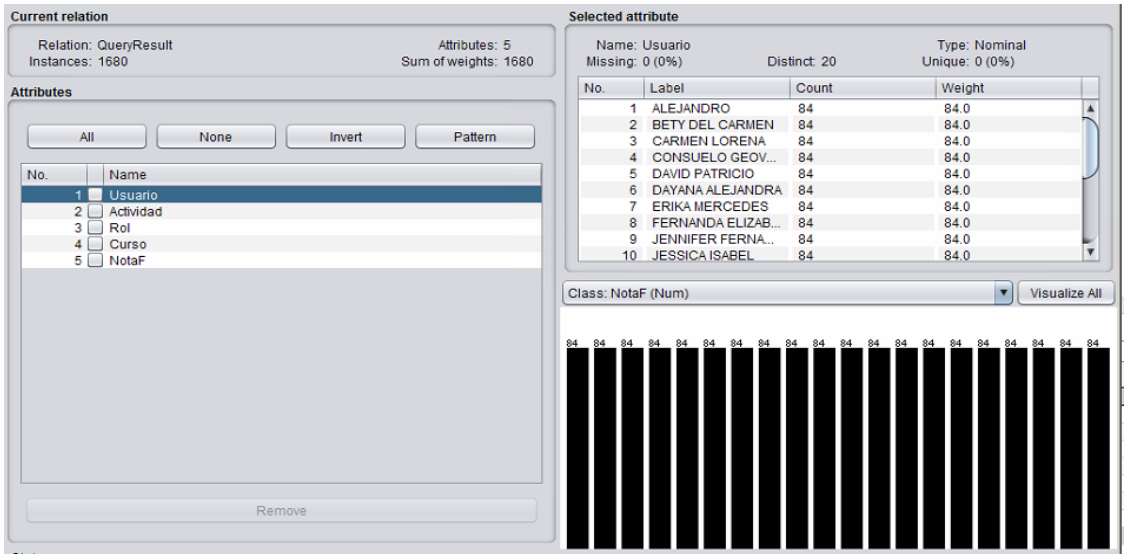


Figura 30. Número de actividades en el curso de Auditoria Informática.

En la figura 31 se muestra un conteo realizado por la herramienta de Weka en relación con un atributo de una base de datos. Weka también ofrece una característica en el momento de realizar sus gráficos, esta característica consiste en realizar un conteo de la cantidad de datos que posee el campo designado en la barra *Class*. En cuanto al diseño de gráficos Weka puede separar cada una de las barras mostradas por colores; para poder dar un bosquejo de las clasificaciones que tengan algún tema en particular. En la figura 31 se muestra un cambio en el campo “class” el cual clasifica a la barra de acuerdo con el atributo seleccionado.

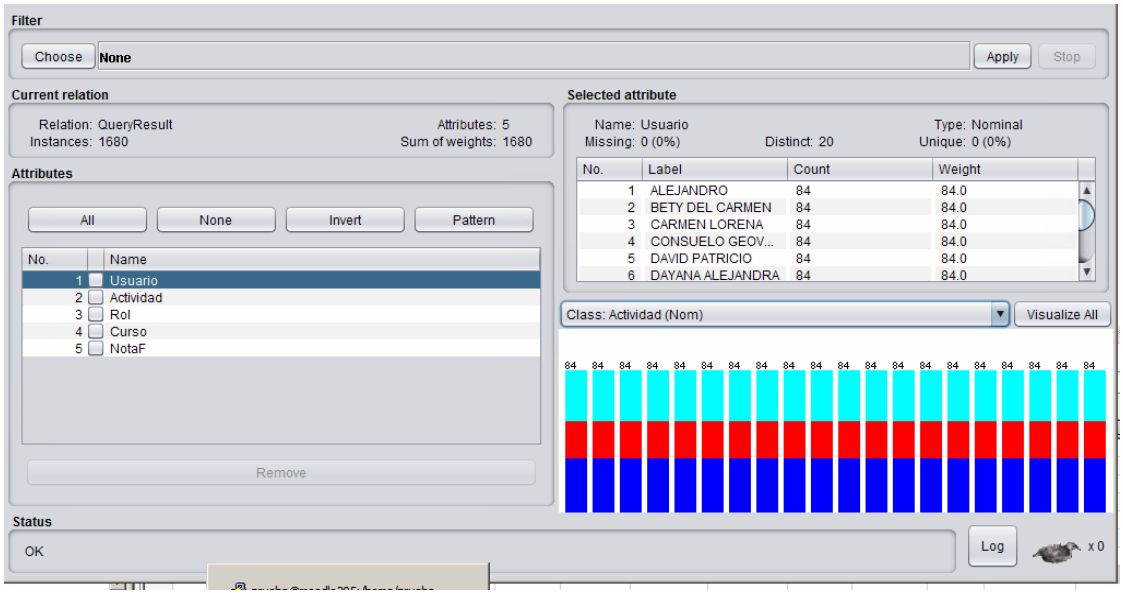


Figura 31. Cantidad de actividades en el curso por actividad.

Para la visualización de patrones es necesario ir a la pestaña “Visualize” y se muestra una especie de matriz, en donde se realiza la comparación entre todos los atributos en formas de cuadros, de acuerdo con la consulta suministrada al inicio. Se selecciona uno de eso cuadros y se despliega un gráfico como la figura 32. Al realizar un clic sobre un punto azul, se despliega una ventana y brinda la información que existe en ese punto. Para poder tener un mejor análisis en cuanto a la toma de decisiones.

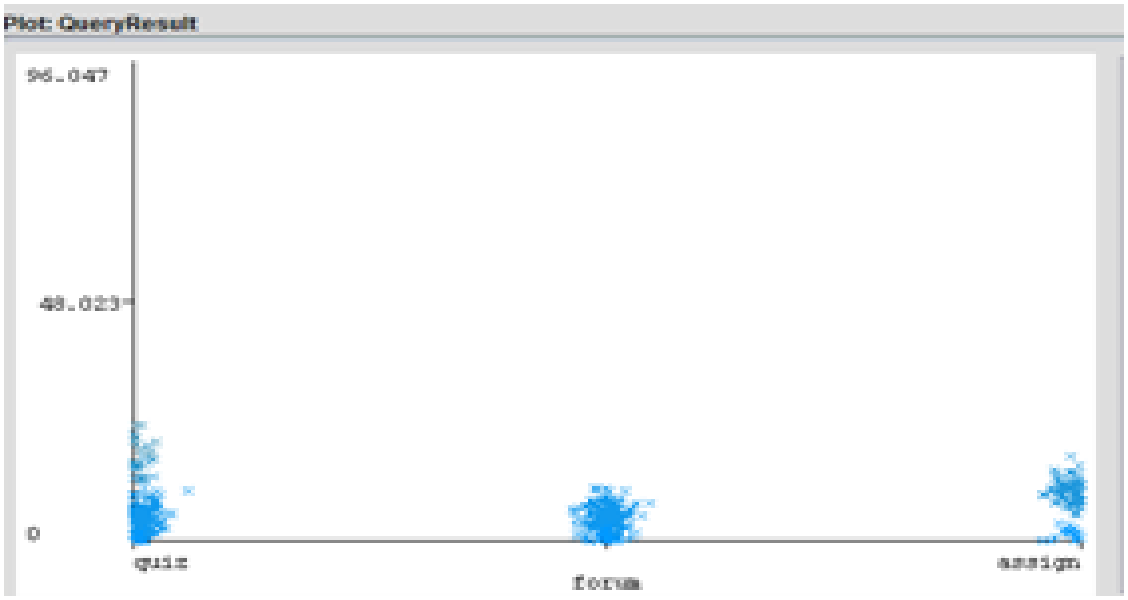


Figura 32. Patrón de notas por foro.

Se puede realizar toda clase de comparaciones dependiendo de la consulta, existen casos en donde se pueden reflejar casos ejemplo, como la figura 33, en donde existe información aislada que representa una anomalía. Los rectángulos resaltados implican un patrón constante en bajo nivel; esto sirve como retroalimentación en cuanto a usuarios.

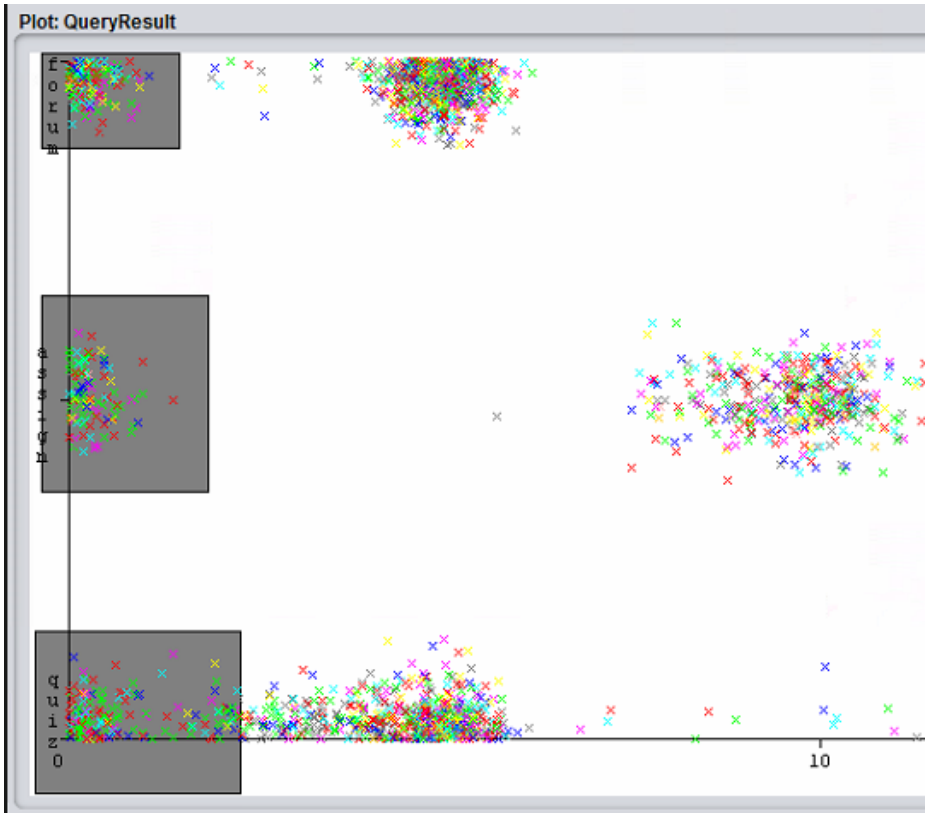


Figura 33. Patrón de usuarios por sus calificaciones y tareas.

4.7 Reportes

Para los reportes se utiliza la herramienta llamada Pentaho Report Designer.

Esta herramienta permite tomar plantillas o diseñar un reporte nuevo. Al diseñar un nuevo reporte se van arrastrando elementos como etiquetas, cuadros de texto, comentarios, imágenes, tablas, gráficos para estadísticas, es decir, todo lo que se necesite para crear un reporte personalizado. Aparte de esto, también se puede realizar nuevas consultas a través de conexiones a diferentes bases de datos y con esto sacar información que se necesita en el reporte.

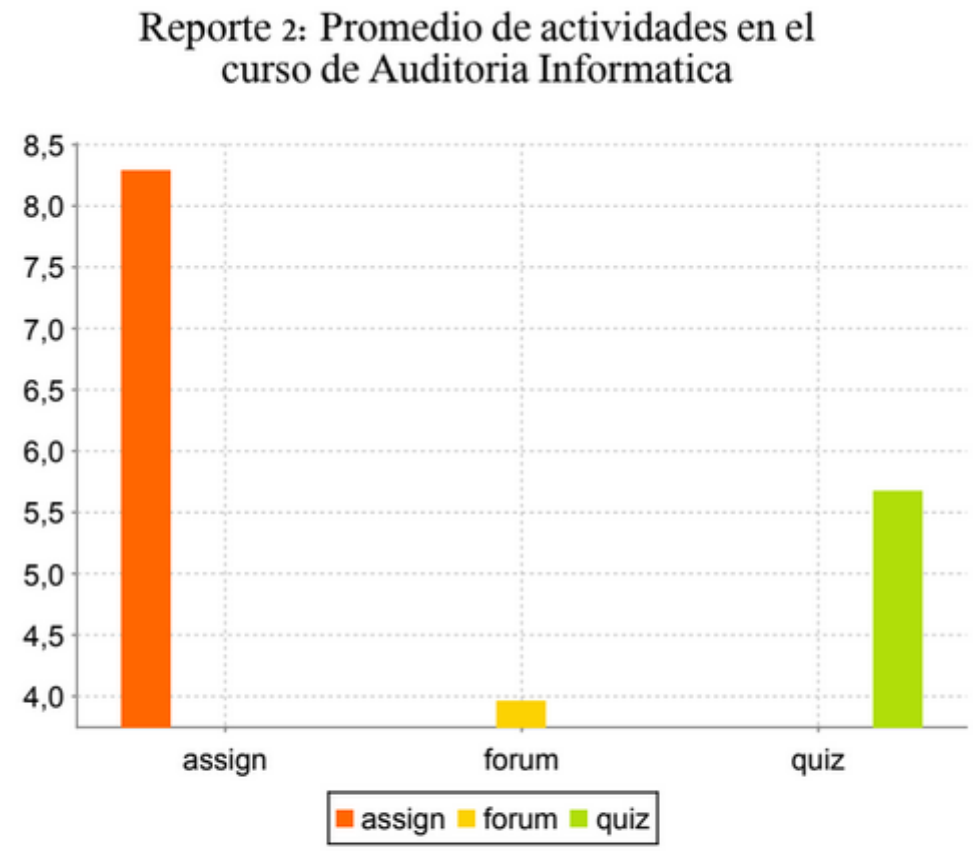
4.7.1 Reporte de promedio por tipo de actividad

En el primer reporte se muestra un promedio de las notas que tienen por tipo de actividad en los diferentes cursos, los reportes se muestran en las figuras 34, 35 y 36.

Lo que reflejan estos reportes es que el promedio de las actividades de tipo pruebas o cuestionarios es demasiado bajo y lo que más llama la atención es que el valor es bastante similar y se repite en los tres cursos. También se debe tomar en cuenta que el promedio de las actividades tipo foro tiene una calificación máxima de cinco, lo que quiere decir que la participación y el contenido de esta actividad es muy bueno.

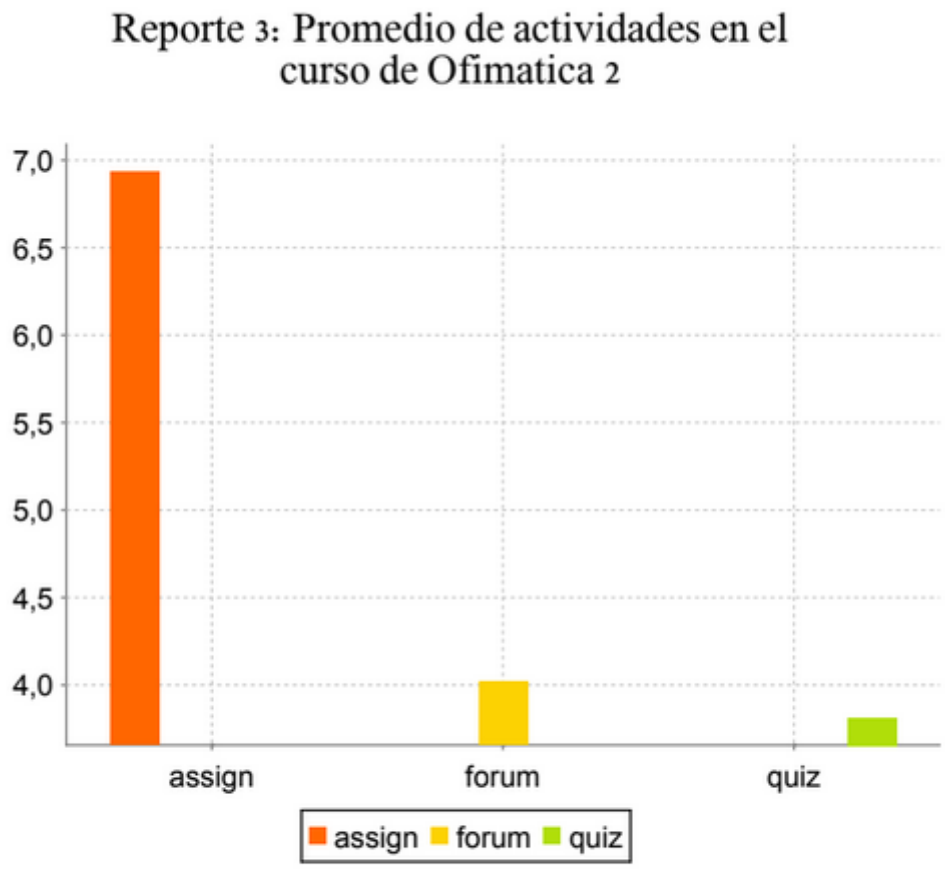
Tipo de Actividad	Promedio	Materia
assign	7,684177215	Ofimtica 3
forum	4,055555555	Ofimtica 3
quiz	4,519038076	Ofimtica 3

Figura 34. Promedio por actividades del curso Ofimática 3.



Tipo de la actividad	Promedio de la actividad	Materia
assign	8,294117647	Auditoria Informtica
forum	3,963800904	Auditoria Informtica
quiz	5,675291262	Auditoria Informtica

Figura 35. Promedio por actividades del curso Auditoría Informática.



Tipo de la actividad	Promedio de la actividad	Materia
assign	6,934042553	Ofimtica 3(2)
forum	4,020833333	Ofimtica 3(2)
quiz	3,813278008	Ofimtica 3(2)

Figura 36. Promedio por actividades del curso Ofimática 2.

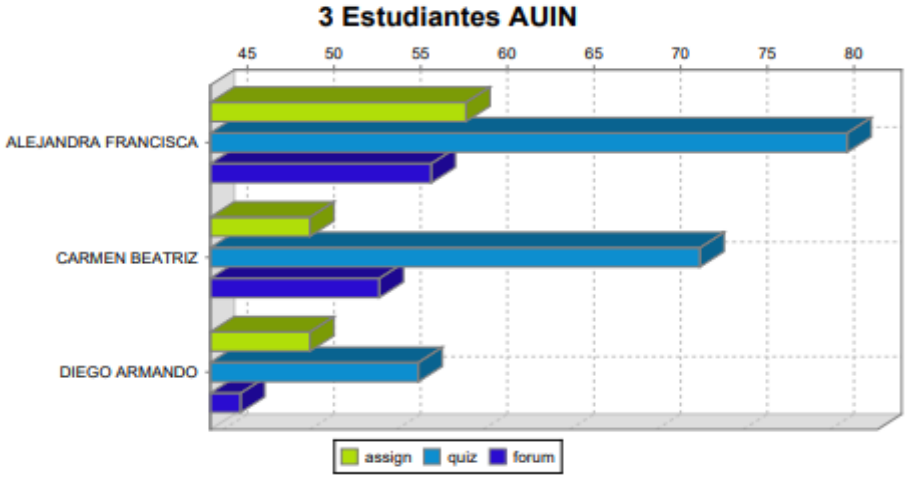
4.7.2 Reporte de promedio de actividades por estudiante

Lo que hace este reporte es brindar información que permita identificar las actividades que cada estudiante está o no realizando. Ver figuras 37, 38 y 39.

Cada uno de los siguientes reportes contiene la cantidad de actividades que cada estudiante está cumpliendo. Esto permite realizar un seguimiento más personalizado y, además, se podría identificar actividades que no están siendo

desarrolladas por ningún estudiante para tomar acciones preventivas y correctivas en el transcurso del período académico.

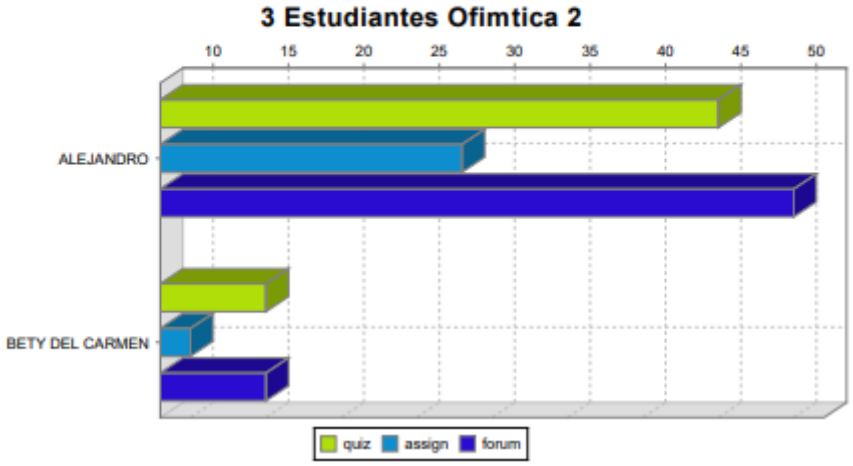
Reporte 4: Notas de actividades de los estudiantes de Auditoria Informatica



Alumno	Tipo de Actividad	Nota Final
ALEJANDRA FRANCISCA	assign	10
ALEJANDRA FRANCISCA	quiz	5
ALEJANDRA FRANCISCA	quiz	5
ALEJANDRA FRANCISCA	forum	5
ALEJANDRA FRANCISCA	quiz	5
ALEJANDRA FRANCISCA	quiz	5
ALEJANDRA FRANCISCA	quiz	4
ALEJANDRA FRANCISCA	quiz	20
ALEJANDRA FRANCISCA	assign	10
ALEJANDRA FRANCISCA	assign	10
ALEJANDRA FRANCISCA	assign	10
ALEJANDRA FRANCISCA	quiz	
ALEJANDRA FRANCISCA	quiz	5
ALEJANDRA FRANCISCA	forum	5

Figura 37. Cantidad de actividades que está cumpliendo cada estudiante del curso de Auditoría Informática.

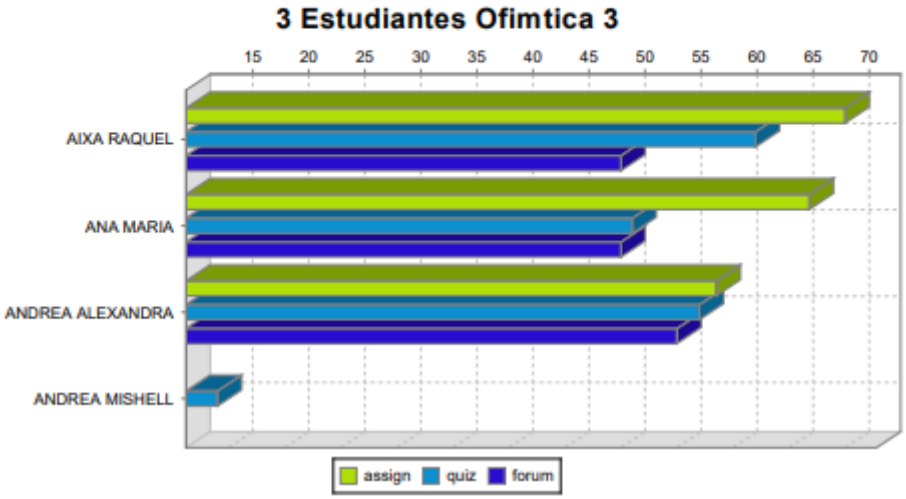
Reporte 6: Notas de actividades de los estudiantes de Ofimatica 2



Alumno	Tipo de Actividad	Nota Final
ALEJANDRO	quiz	
ALEJANDRO	assign	
ALEJANDRO	forum	5
ALEJANDRO	quiz	
ALEJANDRO	assign	10
ALEJANDRO	quiz	0
ALEJANDRO	quiz	2
ALEJANDRO	quiz	
ALEJANDRO	forum	5
ALEJANDRO	quiz	
ALEJANDRO	assign	
ALEJANDRO	quiz	
ALEJANDRO	forum	
ALEJANDRO	quiz	3

Figura 38. Cantidad de actividades que está cumpliendo cada estudiante del curso de Ofimática 2.

Reporte 5: Notas de actividades de los estudiantes de Ofimatica 3



Alumno	Tipo de Actividad	Nota Final
AIXA RAQUEL	assign	10
AIXA RAQUEL	quiz	5
AIXA RAQUEL	assign	10
AIXA RAQUEL	assign	9
AIXA RAQUEL	forum	5
AIXA RAQUEL	quiz	
AIXA RAQUEL	forum	0
AIXA RAQUEL	forum	5
AIXA RAQUEL	forum	5
AIXA RAQUEL	quiz	5
AIXA RAQUEL	quiz	15
AIXA RAQUEL	forum	5
AIXA RAQUEL	quiz	4
AIXA RAQUEL	quiz	5

Figura 39. Cantidad de actividades que está cumpliendo cada estudiante del curso de Ofimática 3.

.

5. Capítulo V. Evaluación de resultados.

5.1 Reportes

En cuanto al primer reporte lo que se puede evidenciar es que las actividades dentro del Moodle se clasifican en tres categorías diferentes, las cuales son assign, fórum y quiz. Cabe mencionar que las calificaciones de los foros tienen una calificación máxima de cinco puntos.

Lo que se puede observar del reporte es que existe un patrón que se repite en los tres cursos, el promedio de calificaciones de los cuestionarios o exámenes es muy bajo comparado con el resto de las actividades esto puede indicar problemas como dificultad para retener la información, falta de preparación por parte de los estudiantes, o también, falencias en el método de enseñanza. Cualquiera que sea la causa del problema puede ser solucionado una vez que sea encontrado y de esta manera mejorar la calidad educativa.

Por otro lado, el segundo reporte muestra la cantidad de actividades que cada estudiante está realizando y sus calificaciones. Esto puede ayudar como alerta temprana para detectar las actividades que los estudiantes no están realizando y de esta manera poder corregirlas a tiempo ya que el reporte se obtiene de cada estudiante. Otro enfoque podría ser clasificar a los estudiantes que tienen la menor cantidad de actividades completadas y ofrecer una ayuda personalizada para evitar futuros problemas como deserciones o pérdidas de la materia.

5.2 Minería de datos

Una vez se realiza la consulta SQL, se puede observar los atributos que está disponible para realizar comparaciones por medio de barras. En la figura 40 se observa que la cantidad de notas por estudiante que existen en el curso *Auditoría Informática* es de 84. La alteración del campo *class* puede alterar la gráfica de barras y brindar otra información distinta.

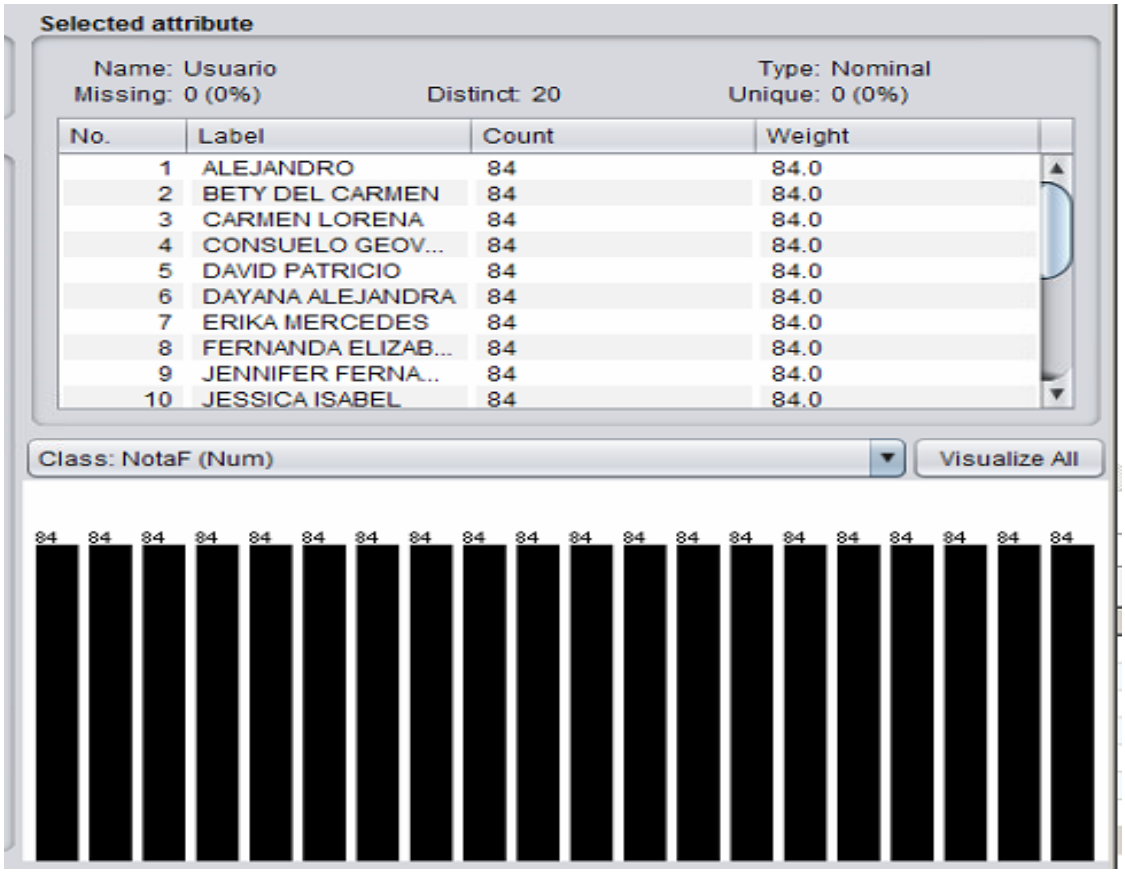


Figura 40. Cantidad de notas por estudiante en un determinado curso.

En la figura 41 se observa el cambio en el campo *class*, en este caso la información de salida brinda la cantidad de notas que tiene un quiz, test y assign. Sin embargo, se delimita como campo *class* al nombre de la actividad y el campo relacionado es el estudiante. La mejor manera de explicar es que la barra esta segmentada en el tipo de actividad. Como se observa existen 3 cursos, los cuales tienen cierta cantidad de calificaciones.

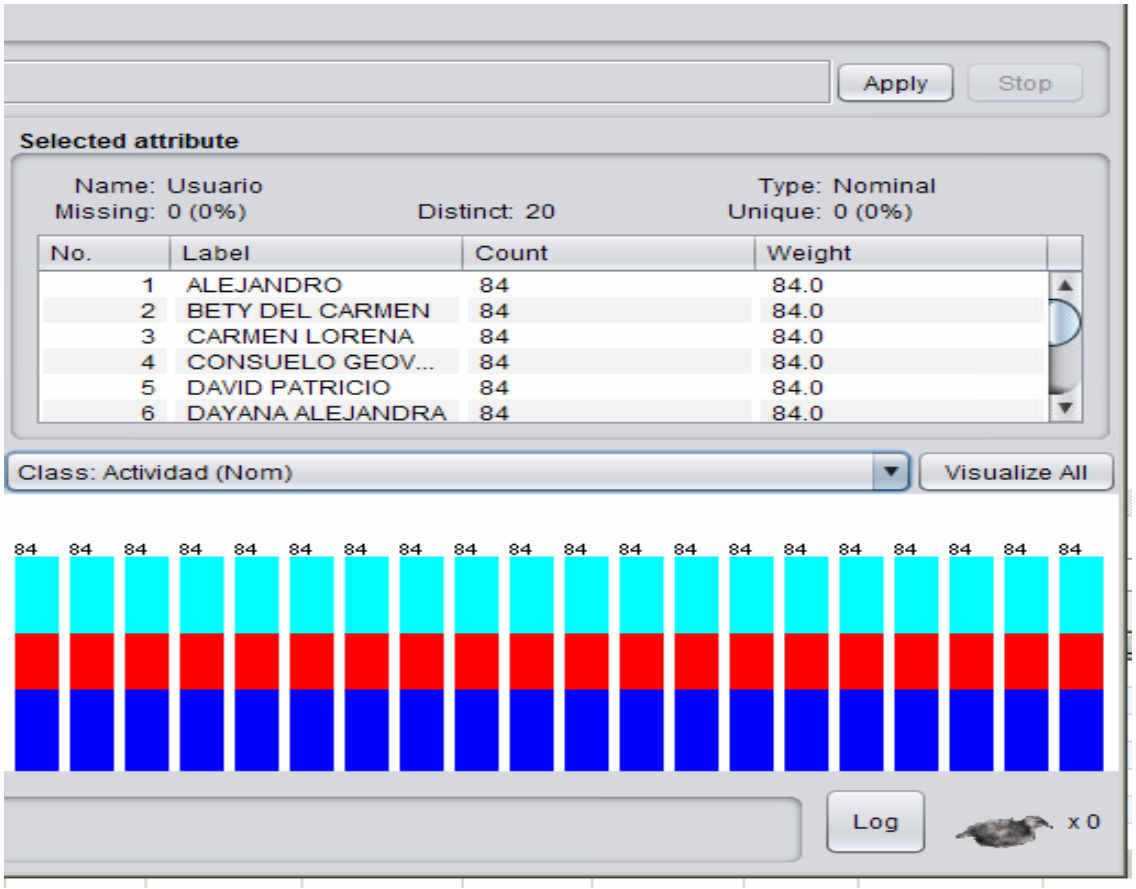


Figura 41. Cantidad de calificaciones por estudiante segmentada por sus actividades.

En la figura 42 con el uso de la herramienta Weka se puede realizar una comparación de información por medio de gráficos los cuales tendrán patrones. Se observa que la comparación se estructura como “Las notas de las actividades por usuario” y se analiza la información. Como se entiende el gráfico los puntos azules representa las “quiz” con una nota sobre 10 puntos, los puntos rojos los “assign” con una nota sobre 10 puntos y los puntos verdes los “forum” con una nota sobre cinco puntos. Se toma en cuenta lo anterior y se llega a la hipótesis de que los puntos azules son constantes en la parte inferior, es decir que la mayoría de los estudiantes no estudian para rendir pruebas o cuestionarios. En los foros, la mayoría de estudiante llegan a tener buenas notas, es un punto que no interesa por el momento. Sin embargo, en las tareas se presenta una variación moderada de eventos.

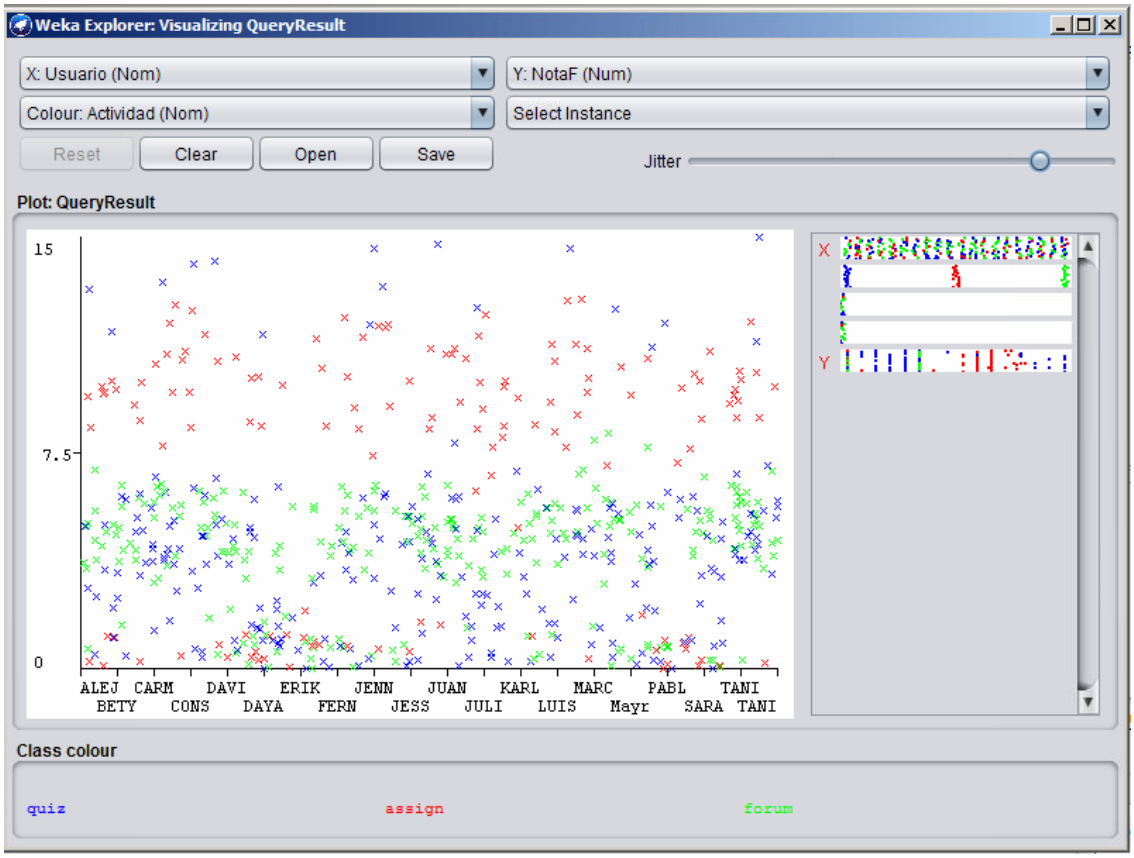


Figura 42. Patrón de actividades de acuerdo con su calificación y estudiante.

En el patrón que se muestra en la figura 43 se puede observar la relación entre el promedio por actividad, si se fija en los puntos azules lo cuales están en el rango de 0 a 10, se encuentran concentrados en la parte inferior. Al dar un clic en la parte concentrada se despliega la tabla con la información de tres estudiantes los cuales poseen notas bajas en las actividades de “quiz”. Mientras que en otro caso se muestra exactamente el mismo patrón y se da clic en la parte de “fórum”, como análisis se obtienen resultados esperados tal como antes se menciona, los foros para los estudiantes no son un problema de rendimiento. El problema son las pruebas y cuestionarios.

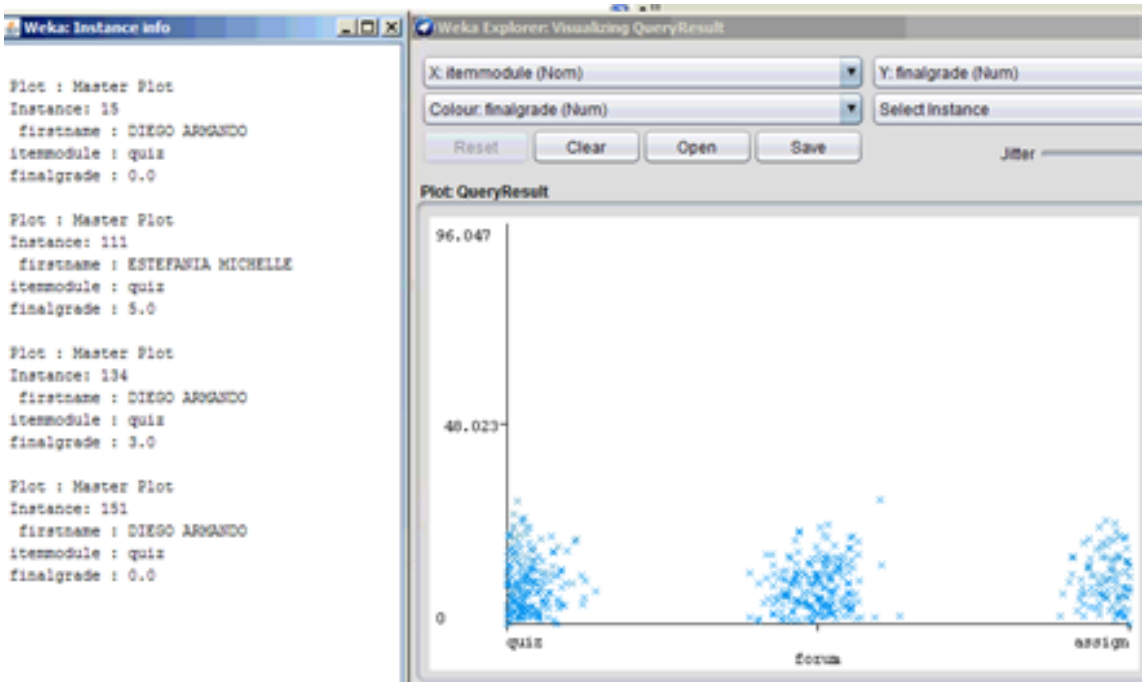


Figura 43. Patrón de calificaciones por actividad.

En la figura 44 se puede determinar el patrón de relación entre las actividades y su promedio por usuario. De igual manera se percibe una figura ploma en la cual se hace énfasis, debido al patrón en donde los estudiantes de color rojo, entre otros, tienden a tener calificaciones de cero puntos, tal y como se encuentra en la ventana izquierda. Para poder realizar una mejor métrica de este resultado sería conveniente tomar en cuenta la asistencia del estudiante a clases, pero esa información no se posee en la base de datos.

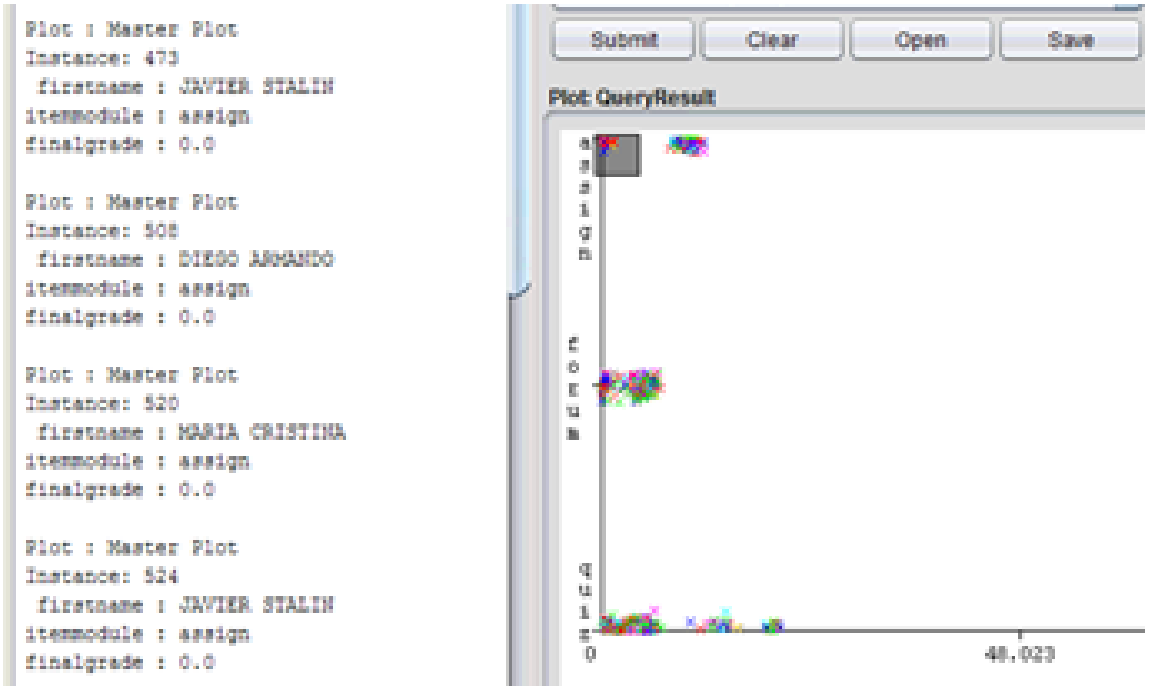


Figura 44. Patrón de usuario por sus calificaciones.

En la toma de decisiones según los árboles de decisión se obtuvo el siguiente resultado como se muestra en la figura 45. Para ello se utiliza el algoritmo J48 basado en arboles diseñado para establecer relaciones y realizar un análisis de datos estadísticos. Como se observa el porcentaje de que la clasificación haya sido correcta es del 70%, es decir que el árbol fue realizado parcialmente bien. El atributo para analizar es el de Actividad (quiz, fórum, assign), el algoritmo lo relaciona con la nota final de cada actividad y desarrolla el árbol que se encuentra en la figura 46. En este árbol se realiza la comparación de la actividad por nota y se llega a la conclusión de poder tomar en cuenta en donde los estudiantes están fallando más; todo esto con el fin de mejorar cada uno de los aspectos de cada actividad para que los estudiantes no tengan inconvenientes al pasar una materia. Se realizan condicionamientos para poder clasificar la información y hacer que pertenezca la nota a la cual pertenece.

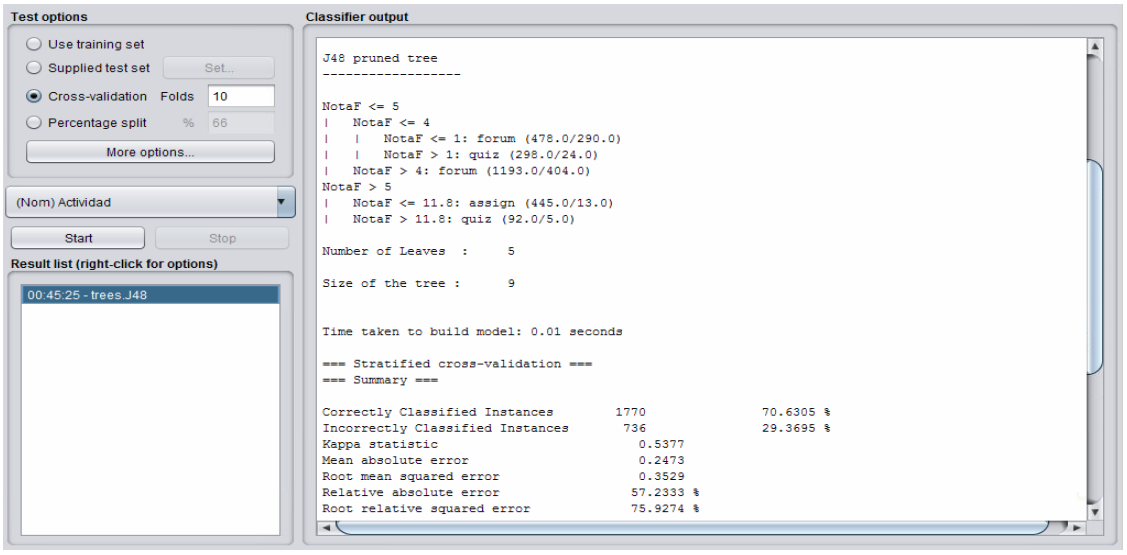


Figura 45. Árbol de decisión de acuerdo con la actividad.

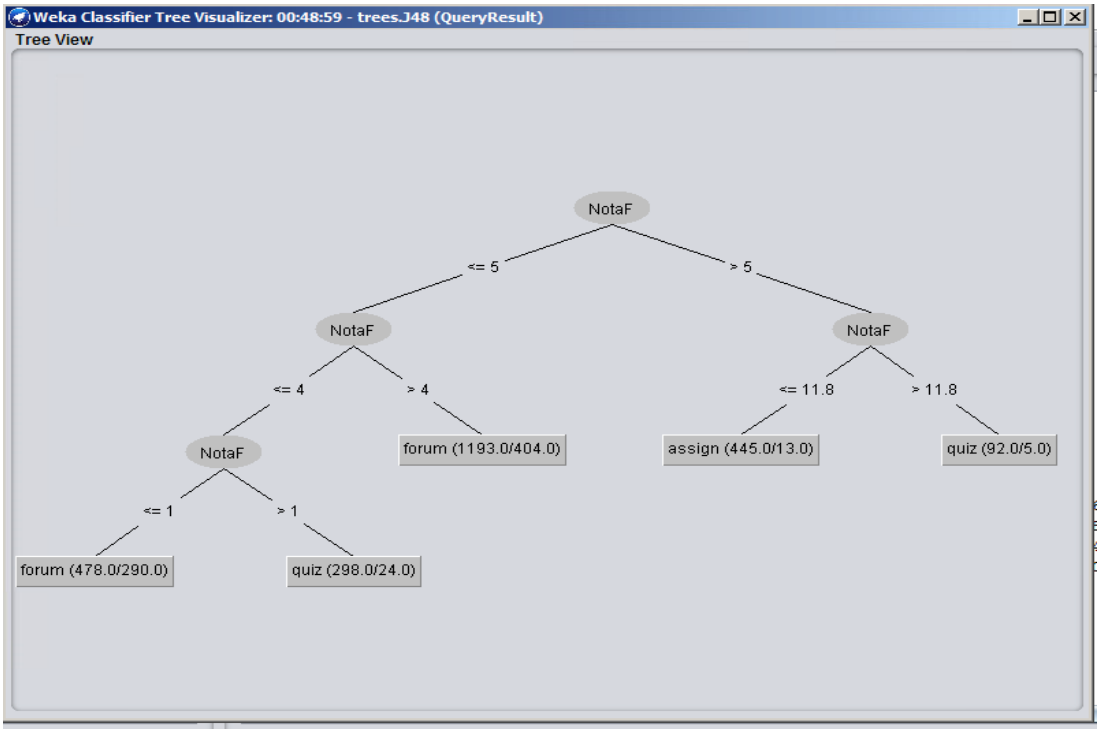


Figura 46. Forma del árbol de decisión.

Finalmente, para la figura 47 se realiza la toma de decisiones con el mismo algoritmo J48. Esta vez se usa una cantidad más extensa de datos con el fin de hacer la clasificación más exacta. Se realiza la consulta y se obtiene un 74% de una correcta clasificación. Lo cual significa que el árbol de decisión de la figura

48 hace la misma clasificación, sin embargo, este ahora incluye el rol del usuario. Cada una de las clasificaciones tiene su respectiva comparación. En punto de realizar una clasificación para la toma de decisiones es poder llegar a que la relación del pie de árbol llegue a uno.

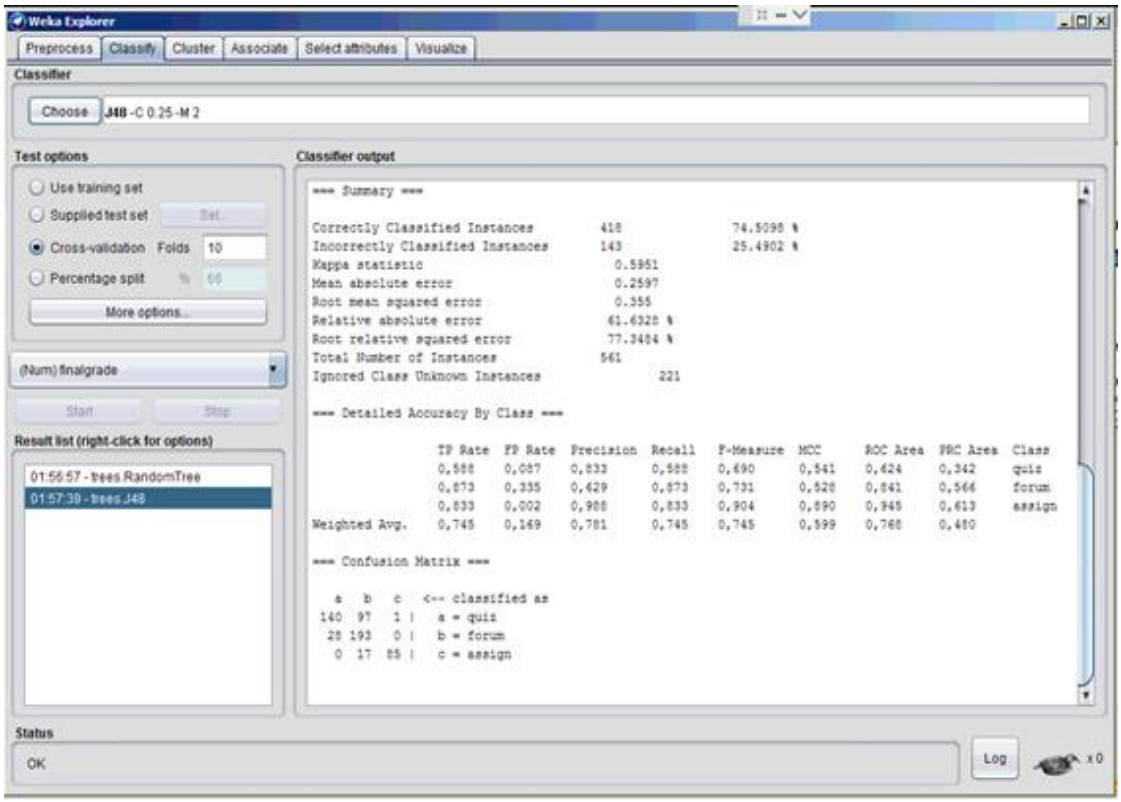


Figura 47. Árbol de decisión dos de acuerdo con la actividad y rol.

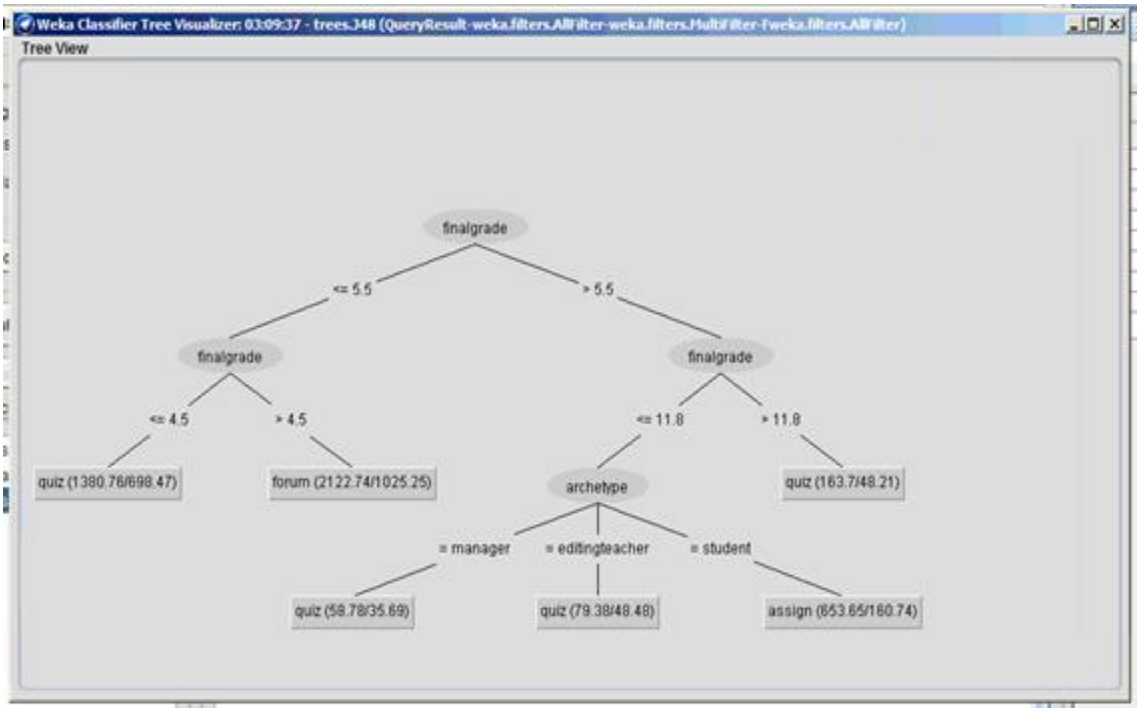


Figura 48. Forma del árbol de decisión dos

6. Capítulo VI. Conclusiones y recomendaciones.

6.1 Conclusiones

El uso de herramientas de análisis predictivo más reportes obtenidos basados en la información de los estudiantes puede ser determinante en el éxito o fracaso de estos.

Con la minería de datos se ha logrado predecir que los estudiantes que tienen mayor participación en las actividades tienen más posibilidades de aprobar las diferentes materias sin mayor problema.

Tener un buen entendimiento de las fuentes de datos, tablas de dimensiones y tabla de hecho garantiza un buen diseño de la base de datos del DW y evita resultados no esperados a la hora de generar reportes o indicadores.

La convergencia de herramientas propietarias y de código libre se ha logrado sin mayor problema, las diferencias están presentes a la hora de buscar información relacionada al soporte o configuración de estas.

El algoritmo J48 de árbol de decisiones ayuda a la toma de decisiones en casos específicos realizando cálculos estadísticos en base a un atributo de tipo string en la base de datos, este algoritmo viene instalado en la herramienta de Weka.

Para poder tener un buen diseño de la tabla de hecho es necesario saber lo que se necesita para el negocio, es decir, determinar la información que se requiere analizar y la razón por la cual se quiere analizar.

Se requiere un alto conocimiento sobre inteligencia de negocios y DW, es decir, toda clase de definiciones que sirvan para el desarrollo de un proyecto y todas las características técnicas que se necesitan para desarrollar un modelo de DW.

Los cubos OLAP son consultas dentro de otras, un cubo posee distintas métricas, dimensiones y una tabla de hecho, la misma que debe ser diseñada de

manera exhaustiva debido a su complejidad, ya que su estructura está compuesta por todas las tablas de dimensiones.

Para un mejor análisis de resultados dentro del DW es necesario incluir más datamarts con el fin de expandir el tamaño de las consultas y hacer que el DW sea aún más eficiente para realizar peticiones y su retardo sea el menor posible.

La predicción de patrones ayuda a la toma de decisiones por parte de profesores e inclusive por la misma universidad.

Si se agrega más información a un DW de la Universidad es posible que genere gran ayuda para poder establecer una toma de decisiones correcta, con el fin de que los estudiantes tengan un mejor rendimiento.

En base a los reportes se ha identificado que en todos los cursos existe un alto porcentaje de participación en las actividades tipo foro; por otro lado, el promedio de las pruebas es muy bajo en los tres cursos lo cual puede generar problemas al final del semestre porque este tipo de actividades generalmente lleva el mayor peso de la nota final para los estudiantes.

Del segundo reporte se ha detectado que la cantidad de estudiantes que realiza todas las actividades es mínima, lo cual se relaciona con el primer reporte en donde se observa que el promedio de calificación de las pruebas es bajo. Esto puede indicar que las horas de estudio fuera de clase no se están cumpliendo.

Para el proceso ETL, PDI es una herramienta muy intuitiva en cuanto al funcionamiento de sus herramientas y la forma de interactuar con ellas con la característica de arrastrar y soltar.

En el Reporte 4: Notas de actividades de los estudiantes de Auditoria informática se tiene el número de actividades realizadas por estudiante, basándose en el ejemplo, se han encontrado tres estudiantes de los cuales uno no ha realizado la misma cantidad de foros en comparación con el resto.

Report Designer es una herramienta completa para la generación de informes, por medio de consultas hacia bases de datos, con una interfaz completa que posee un sin número de características para poder potenciar los resultados.

6.2 Recomendaciones

Antes de implementar una plataforma de inteligencia de negocios es necesario tener un nivel de comprensión alto de la base de datos que se va a manejar para que la fuente de datos sea lo más precisa posible y de esta manera evitar procesamiento innecesario y un mejor diseño del DW.

Se recomienda tomarse el tiempo necesario para definir los requerimientos de manera precisa ya que esta información es muy importante para que la implementación sea lo más precisa posible y sin pérdidas de tiempo.

La involucración de docentes y autoridades en este tipo de plataformas de inteligencia de negocios puede ayudar en gran medida a los estudiantes de la Universidad y a la mejora de la calidad educativa de la misma.

Se recomienda que los docentes pongan cuidado especial y den seguimiento a las actividades que proponen a lo largo del curso con la ayuda de las herramientas planteadas en este proyecto de titulación para que el resultado final sea mucho más fructífero para todas las partes involucradas.

Es importante conocer un tamaño aproximado de las bases de datos para poder definir los requerimientos solicitados por los diferentes sistemas y herramientas que se van a implementar, y no tener problemas de rendimiento cuando el proyecto de titulación se encuentre en marcha.

Tener respaldos de la base de datos original es imprescindible para que la información almacenada no se vea alterada ya que es información real de aulas virtuales.

Se recomienda tener conocimiento acerca de sistemas Linux, debido a la manipulación de la consola dependiendo de donde se instalen los servidores con los motores de base de datos.

Se recomienda la instalación de TeamViewer, si la implementación es en un lugar lejano, gracias a esta herramienta se puede realizar cualquier acción dentro de la máquina en cualquier momento.

Tomar en cuenta todos los requerimientos de software y hardware para la instalación de las herramientas, con el fin de no tener inconvenientes a lo largo de la práctica que se realice.

Tener conocimientos sobre BI para poder realizar cualquier proceso ETL; sin embargo, también se recomienda poseer conocimiento sobre el proceso de minería de datos y DW.

Se recomienda realizar el diseño de las tablas de dimensiones y hecho con antelación para no tener inconvenientes en la parte de la implementación.

Los reportes que se generen de esta plataforma deben ser lo más amigables con el usuario final para que no se vean contrariados a la hora de utilizar esta funcionalidad.

Referencias

Apache. (2010). Apache Tomcat. Recuperado el 10 de marzo de 2019 de http://tomcat.apache.org/

Apache Tomcat. (s.f). Tomcat 8 Software Downloads. Recuperado el 10 de marzo de 2019 de https://tomcat.apache.org/download-80.cgi

Bailón, R. (2010). Experiencias de mi cambio hacía Windows Server 2008 R2. Recuperado el 14 de marzo de 2019 de https://informaticaenred.wordpress.com/2010/01/02/experiencias-de-mi-cambio-hacia-windows-server-2008-r2/

Databricks. (2019). Extract, Transform and Load. Recuperado el 14 de marzo de 2019 de https://databricks.com/glossary/extract-transform-load

Fedora Wiki. (s. f.). EPEL/es. Recuperado el 13 de marzo de 2019 de https://fedoraproject.org/wiki/EPEL/es

Finances Online. (2019). Business Intelligence Software. Recuperado el 14 de marzo de 2019 de https://financesonline.com/

Finances Online. (2019). Business Intelligence Software. Recuperado el 14 de marzo de 2019 de https://business-intelligence.financesonline.com/

Finances Online. (2019). Pentaho. Recuperado el 14 de marzo de 2019 de https://reviews.financesonline.com/p/pentaho/#overview-benefits

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Hidalgo, G. (2016). PROYECTO DE DETECCIÓN DE PATRONES. p 29. Instituto de Postgrado y Educación Continua de la ESPOCH.

Hitachi. (2018). Hitachi Vantara. Recuperado el 14 de marzo de 2019 de https://help.pentaho.com/Documentation/8.2/Products/

Hitachi. (s.f). Hitachi Vantara. Recuperado el 14 de marzo de 2019 de https://help.pentaho.com/Documentation/5.4/0D0/160/000

Hitachi Vantara. (2016). Install Drivers with the JDBC Distribution Tool. Recuperado el 20 de marzo de 2019 de https://help.pentaho.com/Documentation/7.0/0D0/160/010/001

Hitachi Vantara. (2016). Virtual OLAP Cubes. Recuperado el 1 de abril de 2019 de https://help.pentaho.com/Documentation/6.1/0N0/020/040/000

Hitachi Vantara. (2017). Pentaho Dashboard Designer. Recuperado el 5 de abril de 2019 de https://help.pentaho.com/Documentation/8.0/Products/Dashboard_Designer

Hitachi Vantara. (2017). Pentaho Documentation. Recuperado el 10 de abril de 2019 de https://help.pentaho.com/Documentation/5.4/0D0/160/000

Hitachi Vantara. (2018). JDBC Drivers Reference. Recuperado el 12 de abril de 2019 de https://help.pentaho.com/Documentation/7.0/0D0/160/010

Hitachi Vantara. (2018). Pentaho Documentation. Recuperado el 15 de abril de 2019 de https://help.pentaho.com/Documentation/8.2/Products/Aggregation_Designer

Hitachi Vantara. (2018). Pentaho Documentation. Recuperado el 20 de abril de 2019 de https://help.pentaho.com/Documentation/8.2/Products/Schema_Workbench

Hitachi Vantara. (2018). Pentaho Report Designer. Recuperado el 21 de abril de 2019 de https://help.pentaho.com/Documentation/8.2/Products/Report_Designer

Hitachi Vantara. (2018). Petaho Documentation. Recuperado el 21 de abril de 2019 de https://help.pentaho.com/Documentation/8.1/Products/Metadata_Editor

Inoue, H., Amagasa, T., & Kitagawa, H. (2013). An ETL Framework for Online Analytical Processing of Linked Open Data. Recuperado el 21 de abril de 2019 https://link.springer.com/chapter/10.1007/978-3-642-38562-9_12#citeas

Kimball, R., & Caserta, J. (2004). The Data Warehouse ETL Toolkit. Indianapolis: Wiley Publishing, Inc.

Mamani, Y. (2018). Business Intelligence: herramientas para la toma de decisiones en procesos de negocio. Recuperado el 22 de abril de 2019 de https://www.researchgate.net/profile/Yonatan_Mamani/publication/323993348_Business_Intelligence_herramientas_para_la_toma_de_decisiones_en_procesos_de_negocio/links/5ab6bc4ba6fdcc46d3b6b9ee/Business-Intelligence-herramientas-para-la-toma-de-decisiones-en-

Martyn, T. (2004). Reconsidering Mullti-dimensional Schemas. Rensselaer. ACM Sigmod record, 33(1), 83-88.

Microsoft . (2009). Docs. Recuperado el 25 de abril de 2019 de https://docs.microsoft.com/en-us/iis/install/installing-iis-7/install-windows-server-2008-and-windows-server-2008-r2

Microsoft . (2011). Download Center. Recuperado el 28 de abril de 2019 de https://www.microsoft.com/en-us/download/details.aspx?id=11093

Microsoft. (2009). Microsoft Docs. Recuperado el 28 de abril de 2019 de https://docs.microsoft.com/en-us/iis/install/installing-iis-7/install-windows-server-2008-and-windows-server-2008-r2

Microsoft. (2011). Windows Server. Recuperado el 30 de abril de 2019 de https://docs.microsoft.com/en-us/iis/install/installing-iis-7/install-windows-server-2008-and-windows-server-2008-r2

Negash, S., & Gray, P. (2008). Handbook on decision support systems 2. Springer, Berlin, Heidelberg.

Niinimaki, M., & Niemi, T. (2009). An ETL Process for OLAP Using RDF/OWL Ontologies. In journal on data semantics XIII. Springer, Berlin, Heidelberg.

Oracle. (2018). MySQL. Recuperado el 1 de mayo de 2019 de https://www.mysql.com/products/enterprise/database/

Oracle. (2015). Oracle Technology Network. Recuperado el 2 de mayo de 2019 de https://www.oracle.com/technetwork/java/overview-141217.html

Oracle. (s.f). Oracle. Recuperado el 5 de mayo de 2019 de https://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html

Pentaho. (2011). Mondrian Documentation. Recuperado el 5 de mayo de 2019 de https://mondrian.pentaho.com/documentation/schema.php#What_is_a_schema

PowerData. (2013). Procesos ETL La Base de la Inteligencia de Negocio. Barcelona.

Pragmatic. (s.f). Pentaho Features. Recuperado el 10 de mayo de 2019 de http://www.pragtech.co.in/technologies/pentaho/pentaho-community-edition-vs-enterprise-edition.html

Ranjan, V. (2011). A comparative study between ETL (Extract, Transform, Load) and ELT (Extract, Load and Transform) approach for loading data into data warehouse. Recuperado el 12 de mayo de 2019 de http://www. ecst. csuchico. edu/~ juliano/csci693/Presentations/2009w/Materials/Ranjan/Ranjan.

SAP. (2019). SAP PorwerDesigner. Recuperado el 14 de mayo de 2019 de https://www.sap.com/products/powerdesigner-data-modeling-tools.product-capabilities.html

SelectHub. (s.f). IBM Cognos Analytics. Recuperado el 16 de mayo de 2019 de What is IBM Cognos Analytics?: https://selecthub.com/business-analytics-tools/ibm-cognos-analytics/?from_category=69

SelectHub. (s.f). Microsoft Power BI. Recuperado el 18 de mayo de 2019 de https://selecthub.com/business-intelligence-tools/microsoft-bi/?from_category=69

SelectHub. (s.f). MicroStrategy Enterprise Analytics & Mobility. Recuperado el 21 de mayo de 2019 de https://selecthub.com/big-data-analytics-tools/microstrategy-analytics-express/?from_category=69

SelectHub. (s.f). SAP BusinessObjects Business Intelligence. Recuperado el 24 de mayo de 2019 de https://selecthub.com/business-intelligence-tools/sap-business-intelligence/?from_category=69

SelectHub. (s.f). Sisense. Recuperado el 25 de mayo de 2019 de https://selecthub.com/business-intelligence-tools/sisense/?from_category=69

Sergio, L., Juan, T., & Il, Y. (2005). A UML profile for multidimensional modeling in data warehouses. ScienceDirect, pp. 725-769.

Shaker, A., Adeltawab, A., & Ali, H. (2011). A proposed model for data warehouse ETL processes. Journal of king Saud University - Computer and Information Sciences, pp. 92-93.

Sinnexus. (2017). Datawarehouse. Recuperado el 27 de mayo de 2019 de https://www.sinnexus.com/business_intelligence/datawarehouse.aspx

Universidad a Distancia de Madrid. (2014). Recuperado el 31 de mayo de 2019 de https://i.ytimg.com/vi/4VvlgkL25Ks/maxresdefault.jpg

Vargas, A. (2017). Grupo de Investigación GT-IDE UNIANDES. Recuperado el 1 de junio de 2019 de http://uniandesinvestigacion.edu.ec/ide/wp-content/uploads/2017/03/Instalación-Pentaho-1.pdf

Villegas, W., & Luján, S. (2017). Analysis of data mining techniques applied to LMS for personalized education. IEEE, pp. 1-5.

VMware. (2017). Products. Recuperado el 5 de junio de 2019 de https://www.vmware.com/products/esxi-and-esx.html