PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

Análisis de métodos y técnicas de Limpieza de Datos existentes y aplicación en un Sistema CRM para una institución educativa limeña.

Tesis Para optar por el Título de Ingeniero Informático que presenta el bachiller:

Angel Gabriel Sandoval Linares

Asesor: Dr. Héctor Andrés Melgar Sasieta

Lima, Agosto de 2018

A Dios…

A mis padres…

A mi hermano…

A Mons. Adriano Tomassi…

Agradecimientos

A la Institución Educativa y a su
Director, por la buena disposición e
información, tan gentilmente, brindada.

A mi asesor y amigo,
el Dr. Andrés Melgar, por ser un apoyo
cuando más lo necesitaba.

A mi profesor de la universidad,
el Mg. Isaac Yrigoyen, por inculcar en mí
el amor por el área de Sistemas de Información

A todos los que fueron parte de
mi vida universitaria
GRACIAS

Angel Sandoval

Resumen

En la actualidad, las organizaciones emplean varios sistemas y varias fuentes de información para las actividades del día a día, y buscan tener toda esta información reunida e integrada en una única base de datos llamada *data warehouse* ya que permite fortalecer el trabajo del día a día, el análisis de datos y la toma de decisiones.

Sin embargo, la información guardada debe de ser de buena calidad ya que una baja calidad de datos puede impactar severamente en el desempeño de la organización, la satisfacción del cliente, la toma de decisiones y reducir la habilidad de la organización de ejecutar correctamente sus planes estratégicos. En este contexto, aparece un problema crítico: la baja calidad de la información en los sistemas; y lo preocupante es que algunas empresas ignoran los impactos y consecuencias mencionados.

Un sistema de información muy adquirido y usado por organizaciones *Business-to-Consumer* (B2C por su abreviatura en inglés) es el sistema de Gestión de Relación con el Cliente (*Customer Relationship Management* - CRM). Un sistema CRM es un sistema enfocado en la gestión de clientes. Los registros más importantes pertenecen a la entidad "clientes" y esta información es obtenida por las organizaciones a través de varios canales o mediante la compra de bases de datos de terceros. Finalmente, toda la información es almacenada en el *data warehouse* para ser consumida de allí para la toma de decisiones.

Los problemas específicos para un sistema CRM son: registros duplicados de clientes, datos faltantes de un cliente como su teléfono o dirección, datos incorrectos, datos obsoletos que en algún momento fueron correctos y atributos con valores diferentes para un mismo cliente. Mantener estos registros limpios debe ser una actividad vital para la organización.

Las instituciones educativas no son ajenas a esta herramienta de soporte CRM, y con el transcurso de los años, están apostando por adoptar sistemas CRM en las organizaciones (KaptureCRM, 2017). En este contexto, tener los datos de los estudiantes limpios es una tarea primordial para la organización.

El desarrollo de este proyecto se enfoca en un análisis de los algoritmos, técnicas y métodos usados para la limpieza de datos, la implementación de procesos ETL (extracción, transformación y carga) que permitan la limpieza de cada fuente de datos, la integración de la información a una base de datos transaccional, la carga de la información de la base de datos transaccional a un *data warehouse* para su próxima explotación y, adicionalmente, el modelamiento de nuevos procesos de negocio para prevenir y mantener la correcta calidad de los datos en el sistema transaccional, para la institución educativa sobre la cual se realiza el proyecto.

TEMA DE TESIS PARA OPTAR EL TÍTULO DE INGENIERO INFORMÁTICO

TÍTULO:	**Análisis de métodos y técnicas de Limpieza de Datos existentes y aplicación en un Sistema CRM para una institución educativa limeña.**
ÁREA:	Sistemas de Información
ASESOR	Dr. Héctor Andrés MELGAR SASIETA
ALUMNO:	Angel Gabriel SANDOVAL LINARES
CÓDIGO:	20095577
TEMA N°:	_____
FECHA:	San Miguel, 16 de Junio de 2018

DESCRIPCIÓN

En la actualidad, las organizaciones, como instituciones educativas, manejan grandes cantidades de información y varias fuentes de información. Debido a la vasta información presente, la confiabilidad de los datos podría verse comprometida y, por consiguiente, no guiar a la organización a cumplir con sus objetivos de negocio. La información es un recurso muy valioso para las organizaciones y mantenerla con una calidad adecuada es una labor complicada debido a que es un recurso que está en constante actualización. Para una empresa enfocada en sus clientes, el sistema CRM y los datos de los clientes son una herramienta primordial para el negocio y la información aquí guardada debería ser de la mejor calidad posible ya que el “core” de la organización está enfocado a este rubro.

El presente proyecto propone la realización de una técnica que permita mejorar la calidad de los datos y ayudar a que la información se mantenga limpia en los sistemas de la organización.

OBJETIVO GENERAL

El objetivo general del presente proyecto de tesis es proponer una nueva técnica para la limpieza de datos usando una combinación de técnicas que permitan mejorar la calidad de datos de los alumnos en una institución educativa limeña.

OBJETIVOS ESPECÍFICOS

Los objetivos específicos son:

- **OE1:** Integrar y normalizar las entidades de las diversas fuentes y sistemas de información transaccionales en una única base de datos transaccional.
- **OE2:** Diseñar e implementar un algoritmo que permita corregir los datos incorrectos.
- **OE3:** Implementar un módulo de software que permita integrar la información de la nueva base de datos transaccional en un *data warehouse*.
- **OE4:** Implementar y mejorar los procesos de negocio relacionados a los estudiantes para mantener la calidad de los datos en el sistema transaccional.

ALCANCE

El proyecto a desarrollar pertenece al área de sistemas de información. El proyecto consiste en implementar módulos de software que permitan limpiar los datos de los alumnos en institución educativa limeña. Para lograr este objetivo, se implementaron procesos ETL, utilizando una herramienta de software libre llamada Pentaho. El primer ETL permite la limpieza de cada una de las fuentes de datos que maneja la institución educativa. El segundo ETL permite la integración de la información, presente en varias fuentes de datos, a una base de datos transaccional. El tercer ETL permite la carga de la información de la base de datos transaccional a un *data warehouse* para su próxima explotación. Adicionalmente, se proponen nuevos procesos de negocio para prevenir y mantener la correcta calidad de los datos. Las entidades de datos enfocadas en el presente proyecto serán las relacionadas con los sistemas CRM de la institución educativa, específicamente la entidad clientes y los campos principales de la entidad. Para lograr cumplir con lo establecido, se contará con dos bases de datos de distintos proveedores y un archivo Excel con distintos datos y atributos con el fin de integrarlos, y se aplicará un *framework* para el proceso de limpieza.

Tabla de Contenidos

Índice de Imágenes

Índice de Tablas

Capítulo 1. Generalidades

1.1 Problemática

En esta Sección se establece y describe la problemática del presente proyecto de fin de carrera. En primer lugar, se inicia con una caracterización del contexto para situar el problema. Luego, se presenta el problema general que se intenta solucionar. Posteriormente, se deriva el problema en uno más específico enfocado al área de CRM y a las instituciones educativas. Finalmente, se presenta el árbol de problema para identificar el problema, sus causas y los efectos que trae consigo.

1.1.1 Definición del Problema

El mundo se encuentra en la era de la información, era en la cual, el uso de la información y el uso de los canales digitales (formularios en la Web, encuestas en la Web, mails, redes sociales, entre otros) son cada vez más usados por las personas (Pareja & Echeverría, 2014).

Las organizaciones no son ajenas a este recurso y, con el transcurso de los años, desean explotar toda esta información. Para ello, las organizaciones se están apoyando en sistemas transaccionales (*Transaction Processing Systems* - TPS), enfocados en el nivel operativo de la organización y en guardar transacciones de negocio; sistemas de planificación de recursos empresariales (*Enterprise Resource Planning* - ERP), enfocados en administrar las operaciones de negocio; sistemas gerenciales (*Management Information Systems* - MIS), enfocados en proveer información de rutina a administradores en la eficiencia del nivel operativo; y sistemas para toma de decisiones (*Decision Support System* - DSS), enfocados en problemas específicos de naturaleza cambiante y en el apoyo para la toma de decisiones (Stair & Reynolds, 2016). Estos últimos ayudan a generar una ventaja competitiva porque proveen de información importante para la toma de decisiones (Song, Liu, Wu, & Bao, 2015). Los sistemas mencionados tienen muchas utilidades y apoyan a las distintas áreas de la organización: desde el área operativa hasta el área de la alta dirección y son cada vez más grandes e involucran grandes volúmenes de datos (Lucas, Raja, & Ishfaq, 2014).

Asimismo, lejos ha quedado la idea de los sistemas monolíticos (Batini, Cappiello, Francalanci, & Maurino, 2009). En la actualidad, las organizaciones emplean varios sistemas y varias fuentes de información, y buscan tener toda esta información

reunida e integrada en una única base de datos llamada *data warehouse* (Lucas et al., 2014). El *data warehouse* permite fortalecer el trabajo del día a día, el análisis de datos y la toma de decisiones (Song et al., 2015).

Sin embargo, la información guardada debe de ser de buena calidad ya que una baja calidad de datos puede impactar severamente en el desempeño de la organización, la satisfacción del cliente, la toma de decisiones y reducir la habilidad de la organización de ejecutar correctamente sus planes estratégicos (Lucas et al., 2014). Asimismo, un estudio concluyó que los datos inconsistentes, inexactos e inaccesibles ocasionaban en las empresas estadounidenses costos de hasta 600 billones de dólares por año (Eckerson, 2011). En este contexto, aparece un problema crítico: la baja calidad de la información en los sistemas; y lo preocupante es que algunas empresas ignoran sus impactos y consecuencias.

Un sistema de información muy adquirido y usado por organizaciones *Business-to-Consumer* (B2C por su abreviatura en inglés) es el sistema de Gestión de Relación con el Cliente (*Customer Relationship Management* - CRM). Un sistema CRM es un sistema enfocado en la gestión de clientes (Faed, Wu, & Chang, 2010). Los registros más importantes pertenecen a la entidad “clientes” y esta información es obtenida por las empresas a través de varios canales o mediante la compra de bases de datos de terceros. Las campañas de marketing usan la información del cliente para establecer el público objetivo y para ofertar productos que sean atrayentes (Kumar & Reinartz, 2012). Finalmente, toda la información es almacenada en el *data warehouse* para ser consumida de allí para la toma de decisiones.

Los problemas específicos para un sistema CRM son: registros duplicados de clientes, datos faltantes de un cliente como su teléfono o dirección, datos incorrectos, datos obsoletos que en algún momento fueron correctos y atributos con valores diferentes para un mismo cliente (Batini et al., 2009). Mantener estos registros limpios debe ser una actividad vital para la organización. En las organizaciones que utilizan sistemas CRM, un estudio demostró que 2% de los registros de clientes se vuelven obsoletos en 1 mes, lo que podría significar que, en una base de datos de 500 mil clientes, 10 mil se vuelven obsoletos por mes y 120 mil por año (Fan, Geerts, & Wijsen, 2011). El estudio previo muestra cifras preocupantes ya que, en tan solo 2 años, la mitad de la base de datos de clientes tendrá datos obsoletos.

Las instituciones educativas no son ajenas a esta herramienta de soporte, y con el transcurso de los años, están apostando por adoptar sistemas CRM en las organizaciones (KaptureCRM, 2017). Según un estudio realizado por la *American Association of Collegiate Registrars and Admissions Officers* (AACRAO), se llegó a la conclusión que el 64% de las organizaciones del sector educativo utilizan sistemas CRM en los Estados Unidos; y el 42% de las organizaciones que no cuentan con uno, están considerando adquirir uno en el corto plazo (AACRAO, 2016). En este contexto, tener los datos de los estudiantes limpios es una tarea primordial para estas organizaciones.

En la actualidad, existen algoritmos y técnicas para solucionar este problema, tales como: adquisición de nuevos datos, normalización de la base de datos, integración de registros duplicados, entre otros. Sin embargo, las soluciones disponibles en el mercado cuestan mucho dinero ya que empresas líderes a nivel mundial como Oracle y Microsoft ofrecen el servicio, e implementar estas herramientas suponen un riesgo para las empresas por el alto costo que tienen y la posibilidad que el proyecto no culmine correctamente.

El desarrollo de este proyecto se enfoca en un análisis de los algoritmos, técnicas y métodos usados para la limpieza de datos, la implementación de procesos ETL (extracción, transformación y carga) que permitan la limpieza de cada fuente de datos, la integración de la información a una base de datos transaccional, la carga de la información de la base de datos transaccional a un *data warehouse* para su próxima explotación y, adicionalmente, el modelamiento de nuevos procesos de negocio para prevenir y mantener la correcta calidad de los datos en el sistema transaccional, para la institución educativa sobre la cual se realiza el proyecto.

1.1.2 Árbol del Problema

La imagen 1 presenta el árbol de problema para el presente proyecto de fin de carrera.

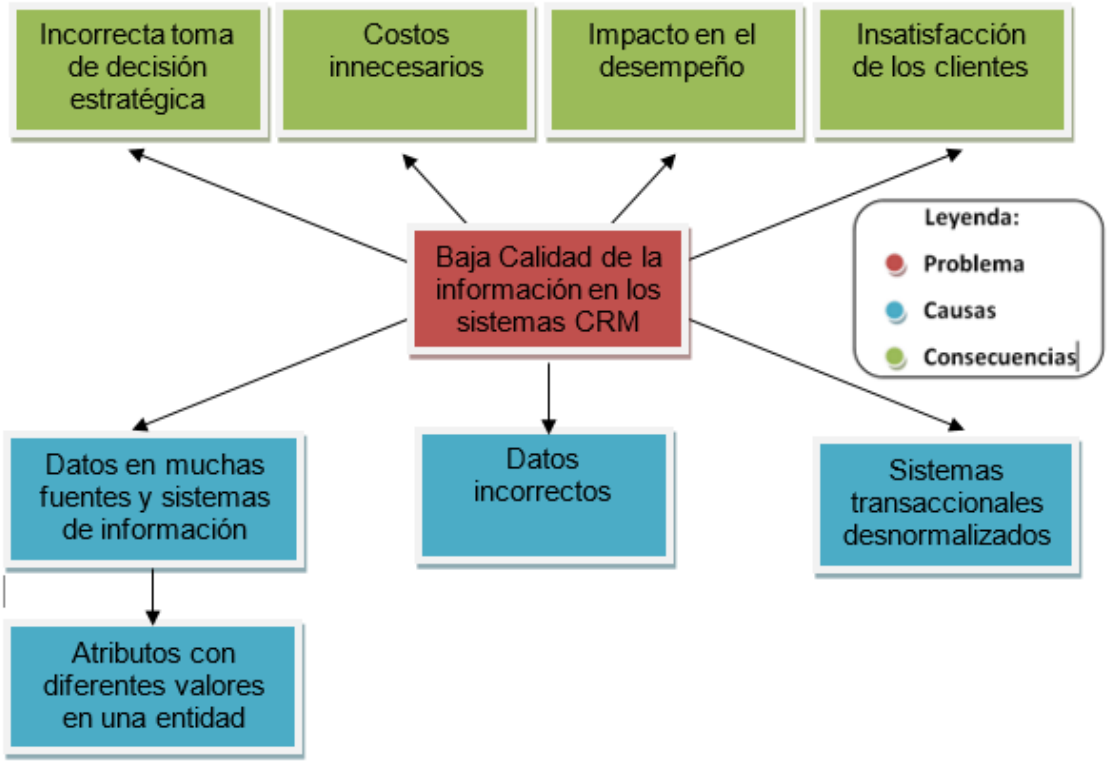

Imagen 1: Árbol del Problema. Imagen de autoría propia.

1.2 Objetivos

1.2.1 Objetivo general

Proponer una nueva técnica para la limpieza de datos usando una combinación de técnicas que permita mejorar la calidad de datos de los alumnos de una institución educativa limeña.

1.2.2 Objetivos específicos

- **OE1:** Integrar y normalizar las entidades de las diversas fuentes y sistemas de información transaccionales en una única base de datos transaccional.
- **OE2:** Diseñar e implementar un algoritmo que permita corregir los datos incorrectos.
- **OE3:** Implementar un módulo de software que permita integrar la información de la nueva base de datos transaccional en un *data warehouse*.
- **OE4:** Implementar y mejorar los procesos de negocio relacionados a los estudiantes para mantener la calidad de los datos en el sistema transaccional.

1.2.3 Resultados esperados

- **R1-1:** 2 bases de datos transaccionales de distintos proveedores y un archivo en Excel. (OE1)
- **R1-2:** Base de datos transaccional normalizada y unificada. (OE1)
- **R2-1:** Algoritmo de limpieza de datos. (OE2)
- **R2-2:** 3 ETL para la lectura de datos de cada fuente de datos, limpieza y carga a las fuentes de datos transaccionales. (OE2)
- **R2-3:** ETL para la lectura de datos de las fuentes de datos, integración y carga a la base de datos transaccional unificada. (OE2)
- **R3-1:** ETL para la lectura de datos de la base de datos transaccional, integración y carga al *data warehouse*. (OE3)
- **R3-2:** Creación de *data warehouse* y *data marts*. (OE3)
- **R3-3:** Información de las fuentes de datos transaccionales integrada en un *data warehouse*. (OE3)
- **R4-1:** Modelado de procesos. (OE4)
- **R4-2:** Mejora de procesos para prevenir la aparición de datos con mala calidad. (OE4)

1.3 Mapeo de objetivos, resultados, verificación y herramientas

Tabla 1: Tabla de Objetivos y Resultados. Tabla de autoría propia.

Objetivo 1: Implementar un módulo de software que permita integrar y normalizar las entidades de las diversas fuentes y sistemas de información transaccionales en una única base de datos transaccional.		
Resultado	**Meta física**	**Medio de verificación**
2 bases de datos transaccionales de distintos proveedores y un archivo en Excel.	1 Base de Datos SQL Server 1 Base de Datos MySQL 1 Archivo Excel	Diagrama DER Diccionario de Datos
Base de datos transaccional normalizada y unificada.	Base de Datos MySQL con data cargada	Diagrama DER
Objetivo 2: Diseñar e implementar un algoritmo que permita corregir los datos incorrectos.		
Resultado	**Meta física**	**Medio de**

		verificación
Algoritmo de limpieza de datos.	Algoritmo Descripción del Algoritmo	Reporte de Inconsistencias y Modificaciones.
ETL para la lectura de datos de las fuentes de datos, limpieza y carga.	Aplicativo de Limpieza de Excel Aplicativo de Limpieza de MySQL Aplicativo de Limpieza de SQL Server	Ejecución de procesos ETL Pruebas de los Aplicativos
ETL para la lectura de datos de las fuentes de datos, integración y carga a la base de datos transaccional.	Aplicativo de Integración en Base de Datos Transaccional	Ejecución de proceso ETL
Objetivo 3: Implementar un módulo de software que permita integrar la información de la nueva base de datos transaccional en un *data warehouse*.		
Resultado	**Meta física**	**Medio de verificación**
ETL para la lectura de datos de la base de datos transaccional, integración y carga al *data warehouse*.	Aplicativo de Integración en *Data Warehouse*	Ejecución de proceso ETL
Creación de *data warehouse*.	1 Data Warehouse MySQL 3 Data Marts	Diagrama DER Diccionario de Datos
Información de las fuentes de datos transaccionales integrada en un *data warehouse*.	1 Data Warehouse MySQL con data cargada 3 Data Marts	Diagrama DER
Objetivo 4: Implementar y mejorar los procesos de negocio relacionados a los estudiantes para mantener la calidad de los datos en el sistema transaccional.		
Resultado	**Meta física**	**Medio de verificación**
Modelado de procesos.	Procesos de negocio legacy	Validación por experto
Mejora de procesos para prevenir la aparición de datos con mala calidad.	Procesos de negocio mejorados	Validación por experto

Tabla 2: Tabla de Herramientas. Tabla de autoría propia.

Resultado Esperado	Herramienta
2 bases de datos transaccionales de distintos proveedores y un archivo en Excel.	• MySQL 5.7 • SQL Server 2016 Express Edition • Hojas de Cálculo • Diagrama Entidad Relación (DER) Aplicaciones: • MySQL Workbench • Microsoft SQL Server Management Studio (SSMS) • Microsoft Excel
Base de datos transaccional normalizada y unificada.	• MySQL 5.7 • Diagrama Entidad Relación (DER) Aplicación: • MySQL Workbench
Algoritmo de Limpieza de Datos.	• BPMN Aplicación: • Bizagi Modeler
Creación de *data warehouse*.	• MySQL 5.7 • Diagrama Entidad Relación (DER) Aplicación: • MySQL Workbench • Draw.io
ETL	• Data Integration • Report Designer • MySQL Connector (Library) • SQLJDBC (Library) • Apache Commons Lang (Library) • Jaro – Winkler Algorithm Aplicación: • Pentaho Data Integration - Kettle
Procesos de negocio.	• BPMN Aplicación: • Bizagi Modeler

En el Anexo 1, se presenta una breve descripción de cada herramienta utilizada. En el Anexo 2, se presentan los métodos y procedimientos a utilizar para el desarrollo del proyecto.

1.4 Alcance y Justificación

1.4.1 Alcance

El proyecto a desarrollar pertenece al área de sistemas de información. El proyecto consiste en implementar módulos de software que permitan limpiar los datos de los alumnos en institución educativa limeña. Para lograr este objetivo, se implementaron procesos ETL, utilizando una herramienta de software libre llamada Pentaho. El primer ETL permite la limpieza de cada una de las fuentes de datos que maneja la institución educativa. El segundo ETL permite la integración de la información, presente en varias fuentes de datos, a una base de datos transaccional. El tercer ETL permite la carga de la información de la base de datos transaccional a un *data warehouse* para su próxima explotación. Adicionalmente, se proponen nuevos procesos de negocio para prevenir y mantener la correcta calidad de los datos. Las entidades de datos enfocadas en el presente proyecto serán las relacionadas con los sistemas CRM de la institución educativa, específicamente la entidad clientes y los campos principales de la entidad. Para lograr cumplir con lo establecido, se contará con dos bases de datos de distintos proveedores y un archivo Excel con distintos datos y atributos con el fin de integrarlos, y se aplicará un *framework* para el proceso de limpieza.

1.4.2 Justificación

En la actualidad, la información es un recurso muy valioso para las instituciones educativas y mantenerla con una calidad adecuada es una labor complicada debido a que es un recurso que está en constante actualización. Para una organización enfocada en los clientes, el sistema CRM y los datos de los clientes son la herramienta principal para hacer negocio y la información aquí guardada debería ser de la mejor calidad posible ya que el "*core*" de la organización está enfocado a este rubro.

Con el presente proyecto de fin de carrera, la institución educativa se verá beneficiada ya que contará con sus procesos principales modelados y mejorados, contará con datos de alumnos limpios, tendrá integrada su información en una única base de datos, y tendrá un *data warehouse*, del cual podrá, próximamente, explotar información.

1.5 Viabilidad

1.5.1 Viabilidad Temporal

A continuación, en la imagen 2, se presenta el diagrama de Gantt a utilizar para el desarrollo del proyecto de fin de carrera.

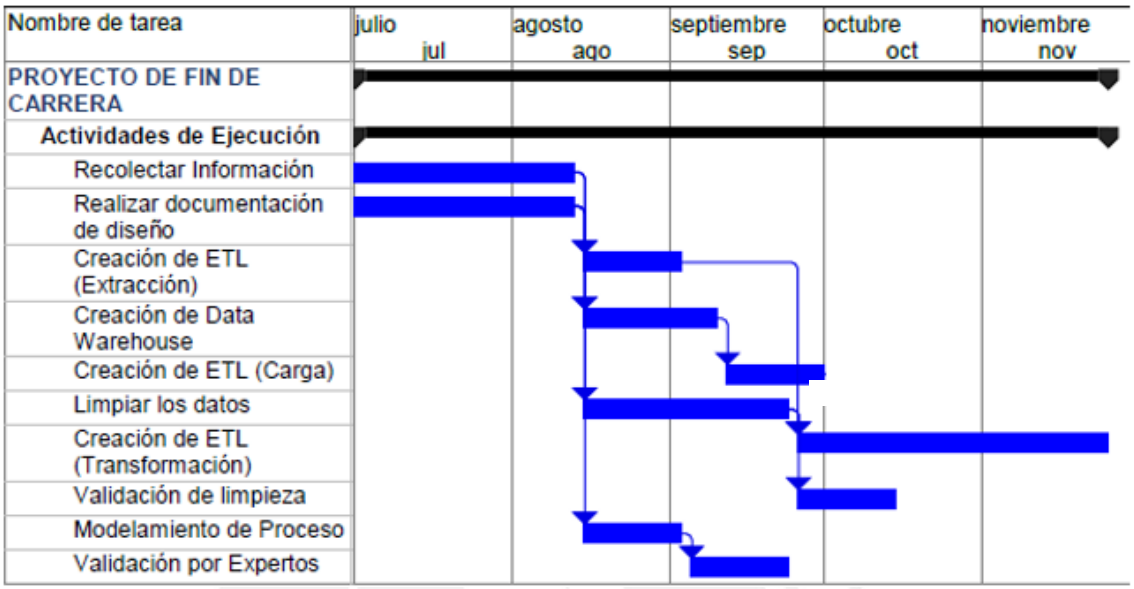

Imagen 2: Diagrama de Gantt. Imagen de autoría propia.

Se puede concluir que el proyecto es viable temporalmente ya que el tiempo para su desarrollo no excede al tiempo establecido.

1.5.2 Viabilidad Económica

Todas las herramientas usadas y propuestas en el proyecto son de uso libre. Por tanto, la organización no tendría que incurrir en costos adicionales.

1.5.3 Viabilidad Técnica

El software Pentaho se encuentra desarrollado en Java y cuenta con compatibilidad en sistemas operativos Microsoft, Linux y Mac OS ya que Java cuenta con una máquina virtual JVM.

Capítulo 2. Marco Conceptual

En esta Sección de detalla el marco conceptual. El marco conceptual permite aclarar los conceptos y la problemática expuesta: la baja calidad de la información en las instituciones educativas.

2.1 Data Warehouse (DW)

Un *data warehouse* es una organización compleja que guarda grandes cantidades de datos a partir de varias fuentes de información o bases de datos (Arora, Pahwa, & Bansal, 2009)(Batini et al., 2009). Es definido como una colección de datos orientados a temas, integrados, variantes en el tiempo y no volátiles que apoya al proceso de toma de decisiones (Inmon, 1980).

2.2 CRM

CRM son las siglas de Gestión de Relación con el Cliente (del inglés: *Customer Relationship Management*) y es una estrategia de negocio que consiste en crear y mantener, en el largo plazo, la relación con los clientes (Faed et al., 2010). CRM es un componente importante del marketing que los expertos toman en consideración en los últimos años (Zhou, Zhang, & Lu, 2011), debido a que el cliente se ha convertido en un consumidor experimentado, exigente y que requiere de un trato personalizado y la empresa no puede ser ajena a este cambio (Petkovic, 2010). Es así como el cliente se convierte en el tema más importante para las organizaciones modernas (Petkovic, 2010).

La satisfacción del cliente implica resultados positivos, en forma de volúmenes de compra mayores, compras repetitivas y generación de nuevos negocios en forma de referencias e identificación de clientes potenciales (*prospects*) (Al-Mudimigh, Ullah, & Saleem, 2009), es decir, empresas más rentables en el largo plazo (Kumar & Reinartz, 2012) y con mejores oportunidades. CRM ayuda a las organizaciones a centrarse en su activo más importante, en los clientes, cambiando de una perspectiva basada en el producto a una basada en sus clientes (Petkovic, 2010). Sin embargo, no es recomendable trabajar sobre todos los clientes de la misma forma ya que estos no tienen las mismas características (Kumar & Reinartz, 2012). Para eso, CRM tiene un proceso llamado segmentación de clientes que consiste en dividir la cartera de clientes en grupos homogéneos de acuerdo a sus

características tales como: su comportamiento de compra, sus gustos, sus deseos y sus necesidades (Zhou et al., 2011).

Para caracterizar correctamente a los clientes, las organizaciones deben recolectar información relevante de los consumidores con el objetivo de entender su comportamiento y sus respectivas necesidades (Petkovic, 2010)(Al-Mudimigh, Saleem, Ullah, & Al-Aboud, 2009). CRM consiste en la utilización de esta información para ofrecer productos destacados (Levine, 2000)(Al-Mudimigh, Ullah, et al., 2009), ofertas pertinentes y ofrecer un valor único al cliente.

2.2.1 Campaña de Marketing

Una campaña de marketing es una herramienta del marketing que es usada para identificar clientes o crear oportunidades que finalmente resultarán en ventas, reconocimiento de la marca u otro tipo de respuestas (Oracle, 2013). Una campaña de marketing contiene contactos, clientes potenciales, ofertas, actividades, entre otros elementos (Oracle, 2013).

Una campaña de marketing consiste en una serie de actividades, dirigidas a un segmento de clientes o clientes potenciales específicos, usadas para promocionar un producto o servicio, a través de ofertas, usando un canal o una serie de canales (Kumar & Reinartz, 2012) (Oracle, 2013).

Una campaña exitosa involucra llegar al cliente correcto, con la oferta correcta, en el tiempo correcto y por el canal correcto (Kumar & Reinartz, 2012). La presencia de estas variables aumenta en forma considerativa la oportunidad de venta para la empresa (Kumar & Reinartz, 2012).

La administración de campañas tiene cuatro fases (Kumar & Reinartz, 2012):

- Planeación: Los objetivos de la campaña son definidos.
- Desarrollo: Se crea la oferta, se escoge el canal o canales a usar y se escogen los miembros de campaña.
- Ejecución: Proceso operacional de lanzar la campaña y controlarla.
- Análisis: Se evalúan los resultados de la campaña contrastándolos con los objetivos.

2.2.2 Canal de Marketing

Los canales son los medios por dónde se transmite la información y pueden ser: páginas Web, *e-commerce*, tiendas, *call-centers*, redes sociales, mails, vendedores

(*sales representatives*), entre otros (Petkovic, 2010)(Kumar & Reinartz, 2012). Los clientes usan diversos canales para comunicarse con la organización, es decir, responder a una campaña, pedir información, plantear sus necesidades, plantear una disconformidad o comprar un producto, y la empresa debe estar preparada para atender al cliente en el tiempo y por el canal adecuados. (Oracle, 2013)(Kumar & Reinartz, 2012).

Los clientes pueden cambiar sus hábitos en el uso de canales. Esta es la razón por la que la organización debe realizar un estudio sobre sus clientes y la empresa debe estar atenta a responder correctamente por el canal adecuado (Kumar & Reinartz, 2012).

2.2.3 Oferta

Una oferta es una proposición o mensaje dirigida a un cliente (Oracle, 2013). El uso de una oferta pertinente para el consumidor generará un incentivo para comprar el producto o preguntar por más información (Kumar & Reinartz, 2012).

Una oferta tiene un producto involucrado, un precio especial (o una lista de precios), un conjunto de canales de marketing asociados y una serie de beneficios y condiciones. Por ejemplo, una oferta podría ser una tarjeta de crédito con línea de crédito de 30 mil soles y una tasa de interés de 39% anual, el producto sería la tarjeta de crédito, el canal asociado podría ser banca telefónica, el beneficio sería la acumulación de puntos y finalmente las condiciones podrían ser sujetos con evaluación crediticia en verde (buenos pagadores de deudas).

2.2.4 Público Objetivo

El público objetivo lo conforman los contactos que serán parte de la campaña (también llamados miembros de campaña) y es escogido durante la fase de planeación de la campaña (Kumar & Reinartz, 2012). Un miembro de campaña puede ser un cliente potencial o un cliente existente que podría transformarse en una oportunidad (Oracle, 2013). Por tanto, la organización tiene tres estrategias: seguir una estrategia de retención de clientes existentes, seguir una estrategia de adquisición de nuevos clientes o seguir una estrategia mixta (Kumar & Reinartz, 2012).

Se puede observar en la imagen 3, cómo se relacionan los conceptos de CRM previamente descritos.

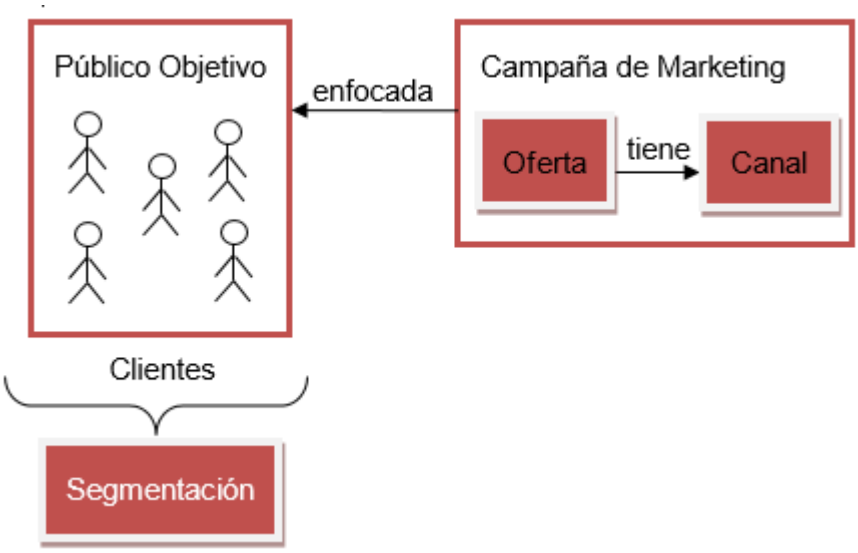

Imagen 3: Resumen de conceptos de CRM. Imagen de autoría propia.

2.3 Instituciones Educativas

Es un conjunto de personas y bienes promovida por las autoridades públicas o por particulares, cuya finalidad será prestar un año de educación preescolar y nueve grados de educación básica como mínimo y la media (Minedu, n.d.). Para el caso del proyecto de fin de carrera, la institución educativa será un colegio de Lima que brinda educación primaria y secundaria.

2.3.1 Profesor

Persona que ejerce o enseña una ciencia o arte (RAE, 2014). Para el caso de la institución educativa, el profesor es aquella persona que dicta un curso en específico para un grado en un bimestre académico.

2.3.2 Estudiante

Persona que cursa estudios en un establecimiento de enseñanza (RAE, 2014). Para el caso de la institución educativa, el estudiante es aquella persona que está matriculado en el año escolar y que lleva los cursos estipulados por la institución educativa.

2.4 Conclusión

Se puede concluir que los conceptos mostrados permiten entender la problemática expuesta: la baja calidad de la información en las instituciones educativas, específicamente, dentro de una base de datos de clientes. Asimismo, los conceptos

describen y clarifican conceptos relacionados a los sistemas CRM y a las instituciones educativas.

Capítulo 3. Estado del Arte

En esta Sección se detallan las últimas investigaciones realizadas en el campo académico. Para el estado del arte se utiliza una metodología de investigación bibliográfica llamada Revisión Sistemática. La Revisión Sistemática es una metodología que resume y sintetiza la evidencia existente relacionada a un campo de interés (Biolchini, Mian, Natali, & Travassos, 2005) y ayuda a identificar vacíos para sugerir áreas de investigación (Kitchenham & Charters, 2007). Al ser una metodología, cuenta con una serie de pasos específicos para su elaboración y de esa manera proporciona una menor parcialidad con una estrategia de búsqueda exhaustiva y explícita (Melgar, 2013).

3.1 Resultados encontrados

En el Anexo 3, se encuentra el método usado para la revisión del estado del arte. En la imagen 4: “Número de Estudios Primarios por Año después de la Ejecución del Criterio de Selección”, se presentan las publicaciones que serán consideradas como estudios primarios para esta revisión. Después de un análisis de los estudios primarios, se pudo identificar 20 estudios primarios usando el criterio de selección de inclusión y exclusión detallado en la sección 3.1. Asimismo, se encontró que en el año 2011 se produjo un incremento en la cantidad de estudios primarios publicados, llegando a 5 publicaciones. Para el año 2008 solo se encontró 1 publicación. Sin embargo, en promedio se encontró 2.5 publicaciones por año, un número bastante pequeño, lo cual confirma lo mencionado en los estudios primarios que hay poca investigación sobre la limpieza y calidad de datos relacionada a un sistema en específico. Los estudios primarios que fueron excluidos no aportaban información relevante al proceso de investigación.

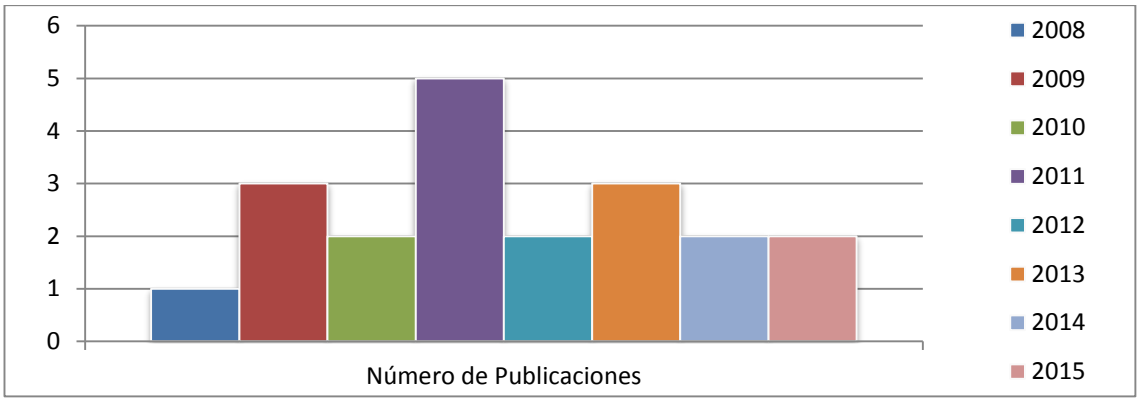

Imagen 4: Número de Estudios Primarios por Año después de la Ejecución del Criterio de Selección. Cuadro de autoría propia.

A continuación, se procede a responder a las preguntas de revisión usando los estudios primarios.

- **Q1: ¿Cuáles son los principales algoritmos, herramientas y técnicas usados en la limpieza de datos?**

  Se encontró que existe una serie de algoritmos, técnicas y métodos usados para solucionar los problemas más comunes de la calidad de datos: inexactitud, inconsistencia e inaccesibilidad (Mezzanzanica, Boselli, Cesarini, & Mercorio, 2015) traducido en datos ubicados en diferentes sistemas, registros duplicados, datos faltantes, datos incorrectos, datos obsoletos y atributos con valores diferentes para una misma entidad (Mezzanzanica et al., 2015)(Bleiholder & Naumann, 2008)(Fan et al., 2011)(Khan, Rauf, Shah, & Khusro, 2011).

  Existen 3 fases para lograr calidad en los datos: (i) la primera fase es la reconstrucción del estado, que consiste en obtener información contextual relacionada a los procesos de la organización; (ii) la segunda fase es la evaluación o medición, que consiste en analizar los datos y comparar los resultados de ese análisis con valores referenciales; (iii) la tercera y última fase es la mejora, que consiste en evaluar los costos, identificar las causas de los errores y escoger estrategias y técnicas para lograr datos de buena calidad (Batini et al., 2009).

  Según Batini y otros, las estrategias se dividen en dos: basadas en los datos, que se enfocan en modificar el valor del dato obsoleto por uno más actual; y basadas en el proceso, que se enfocan en rediseñar los procesos que

manipulan los datos (Batini et al., 2009). Sin embargo, no es una división tajante ya que existen estrategias que pertenecen a ambos grupos.

Asimismo, Liu y otros, establecen otra división para las estrategias, afirmando que existen 3 tipos de estrategias para limpiar y unificar la data duplicada proveniente de varias fuentes de información: *De-Duplication Prior Strategy* (DSS), la cual se ejecuta antes de invocar al proceso ETL (Extracción, Transformación y Carga) (proceso en el que se extraen datos de las distintas fuentes, se limpia, se personaliza y se inserta en un *data warehouse*); *Real Time Scheduling* (RS), la cual se ejecuta inmediatamente después que el proceso ETL terminó; y *ETL Prior Scheduling* (EPS), la cual se ejecuta solo cuando se necesita (Song et al., 2015). Cada uno de estos puede ser usado para solucionar el problema dependiendo de las necesidades específicas.

A continuación, se establecen los métodos y técnicas más usados:

El método más usado es de reglas de negocio; sin embargo, depende mucho de la organización y los tipos de datos que posea (Prasad et al., 2011).

Las estrategias más comunes son: adquisición de nueva data, normalización de la base de datos, integración de registros duplicados, integración de esquemas y datos, uso de fuentes confiables, localización y corrección de errores y optimización de costos (Batini et al., 2009).

Si la organización cuenta con muchas fuentes de información, un método utilizado es la integración de registros, en la cual, los registros provenientes de varias fuentes tienen su propia información. Se utilizan operadores "*join*" y "*union*" y técnicas relacionadas con estos operadores. El enfoque "*Join*" completa todos los atributos faltantes usando las otras fuentes de información. Lo negativo de este enfoque es que al usar un "*full outerjoin*" no se puede garantizar que todos los registros posean todos los atributos, teniendo así, registros con datos faltantes. El enfoque "*Union*" devuelve todas las tuplas de las relaciones. Este enfoque, a diferencia del enfoque "*Join*", no es tan conciso. Otras técnicas son: "*Considering All Posibilities*" y "*Considering only Consistent Possibilities*". Estas técnicas no están basadas en "joins" ni en "unions" y se basan en extender el modelo relacional (Bleiholder & Naumann, 2008). A continuación, en la Tabla 3 se presenta un ejemplo de los enfoques "*join*" y "*union*" usando los datos de un cliente.

Tabla 3: Ejemplos de enfoques para fusión de registros. Tabla adaptada de: Bleiholder & Naumann, 2008.

Enfoque "*Join*"	SELECT U1.Name, U2.Name, U1.Age, U2.Age, U1.Status, U2.Status, U1.Address, U2.Address, U1.Field, U2.Field, U1.Library, U2.Phone FROM U1 FULL OUTER JOIN U2 ON U1.Name=U2.Name
Enfoque "*Union*"	(SELECT Name, Age, Status, Address, Field, Library, NULL as Phone FROM U1) UNION (SELECT Name, Age, Status, Address, Field, NULL as Library, Phone FROM U2)

Asimismo, se puede modificar el proceso ETL para que éste tenga en consideración la calidad de los datos. En la imagen 5, se puede observar cómo sería el flujo del proceso ETL (Lucas et al., 2014).

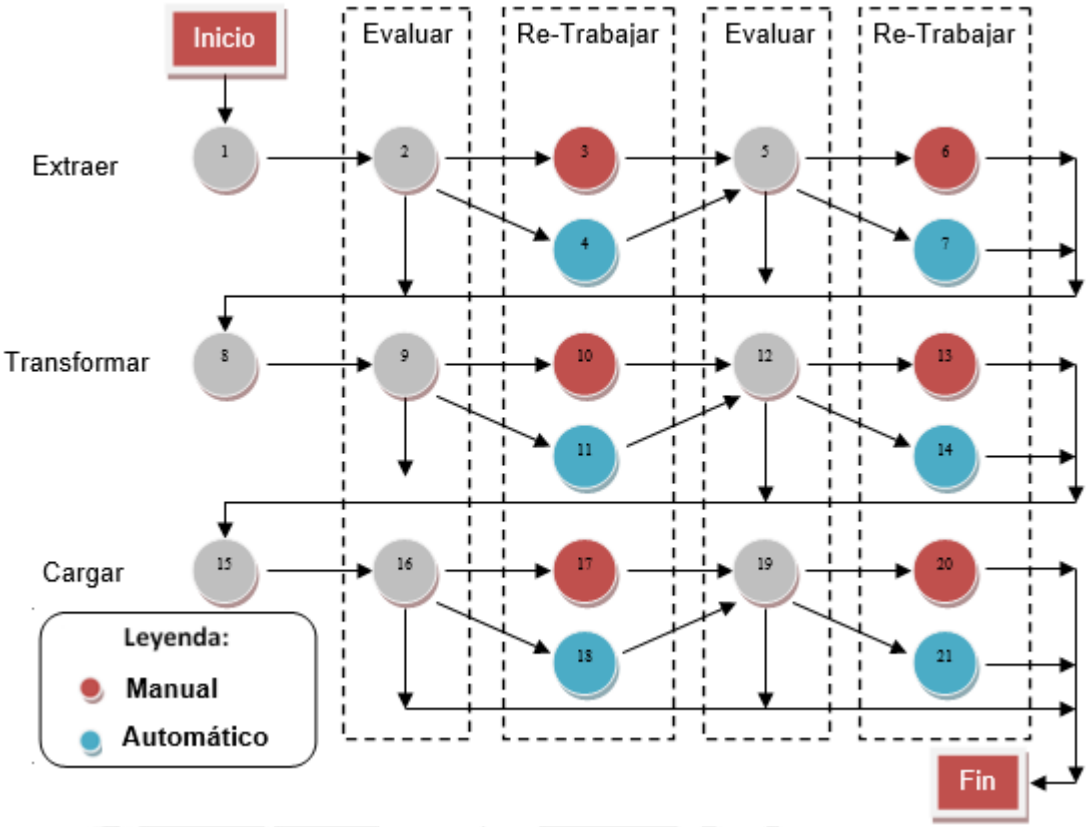

Imagen 5: Proceso ETL. Imagen de adaptada de: Lucas et al., 2014.

Ali y Warraich proponen un *framework* pragmático, interactivo y sencillo que está basado en bases de datos de configuración de reglas (RCDB por sus siglas en inglés). Este *framework* implica que todo el proceso de limpieza se realizará dentro del proceso ETL y se realizará antes de que la data ingrese al *data warehouse*. El *framework* tiene dos partes: la limpieza usando la configuración de reglas y el proceso de limpieza (Ali & Warraich, 2010). En la imagen 6, se puede observar el procedimiento propuesto.

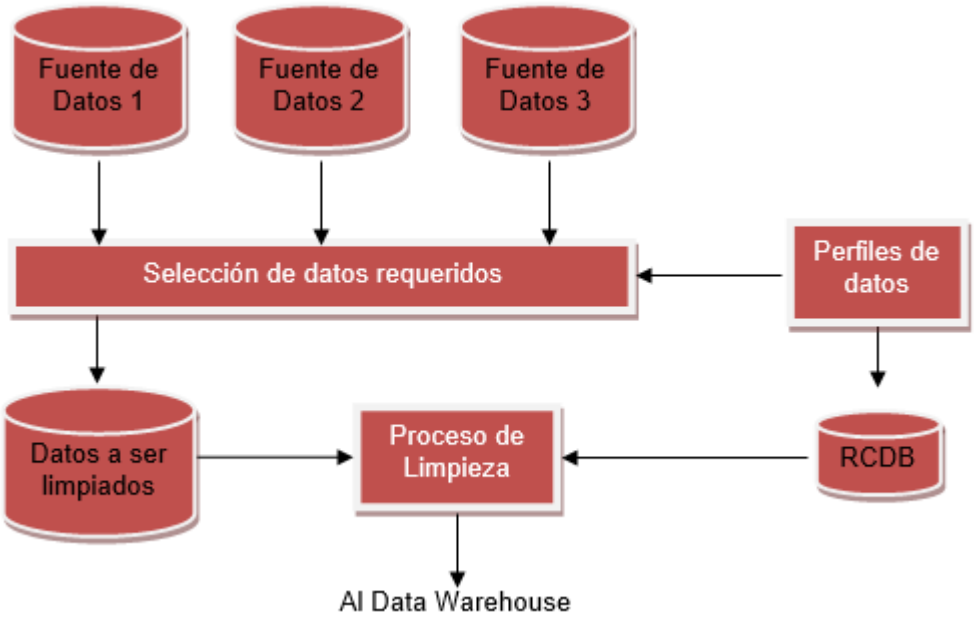

Imagen 6: Framework de limpieza. Imagen de adaptada de: Ali & Warraich, 2010.

Otro método utilizado es interpretar un registro de una base de datos como un evento y un conjunto de registros ordenados como una secuencia de eventos para, de esta forma, utilizar las técnicas de verificación usadas en sistemas basados en eventos (Mezzanzanica et al., 2015).

Finalmente, se presentan dos algoritmos propuestos por Khan y otros. El primero tiene como objetivo homogeneizar el formato; y el segundo tiene como objetivo limpiar los registros duplicados.

Algoritmo 1:

```
Begin
For attribute j=1 to last attribute, n
          For row i=1 to last row, m
          Bring attribute values into uniform format
                Remove the special character
                Remove the variation of attribute values
                Expand abbreviations
                Convert all the values into numeric form
                Put the numeric value into appended
                attribute separated with comma
          End For
     End For
End
```

Algoritmo 2:

```
Begin
      For row=1 to last row, n
            (1)For v, p=1 to q
                  (2)If v>1 then
                        Go to step 7
                  (3)Else Compare all the value with the
                  corresponding value of other row
                  End If
                  (4)If match found b/w values then
                        (5)Calculate the d
                        Go to step 8
                  (6)Else
                        Go to step 13
                        (7)Divide (p+q)/2
                        Go to Step 2
                  End If
                  (8)If d=100% then
                        Discarded the duplicated record
                        Go to End
                  (9)Else if d>=? Then
                        (10)Display the records
                        Mention the attributes values
                        having difference b/w values
                  End If
                  (11)If difference is due to data quality
                  then
                        Correct the entities
                        Go to Step 8
                  (12)Else Go to End
            End For
      End For
End
```

- **Q2: ¿Cómo afectan los datos faltantes, inconsistentes e incorrectos a los sistemas CRM?**

  Se encontró que existe una serie de problemas y costos relacionados a no tener buena calidad en los datos.

  Si bien tener data inconsistente e incorrecta en una base de datos transaccional puede no significar un problema severo para las actividades operativas, sí supone un riesgo cuando se utilizan datos consolidados. Toda la información de las actividades del día a día, es integrada en una base de datos común llamada *data warehouse* para ser usado como almacén general de información. A partir de éste almacén de datos, se obtienen los datos consolidados para la toma de decisiones (Mezzanzanica et al., 2015)(Song et al., 2015)(Lucas et al., 2014)(Khan et al., 2011)(Ali & Warraich, 2010) y así evaluar estrategias.

Asimismo, tener data inconsistente tiene un impacto severo en la performance de los sistemas de la organización y en la satisfacción de los clientes (Lucas et al., 2014). El cliente no recibe la información necesaria por el canal correcto, ya que éste, se encuentra desactualizado. Por ejemplo, un banco cuenta con las direcciones de sus clientes en su base de datos; sin embargo, éstas al no estar actualizadas, el cliente nunca recibe las notificaciones enviadas (Fan et al., 2011). Se puede concluir que no genera beneficio para la empresa ni para el cliente (Shu-Hui & Hong-Nan, 2013).

- **Q3: ¿Qué algoritmos, herramientas y técnicas de limpieza de datos pueden ser usados en sistemas CRM?**

  Ya que el uso de algoritmos, herramientas y técnicas explicadas en Q1 son indiferentes al tipo de registro utilizado, se puede llegar a la conclusión que también es de utilidad para sistemas CRM con bases de datos inexactas, inconsistentes e inaccesibles.

3.2 Conclusiones

El presente proyecto de tesis usa una combinación de técnicas encontradas en el estado del arte para limpiar los datos. La técnica elegida es un proceso ETL que cargue a la base de datos transaccional, la información proveniente de varias fuentes y bases de datos y posteriormente, cargue a un *data warehouse*. El proceso ETL sigue el enfoque propuesto por Ali y Warraich (Ali & Warraich, 2010). Se genera un reporte de inconsistencias identificadas en la fuente de información y los cambios realizados. Adicionalmente, se proponen nuevos procesos de negocio mejorados para prevenir y mantener la correcta calidad de los datos de la institución educativa.

Capítulo 4. Bases de Datos Transaccionales

Este Capítulo está enfocado en mostrar los resultados esperados del objetivo específico 1: *Integrar y normalizar las entidades de las diversas fuentes y sistemas de información transaccionales en una única base de datos transaccional:*

- **R1-1:** *2 bases de datos transaccionales de distintos proveedores y un archivo en Excel.*
- **R1-2:** *Base de datos transaccional normalizada y unificada.*

En esta Sección se presentan los Diagramas Entidad Relación (DER) de las fuentes de datos transaccionales utilizadas por la institución educativa, y los diccionarios de datos. Asimismo, se presenta la nueva base de datos transaccional propuesta.

4.1 Bases de Datos Transaccionales Legacy

La empresa cuenta con dos bases de datos transaccionales donde se registra la información relacionada a los alumnos. Las bases de datos transaccionales son MySQL en la versión 5.7.17 y SQL Server 2016 Express Edition en la versión 13.1.4001.0.

La base de datos MySQL posee toda la información de las notas de los alumnos, es decir, el registro de las notas auxiliares por curso, el registro de las notas por competencia por curso y el registro del promedio de notas por curso para cada bimestre académico. Es necesario mencionar que este modelo de datos ha sido re-diseñado utilizando información recopilada de la institución educativa. Se re-diseñó debido a que ésta base de datos es administrada por una empresa tercerizada (*outsourcing)* y no se pudo obtener acceso a su información.

La base de datos SQL Server posee toda la información del pago de pensiones que debe realizar cada uno de los alumnos. En el caso que el alumno posea algún tipo de beca, se realiza el registro manual del monto que deberá pagar el alumno. La información proveniente de esta base de datos es entregada a los bancos: *BBVA Banco Continental*, *Banco de Crédito del Perú*, *Interbank* y *Scotiabank*; para que éstos realicen el cobro respectivo.

En las imágenes 7 y 8, se puede observar el Diagrama Entidad Relación (DER) de las bases de datos MySQL y SQL Server 2016, en los cuales se puede apreciar la relación que tiene la entidad “alumnos” con otras entidades. Los respectivos diccionarios de datos se pueden observar en el Anexo 4.

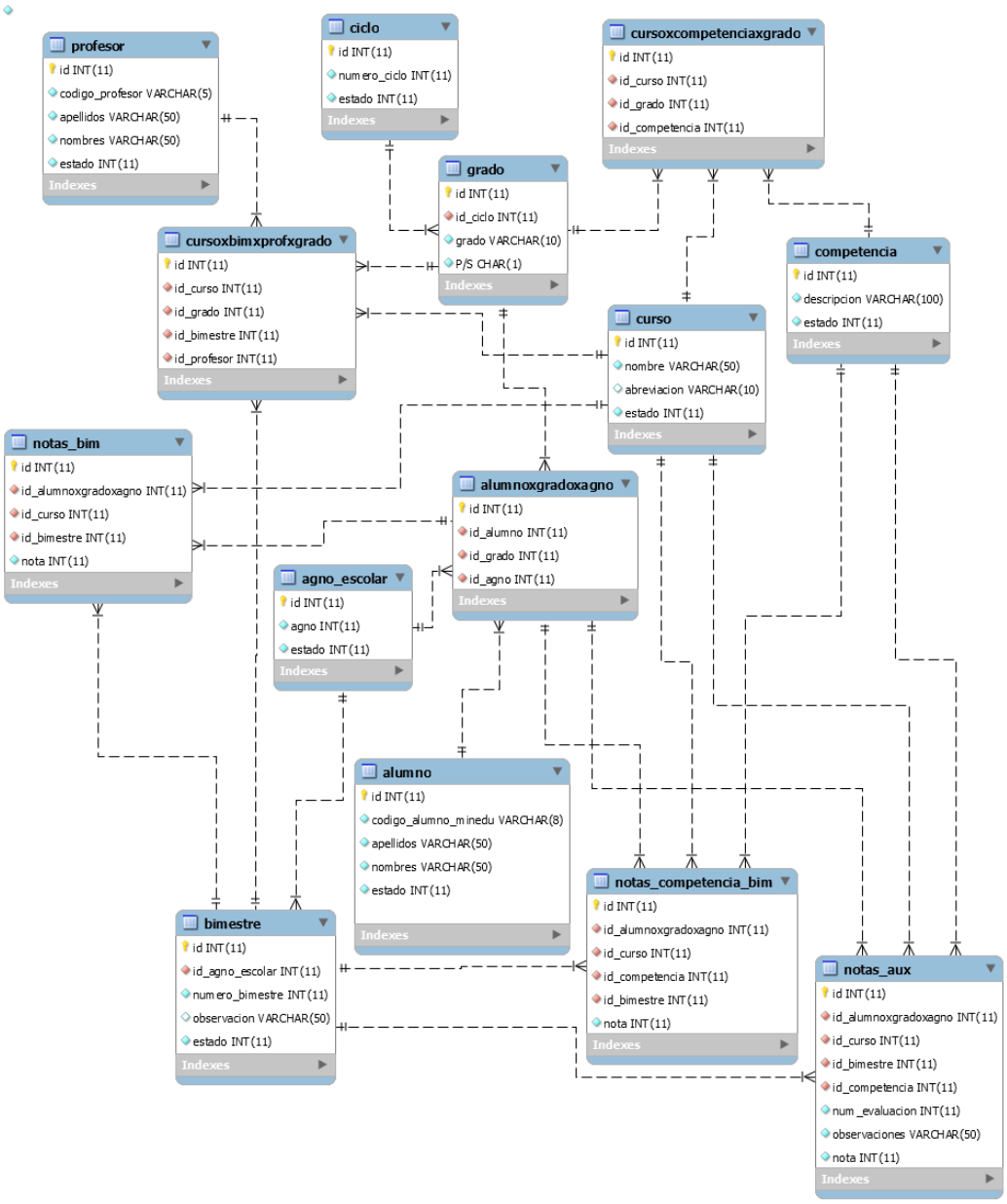

Imagen 7: DER de la Base de Datos MySQL. Imagen de autoría propia.

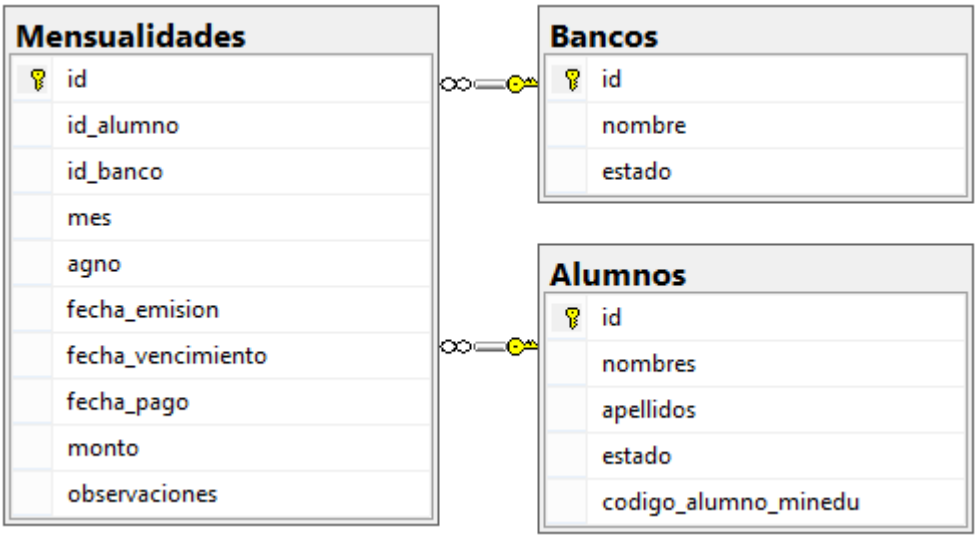

Imagen 8: DER de la Base de Datos SQL Server. Imagen de autoría propia.

Asimismo, la organización cuenta con un archivo en Excel donde guarda la información de los alumnos. Este archivo es la base de datos principal de alumnos utilizada por el colegio y es actualizado al inicio del año escolar por la secretaría de la Dirección.

En el Anexo 4, se puede observar los campos que tiene el archivo Excel.

4.2 Nueva Base de Datos Transaccional MySQL unificada

Esta nueva base de datos tiene unificadas todas las fuentes de datos que, en la actualidad, la institución educativa utiliza. Se propone un nuevo modelo de datos adecuado con el objetivo de tener unificada la información y prevenir la inserción de datos duplicados. Está normalizada hasta la tercera forma normal ya que cumple con las características presentadas en la tabla 4.

Tabla 4: Campos del archivo Excel. Tabla de autoría propia basada en información de la institución educativa.

1era Forma Normal:	• Grupos repetidos eliminados de las tablas individuales. • Tablas independientes para cada conjunto de datos relacionados. • Conjuntos de datos relacionados con una clave principal.
2da Forma Normal:	• Tablas independientes para conjuntos de valores que se apliquen a varios registros. • Tablas relacionadas con una clave externa.
3era Forma Normal:	• Campos eliminados que no dependan de la clave principal.

En la imagen 9, se puede observar el Diagrama Entidad Relación (DER) normalizado de la base de datos, en el cual se puede apreciar la nueva relación de entidades y los nuevos atributos para cada entidad. El respectivo diccionario de datos se puede observar en el Anexo 5.

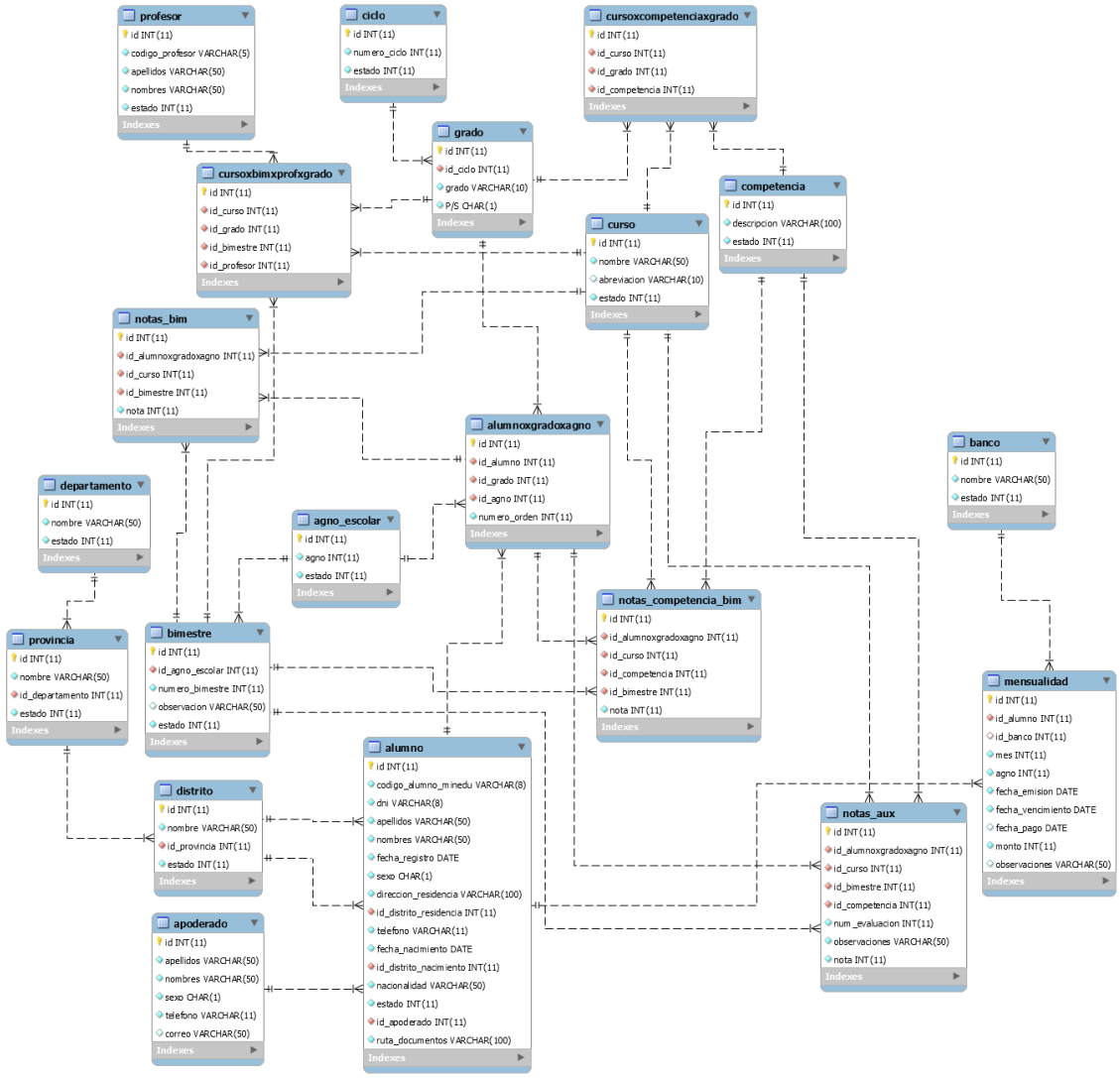

Imagen 9: Base de Datos transaccional MySQL normalizada. Imagen de autoría propia.

4.3 Discusión

En el Capítulo se presentaron los Diagramas Entidad Relación de las bases de datos y el Diagrama Entidad Relación de la nueva base de datos transaccional propuesta. Las realizaciones de estos resultados esperados validan el logro del objetivo específico 1: *Integrar y normalizar las entidades de las diversas fuentes y sistemas de información transaccionales en una única base de datos transaccional.*

Este objetivo específico apoya al objetivo general debido a que dos de las causas de la problemática: *Datos en muchas fuentes y sistemas de información* y *Sistemas transaccionales no normalizados* son solucionadas.

Capítulo 5. ETL de Limpieza de Fuentes de Datos y ETL de Integración en Base de Datos Transaccional

Este Capítulo está enfocado en mostrar los resultados esperados del objetivo específico 2: *Diseñar e implementar un algoritmo que permita corregir los datos incorrectos:*

- **R2-1:** *Algoritmo de limpieza de datos.*
- **R2-2:** *3 ETL para la lectura de datos de cada fuente de datos, limpieza y carga a las fuentes de datos transaccionales.*
- **R2-3:** *ETL para la lectura de datos de las fuentes de datos, integración y carga a la base de datos transaccional unificada.*

En esta Sección se definen los procesos ETL de limpieza e integración a utilizar para la realización del presente proyecto de fin de carrera. Se presenta la configuración ETL para cada una de las aplicaciones de limpieza que forman parte del proyecto y la configuración ETL para el proceso de integración de datos. El algoritmo de limpieza de datos implementado se presenta, también, a modo de diagrama y secuencia de pasos. El software Pentaho y el software Bizagi son las herramientas a utilizar en este capítulo.

5.1 Enfoque General de Solución

En la imagen 10, se muestra el enfoque general que se sigue. En primer lugar, se realiza la limpieza de cada fuente de datos. Luego, se realiza la integración de la información en la nueva base de datos transaccional. Finalmente, se realiza la integración a un *data warehouse*.

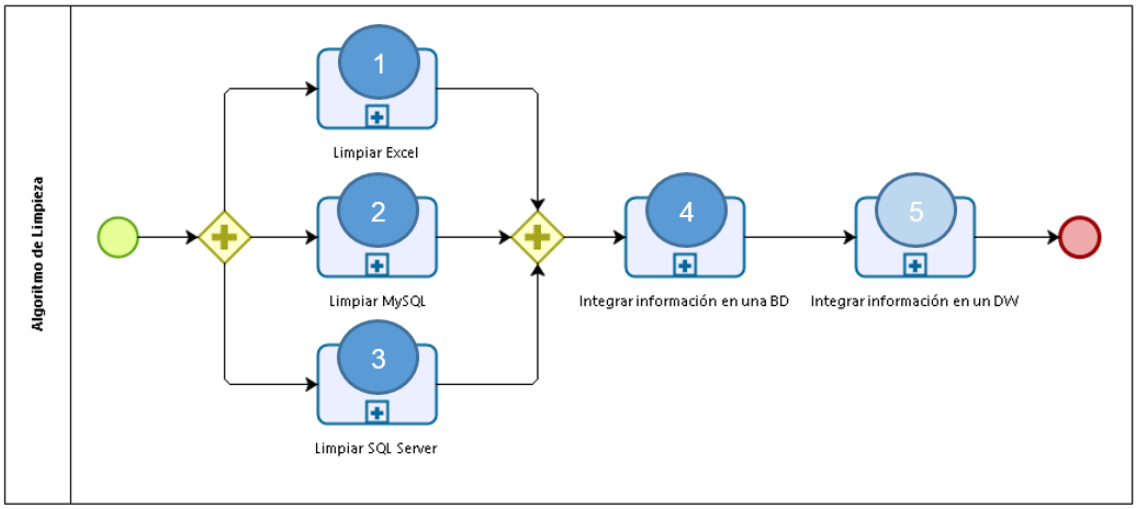

Imagen 10: Enfoque General de la Solución. Imagen de autoría propia.

Las aplicaciones descritas en este Capítulo son las cuatro primeras. Las aplicaciones de limpieza (números 1, 2 y 3) serán ejecutadas cada vez que la organización necesite limpiar sus datos transaccionales. Estas 3 aplicaciones recorrerán todos los registros en cada ejecución. La aplicación de integración (número 4) será ejecutada cuando la organización decida migrar la información a la nueva base de datos transaccional. Esta aplicación es incremental, es decir, solo recorre aquellos registros nuevos y los antiguos no los modifica.

5.2 Aplicación de Limpieza de Excel

En esta Sección se realiza la limpieza de la lista de alumnos. Esta aplicación no tiene pre-requisitos de ejecución y puede ser ejecutada cuando la organización la necesite.

5.2.1 Modelado del Algoritmo

En la imagen 11, se presenta la secuencia de pasos para la limpieza. En primer lugar, se separan los campos nombres y apellidos del alumno, nombres y apellidos del apoderado, lugar de nacimiento y dirección de residencia. Luego, se realiza una serie de validaciones. Finalmente, se vuelve a cargar toda la información al archivo Excel.

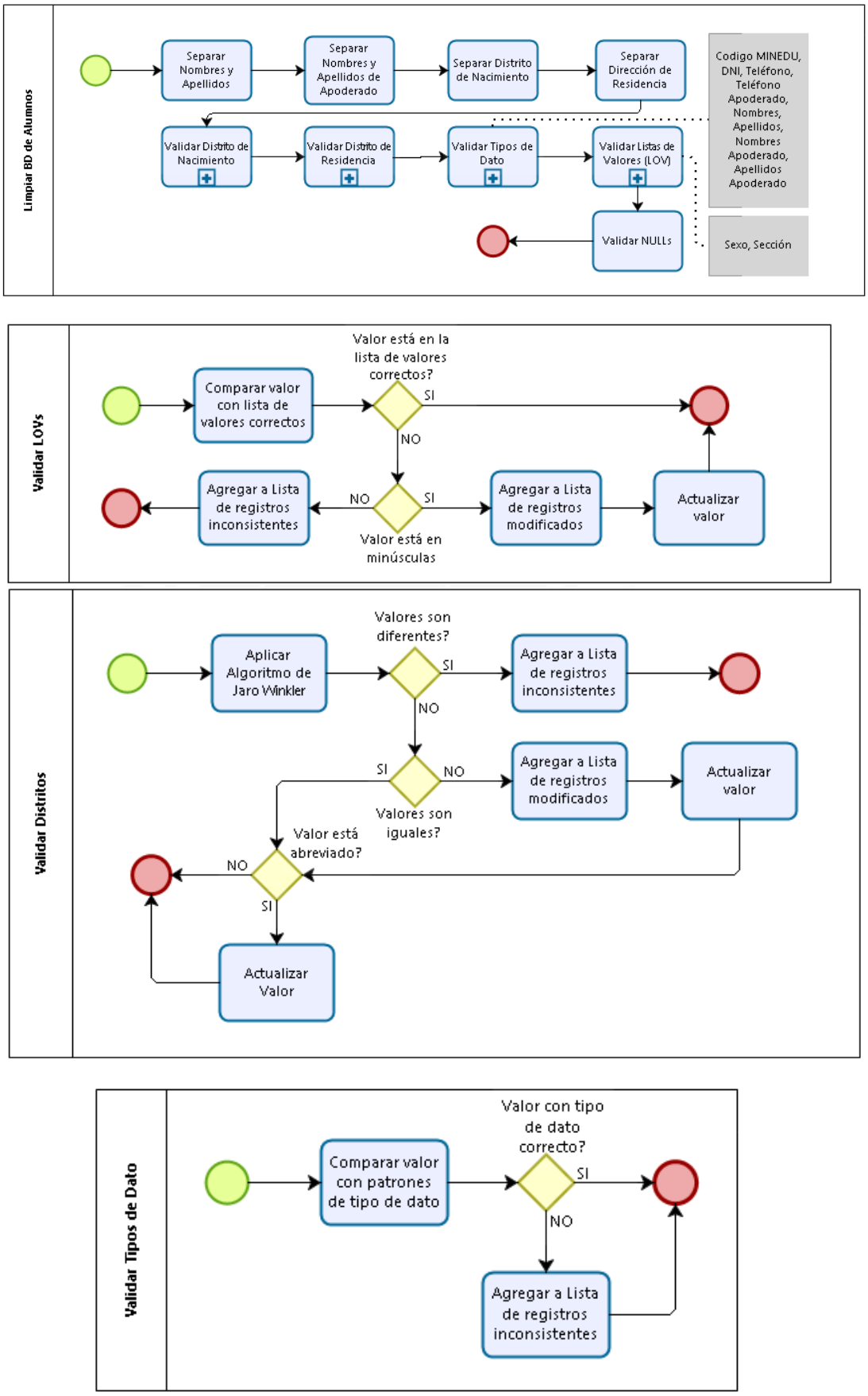

Imagen 11: Actividades para limpiar el archivo Excel. Imágenes de autoría propia.

5.2.2 Proceso ETL

Con el objetivo de tener los datos de los alumnos del archivo Excel limpios, se propone la configuración mostrada en la imagen 12. En primer lugar, se borran archivos temporales del directorio de trabajo. Luego, se realiza la limpieza del archivo Excel. Finalmente, se generan los reportes de inconsistencias y modificaciones.

Imagen 12: Configuración del proceso ETL para la limpieza del archivo Excel. Imagen de autoría propia.

En el Anexo 6, se describen con mayor detalle las transformaciones mostradas en la imagen 12.

5.2.3 Pruebas

Se realizaron 27 pruebas manuales al algoritmo y todas culminaron de acuerdo a lo establecido, demostrando que el algoritmo utilizado en la aplicación no tiene errores. Cada caso de prueba corresponde a un error específico en los registros. En el Anexo 7, se presenta el catálogo de pruebas y la ejecución de las pruebas.

5.3 Aplicación de Limpieza de BD MySQL

En esta Sección se realiza la limpieza de la base de datos de notas. Esta aplicación no tiene pre-requisitos de ejecución y puede ser ejecutada cuando la organización la necesite.

5.3.1 Modelado del Algoritmo

En la imagen 13, se presenta la secuencia de pasos para la limpieza. En primer lugar, se leen todos los registros de la entidad “alumnos” y de la entidad “profesores”. Luego, se realiza una serie de validaciones. Finalmente, se vuelve a cargar toda la información a la base de datos.

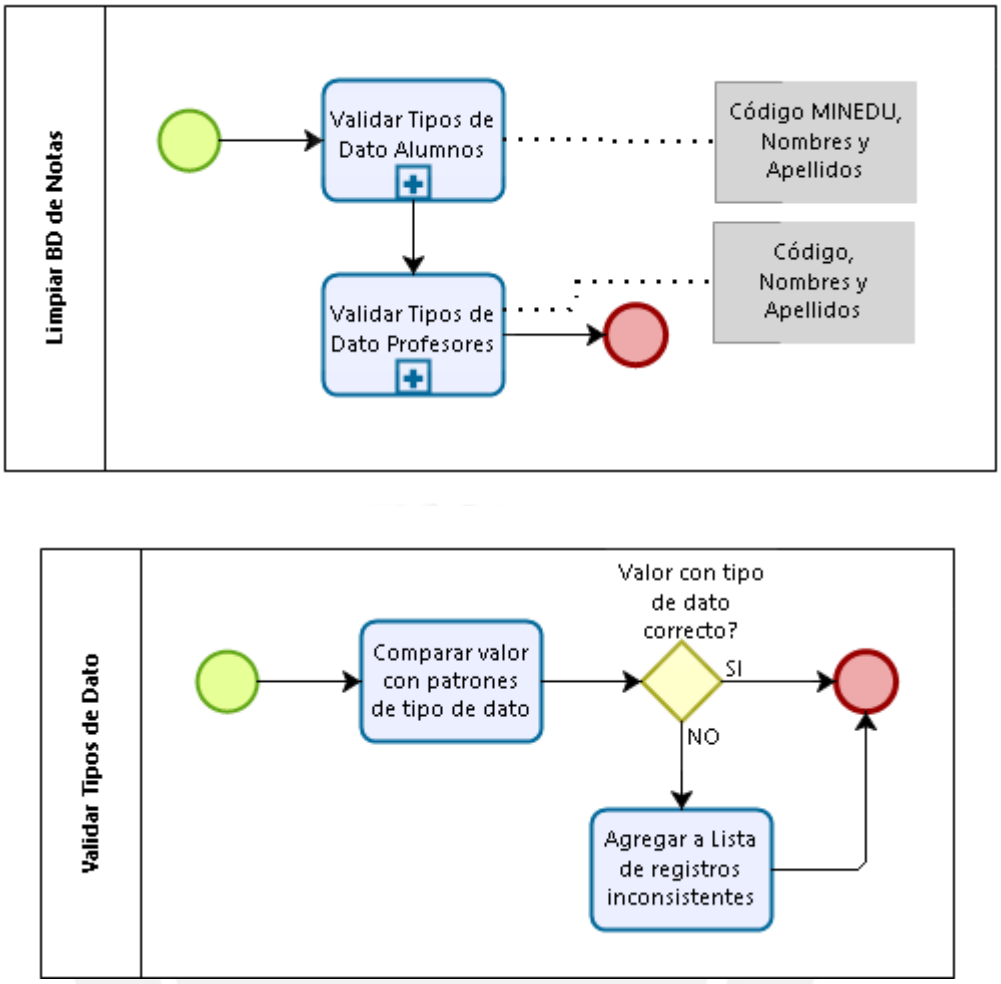

Imagen 13: Actividades para limpiar la Base de Datos de Notas. Imágenes de autoría propia

5.3.2 Proceso ETL

Con el objetivo de tener los datos de los alumnos de la base de datos MySQL limpios, se propone la configuración mostrada en la imagen 14. En primer lugar, se borran archivos temporales del directorio de trabajo. Luego, se realiza la limpieza de la base de datos de notas (limpieza de entidad de alumnos y profesores). Finalmente, se generan los reportes de inconsistencias para alumnos y profesores.

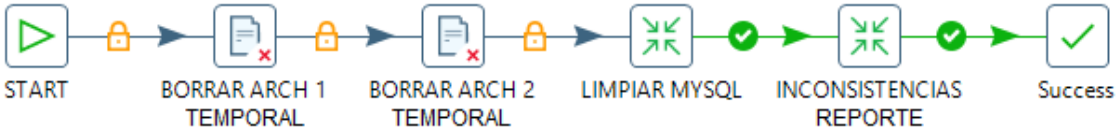

Imagen 14: Configuración del proceso ETL para la limpieza de la base de datos MySQL. Imagen de autoría propia.

En el Anexo 8, se describen con mayor de talle las transformaciones mostradas en la imagen 14.

5.3.3 Pruebas

Se realizaron 10 pruebas manuales al algoritmo y todas culminaron de acuerdo a lo establecido, demostrando que el algoritmo utilizado en la aplicación no tiene errores. En el Anexo 9, se presenta el catálogo de pruebas y la ejecución de las pruebas.

5.4 Aplicación de Limpieza de BD SQL Server

En esta Sección se realiza la limpieza de la base de datos del pago de pensiones. Esta aplicación no tiene pre-requisitos de ejecución y puede ser ejecutada cuando la organización la necesite.

5.4.1 Modelado del Algoritmo

En la imagen 15, se presenta la secuencia de pasos para la limpieza. En primer lugar, se leen todos los registros de la entidad “alumnos” y de la entidad “bancos”. Luego, se realiza una serie de validaciones. Finalmente, se vuelve a cargar toda la información a la base de datos.

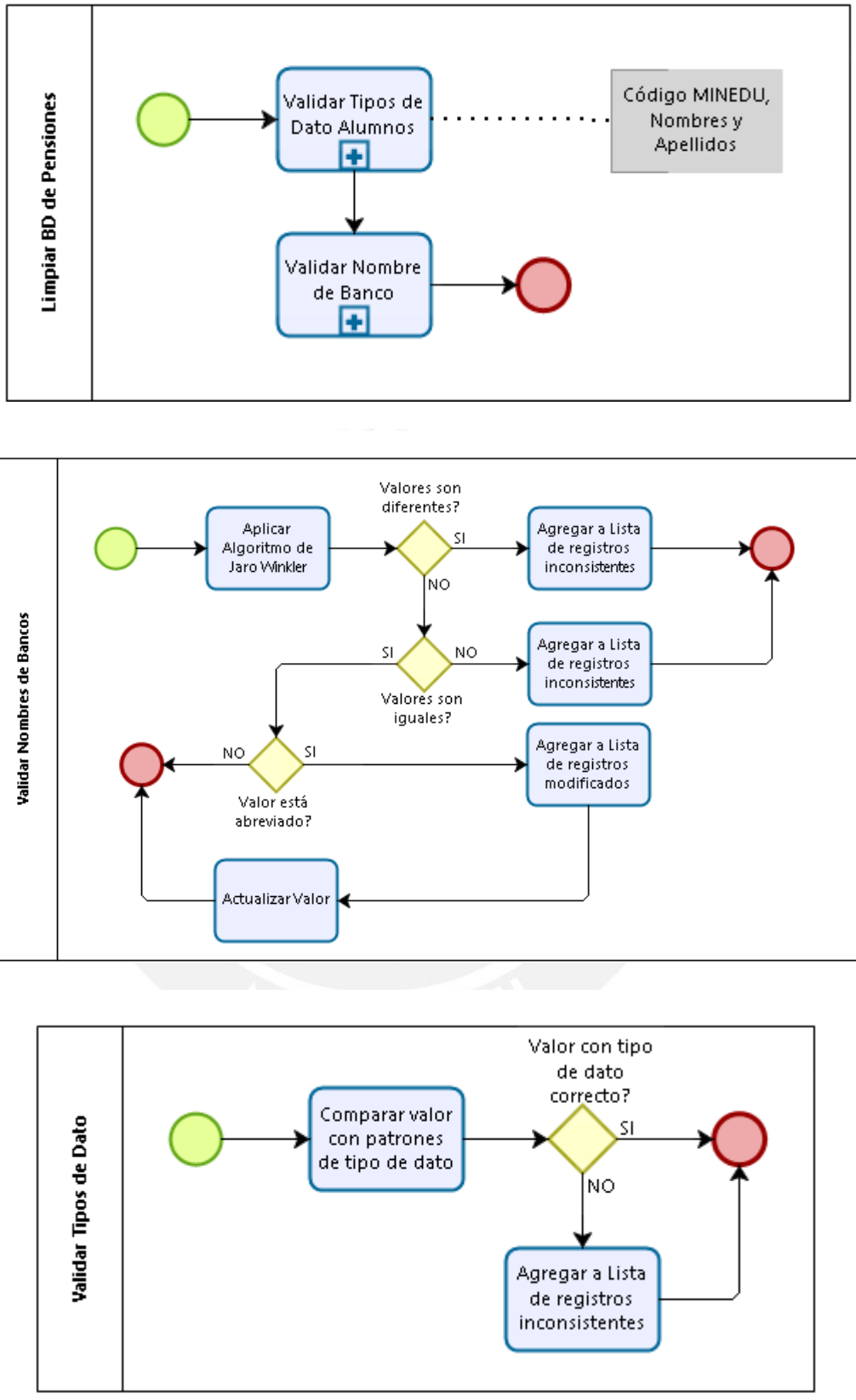

Imagen 15: Actividades para limpiar la Base de Datos de Pagos. Imágenes de autoría propia.

5.4.2 Proceso ETL

Con el objetivo de tener los datos de los alumnos de la base de datos SQL Server limpios, se propone la configuración mostrada en la imagen 16. En primer lugar, se borran archivos temporales del directorio de trabajo. Luego, se realiza la limpieza de la base de datos de pagos (limpieza de entidad de alumnos y bancos). Finalmente, se generan los reportes de inconsistencias y modificaciones.

Imagen 16: Configuración del proceso ETL para la limpieza de la base de datos SQL Server. Imagen de autoría propia.

En el Anexo 10, se describen con mayor detalle las transformaciones mostradas en la imagen 16.

5.4.3 Pruebas

Se realizaron 8 pruebas manuales al algoritmo y todas culminaron de acuerdo a lo establecido, demostrando que el algoritmo utilizado en la aplicación no tiene errores. En el Anexo 11, se presenta el catálogo de pruebas y la ejecución de las pruebas.

5.5 Aplicación de Integración en BD Transaccional

En esta Sección se realiza la integración de todas las fuentes de datos en una única base de datos. Esta aplicación tiene como pre-requisito haber realizado la limpieza de cada fuente de datos con anterioridad y el levantamiento de todos los errores encontrados en los reportes de inconsistencias generados.

5.5.1 Modelado del Algoritmo

Se siguen las reglas de negocio definidas en la Tabla 5.

Tabla 5: Reglas de Negocio. Tabla de autoría propia basada en información de la institución educativa.

Códigos MINEDU	Nombres y Apellidos	¿UNIFICAR?
IGUALES	IGUALES	SI
IGUALES	PARECIDOS	SI
IGUALES	DIFERENTES	NO
PARECIDOS	IGUALES	SI
PARECIDOS	PARECIDOS	NO
PARECIDOS	DIFERENTES	NO
DIFERENTES	IGUALES	NO
DIFERENTES	PARECIDOS	NO
DIFERENTES	DIFERENTES	NO

En la imagen 17, se presenta el algoritmo de integración. Se leen todos los registros de la entidad alumnos y cada uno de estos registros es comparado con los registros del archivo Excel, con el objetivo de encontrar similitudes entre las fuentes de datos. En el caso que el algoritmo encuentre dos registros que tengan similitudes, el algoritmo procede a integrarlos.

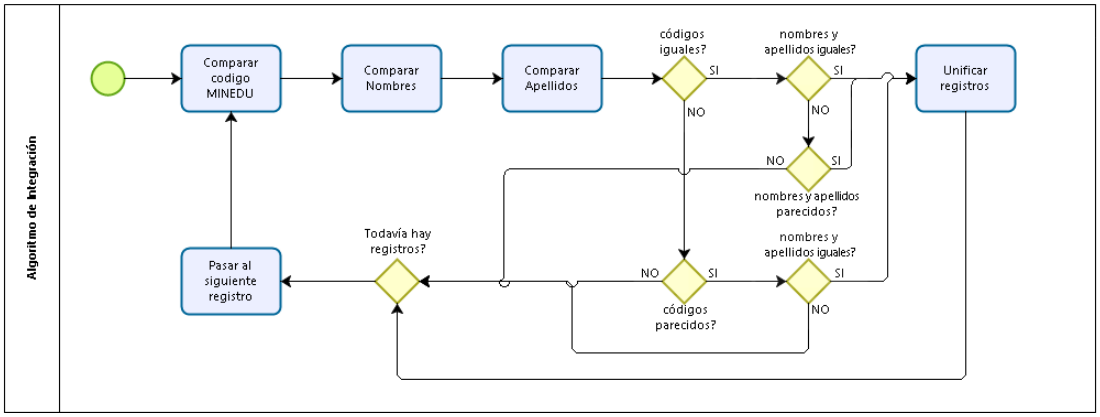

Imagen 17: Algoritmo de Integración. Imagen de autoría propia.

En las imágenes 18 y 19, se presenta la secuencia de pasos para la integración. En primer lugar, se intenta integrar los registros de la base de datos MySQL con los registros del archivo Excel usando el algoritmo de la imagen 17. Luego, se intenta integrar los registros de la base de datos SQL Server con los registros del archivo Excel usando, también, el algoritmo de la imagen 17. Posteriormente, se integran los registros de las 3 fuentes de datos. Finalmente, se realiza la carga a la nueva base de datos transaccional.

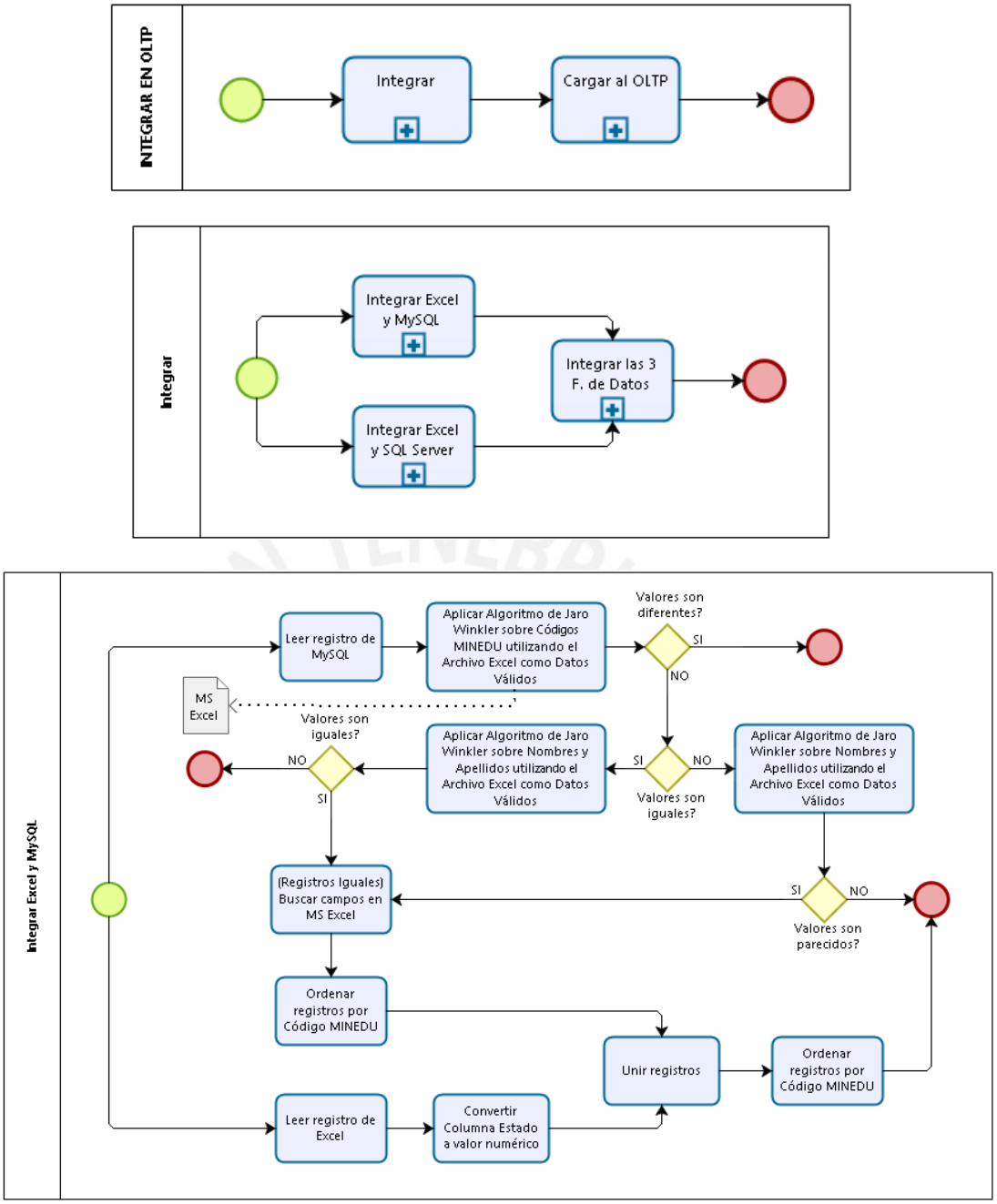

Imagen 18: Actividades para hacer la integración en una única base de datos. Imágenes de autoría propia.

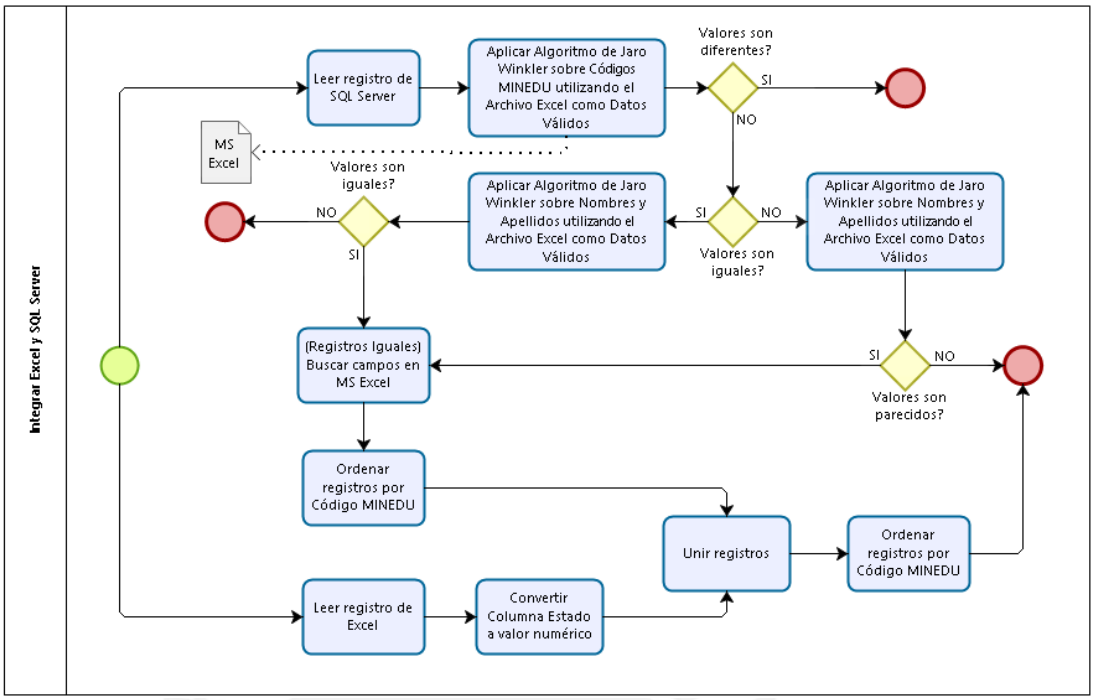

Imagen 19: Actividades para hacer la integración en una única base de datos. Imágenes de autoría propia.

5.5.2 Proceso ETL

Con el objetivo de tener los datos de los alumnos de las fuentes de datos unificados en una sola base de datos transaccional, se propone la configuración mostrada en la imagen 20. En primer lugar, se realiza la integración de las 3 fuentes de datos Finalmente, se realiza la carga de todas las entidades de las bases de datos MySQL y SQL Server a la nueva base de datos transaccional.

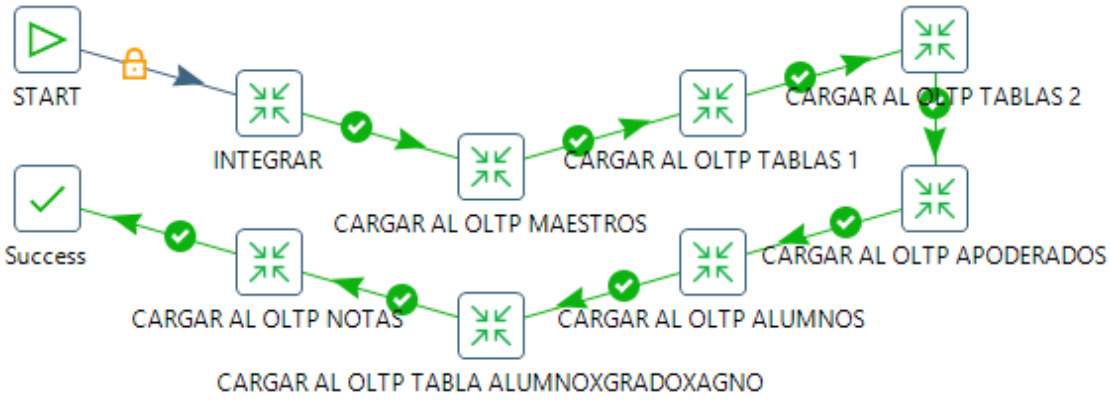

Imagen 20: Configuración del proceso ETL para la integración en una base de datos única. Imagen de autoría propia.

En el Anexo 12, se describen con mayor detalle las transformaciones mostradas en la imagen 20.

5.6 Discusión

En el Capítulo se presentó el enfoque general de la limpieza de datos, que incluye los ETL para la limpieza de las 3 fuentes de datos (parte de este capítulo), el ETL para la integración de los datos en una única base de datos transaccional (parte de este capítulo) y el ETL para la integración de los datos en un *data warehouse* (parte del capítulo 6). Los resultados esperados fueron validados utilizando casos de prueba. Las realizaciones de estos resultados esperados validan el logro del objetivo específico 2: *Diseñar e implementar un algoritmo que permita corregir los datos incorrectos.* Este objetivo específico apoya al objetivo general debido a que una de las causas de la problemática: *Datos incorrectos* es solucionada.

Capítulo 6. ETL de Integración en *Data Warehouse*

Este capítulo está enfocado en mostrar los resultados esperados del objetivo específico 3: *Implementar un módulo de software que permita integrar la información de la nueva base de datos transaccional en un data warehouse:*

- **R3-1:** *ETL para la lectura de datos de la base de datos transaccional, integración y carga al data warehouse.*
- **R3-2:** *Creación de data warehouse* y *data marts.*
- **R3-3:** *Información de las fuentes de datos transaccionales integrada en un data warehouse.*

En esta Sección se define el último proceso ETL a utilizar para la realización del presente proyecto de fin de carrera. Se presenta la configuración ETL para el proceso de integración de datos de la base de datos transaccional a un *data warehouse*. Asimismo, se muestra el d*ata warehouse* con un esquema en estrella. El software Pentaho y MySQL Workbench serán las herramientas a utilizar en este capítulo.

6.1 Enfoque General de Solución

En la imagen 21, se muestra, nuevamente, el enfoque general que se sigue.

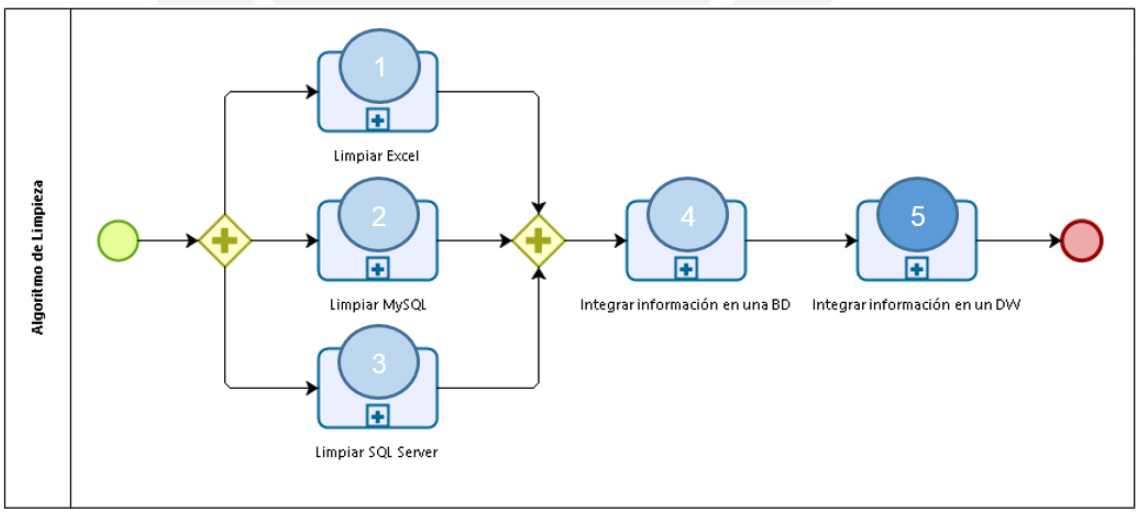

Imagen 21: Enfoque General de la Solución. Imagen de autoría propia.

La aplicación descrita en este Capítulo es la última del proceso. La aplicación de integración al *data warehouse* (número 5) será ejecutada cuando la organización

decida empezar a utilizar el *data warehouse* implementado. Esta aplicación tiene como pre-requisito haber integrado los datos en la nueva base de datos transaccional con anterioridad.

Se genera 1 *data warehouse* con la información integrada. El Diagrama Entidad Relación del *data warehouse* se presenta en el Anexo 13. Se generan 3 *data marts* con información relevante para la institución educativa. Los Diagramas Entidad Relación de los *data marts* son presentados a continuación.

6.1.1 *Data Mart*: Profesores x Notas x Tiempo

En la imagen 22, se presenta el Diagrama Entidad Relación (DER) del *data mart*. Se puede apreciar la tabla de hechos y las tablas de dimensiones que pueblan la tabla de hechos.

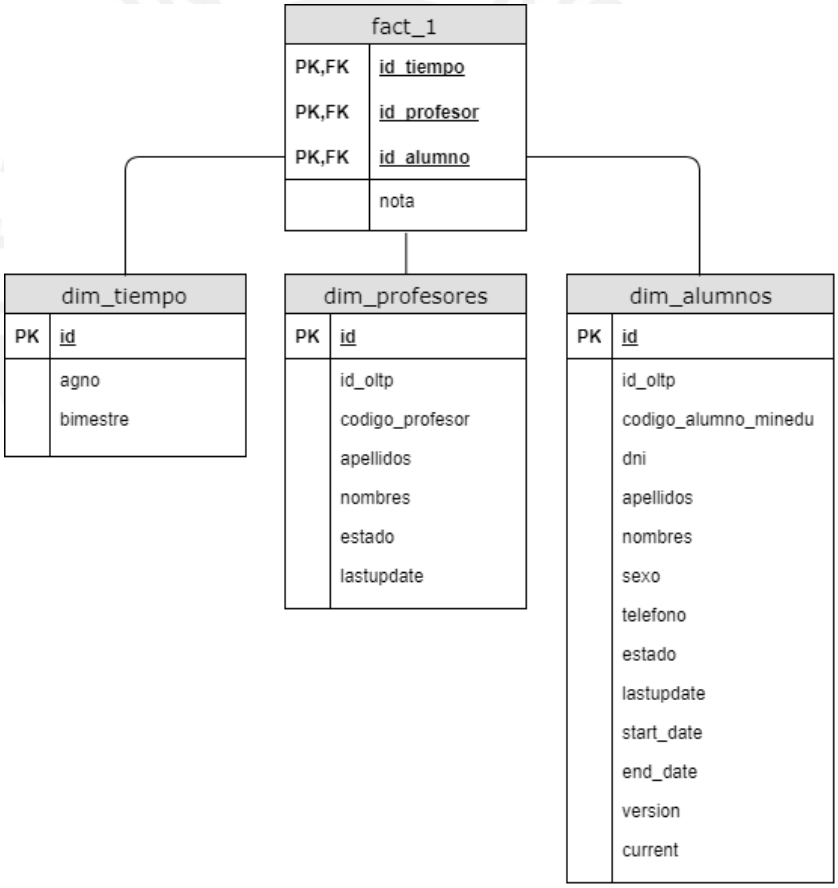

Imagen 22: DER del *Data Mart*: Profesores x Notas x Tiempo. Imagen de autoría propia.

6.1.2 *Data Mart*: Cursos x Notas x Tiempo

En la imagen 23, se presenta el Diagrama Entidad Relación (DER) del *data mart*. Se puede apreciar la tabla de hechos y las tablas de dimensiones que pueblan la tabla de hechos.

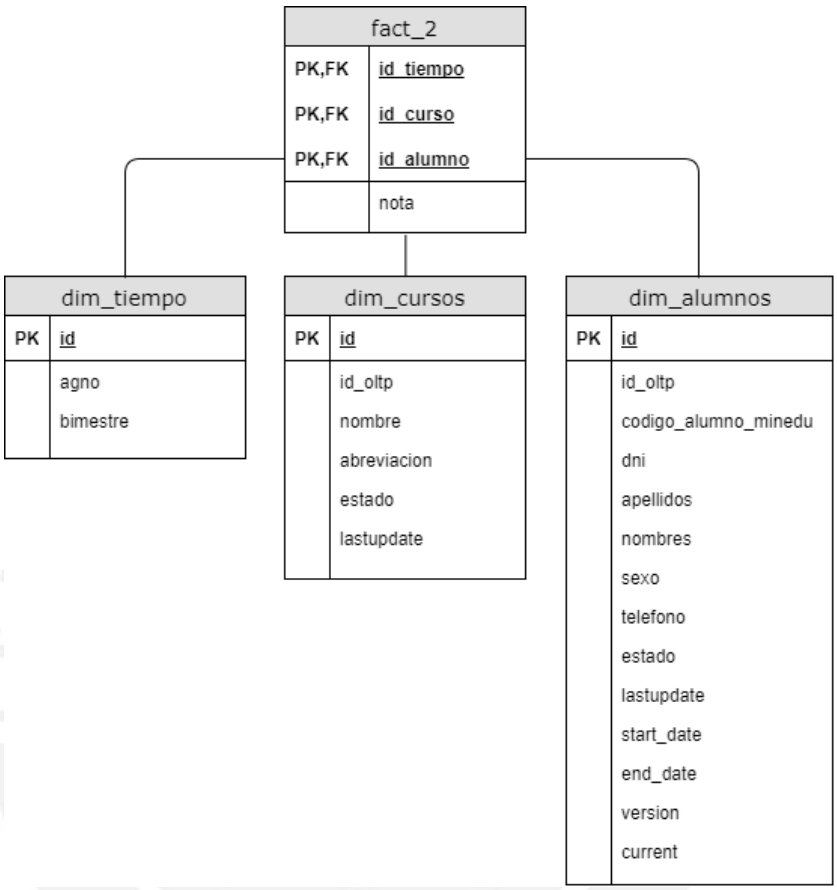

Imagen 23: DER del *Data Mart*: Cursos x Notas x Tiempo. Imagen de autoría propia.

6.1.3 *Data Mart*: Distritos x Notas x Tiempo

En la imagen 24, se presenta el Diagrama Entidad Relación (DER) del *data mart*. Se puede apreciar la tabla de hechos y las tablas de dimensiones que pueblan la tabla de hechos.

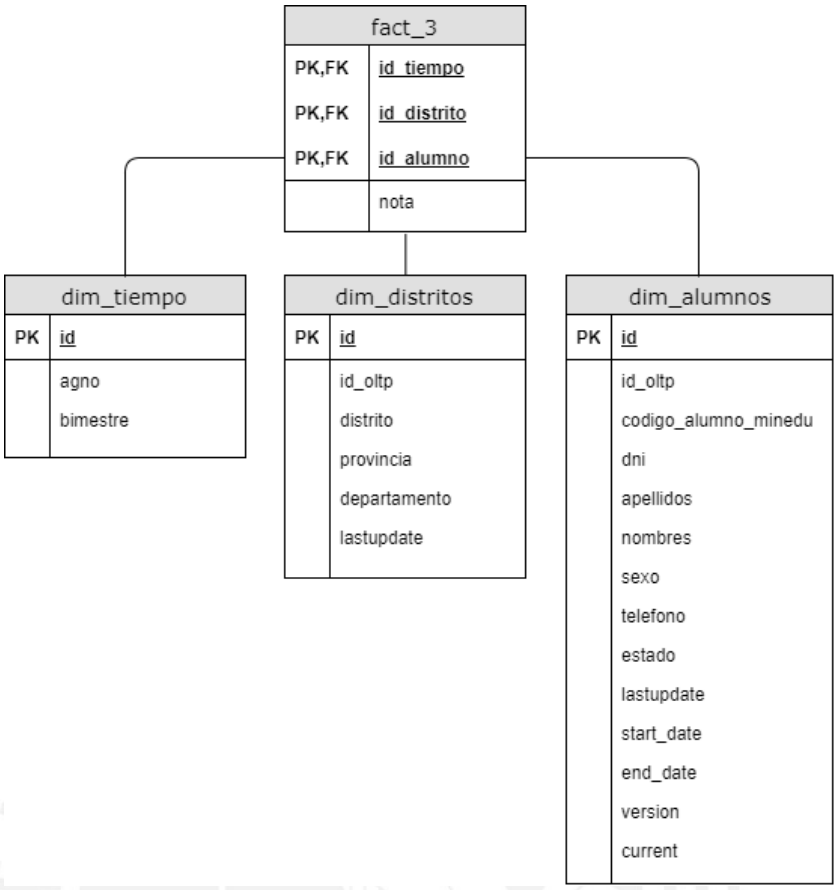

Imagen 24: DER del *Data Mart*: Profesores x Notas x Tiempo. Imagen de autoría propia.

6.2 Proceso ETL

Con el objetivo de tener los datos de los alumnos de las fuentes de datos unificados en un *data warehouse*, para su posterior explotación, se propone la configuración mostrada en las imágenes 25, 26 y 27. En primer lugar, se cargan todas las tablas de dimensiones y, posteriormente, se cargan todas las tablas de hechos, usando los datos provenientes de la nueva base de datos transaccional.

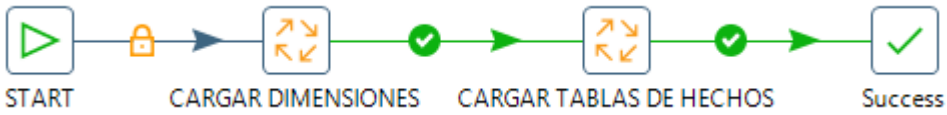

Imagen 25: Configuración del proceso ETL para la integración en un *data warehouse*. Imagen de autoría propia.

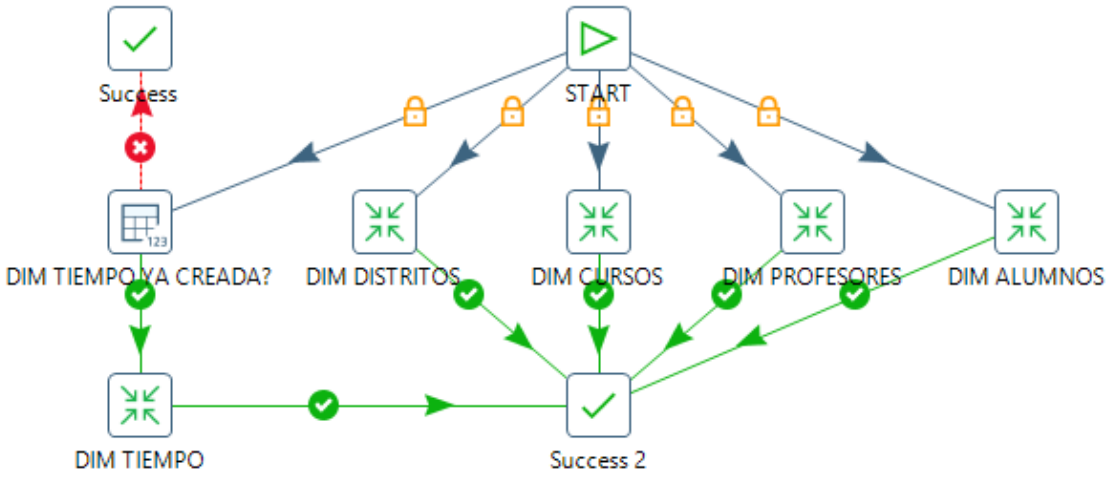

Imagen 26: Configuración del proceso ETL para las cargas de las dimensiones. Imagen de autoría propia.

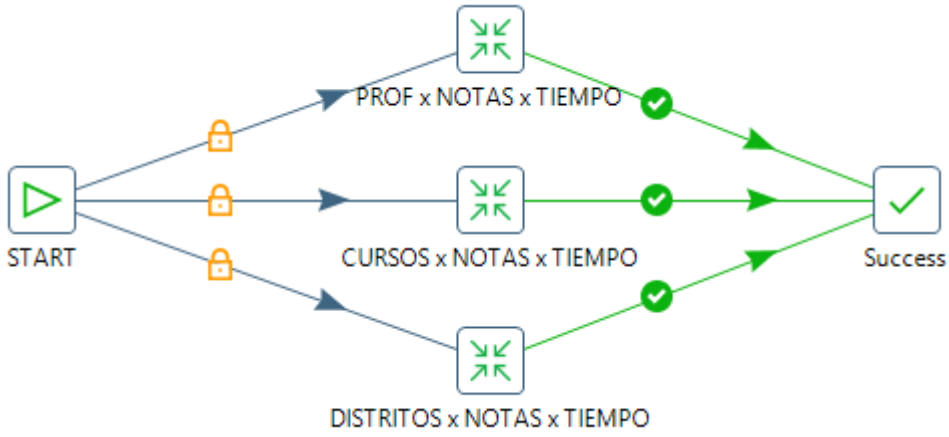

Imagen 27: Configuración del proceso ETL para las cargas de las tablas de hechos. Imagen de autoría propia.

En el Anexo 14, se describen con mayor detalle las transformaciones mostradas en las imágenes 26 y 27.

6.3 Discusión

En el Capítulo se presentó, nuevamente, el enfoque general de la limpieza de datos, que incluye el ETL para la integración de los datos en un *data warehouse* (parte de este capítulo). Asimismo, se presentó el *data warehouse* y *los data marts* con sus tablas dimensionales y de hechos. Las realizaciones de estos resultados esperados validan el logro del objetivo específico 3: *Implementar un módulo de software que permita integrar la información de la nueva base de datos transaccional en un data warehouse.* Este objetivo específico apoya al objetivo

general debido a que una de las causas de la problemática: *Datos en muchas fuentes y sistemas de información* es solucionada.

Capítulo 7. Procesos de Negocio

Este capítulo está enfocado en mostrar los resultados esperados del objetivo específico 4: *Implementar y mejorar los procesos de negocio relacionados a los estudiantes para mantener la calidad de los datos en el sistema transaccional:*

- **R4-1:** *Modelado de procesos.*
- **R4-2:** *Mejora de procesos para prevenir la aparición de datos con mala calidad.*

7.1 Procesos de la empresa

La empresa tiene un conjunto de procesos de negocio relacionados a la inserción y modificación de datos de alumnos. Los procesos no se encontraron documentados, así que se propone una documentación formal usando notación BPMN. Los más importantes son los mostrados a continuación.

7.1.1 Proceso: Admisión de nuevos alumnos

En la imagen 28, se puede observar el proceso de negocio de admisión utilizado por la institución educativa en la actualidad. El proceso se realiza de forma manual y no se tiene automatizada ninguna actividad. El proceso se encuentra validado por el cliente.

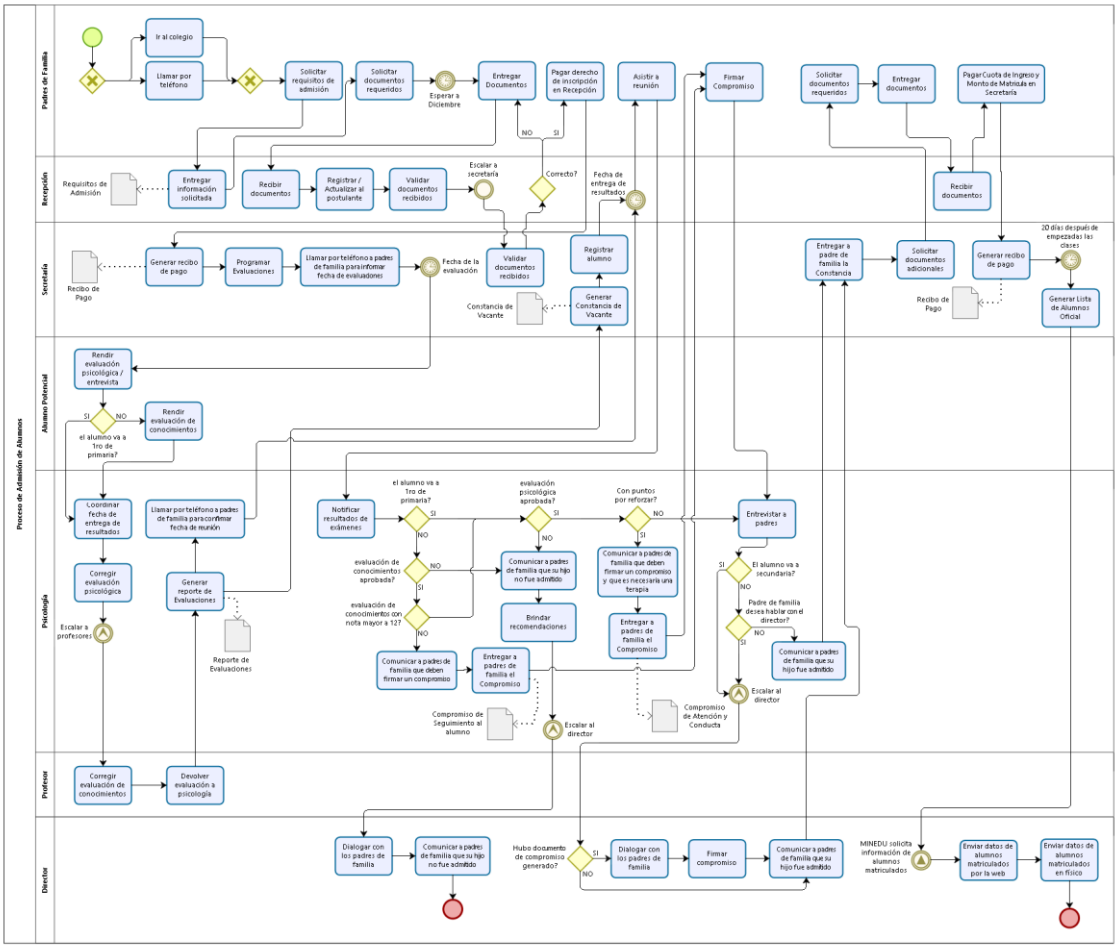

Imagen 28: Proceso de Negocio de Admisión. Imagen de autoría propia.

7.1.2 Proceso: Subir Notas

En la imagen 29, se puede observar el proceso de negocio de subir notas utilizado por la institución educativa en la actualidad. El proceso se realiza en conjunto con una empresa tercerizada. El proceso se encuentra validado por el cliente.

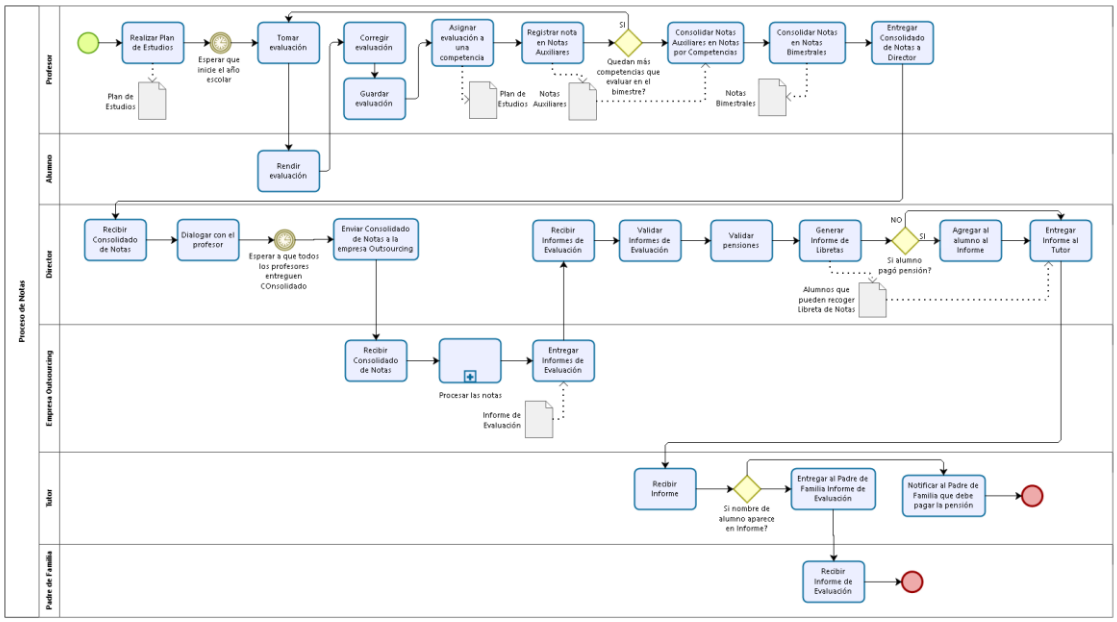

Imagen 29: Proceso de Negocio de Subir Notas. Imagen de autoría propia.

7.2 Nuevos Procesos de Negocio

En el presente proyecto de fin de carrera se proponen nuevos procesos con el objetivo de prevenir la inserción de datos inconsistentes. Estos procesos fueron desarrollados tomando información de empresas líderes en el sector y utilizando buenas prácticas de procesos de negocio.

7.2.1 Nuevo proceso de Admisión de nuevos alumnos

En la imagen 30, se puede observar el proceso de negocio mejorado y automatizado para la institución educativa. Se propone la utilización de la nueva base de datos transaccional integrada.

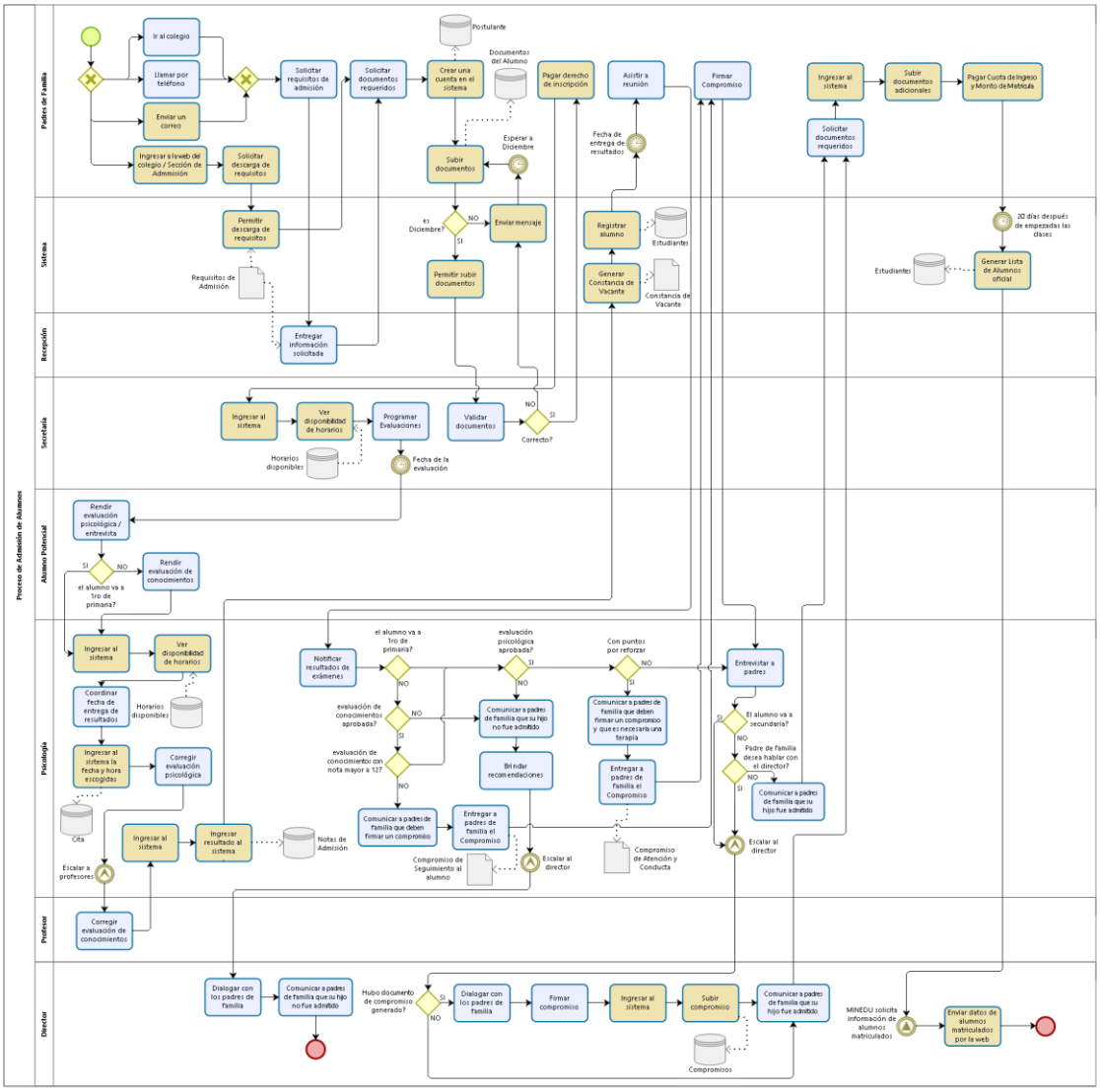

Imagen 30: Proceso de Negocio Mejorado y Automatizado. Imagen de autoría propia.

(AMARILLO: Parte del proceso que se automatiza**)**

7.2.2 Nuevo proceso de Subir Notas

En la imagen 31, se puede observar el proceso de negocio mejorado y automatizado para la institución educativa. Se propone la utilización de la nueva base de datos transaccional integrada.

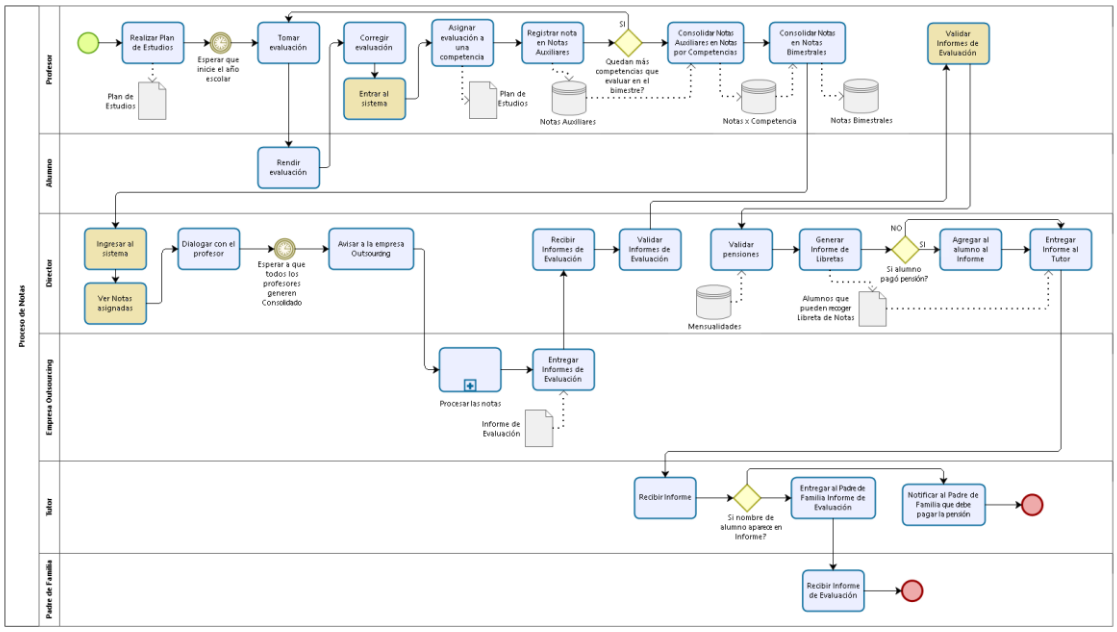

Imagen 31: Proceso de Negocio Mejorado y Automatizado. Imagen de autoría propia.

(AMARILLO: Parte del proceso que se automatiza**)**

7.3 Discusión

En el Capítulo se presentaron los procesos de negocio relacionados a los estudiantes que utiliza la institución educativa y los mismos procesos de negocio, pero mejorados. Los resultados esperados fueron validados mediante el juicio experto del cliente. Las realizaciones de estos resultados esperados validan el logro del objetivo específico 4: *Implementar y mejorar los procesos de negocio relacionados a los estudiantes para mantener la calidad de los datos en el sistema transaccional.* Este objetivo específico apoya al objetivo general debido a que, según el estado del arte, un enfoque de proceso también es recomendable para la limpieza de datos de sistemas.

Capítulo 8. Conclusiones y trabajos futuros

En esta Sección se presentan las conclusiones obtenidas luego de haber logrado los objetivos específicos planteados. Posteriormente, se presentan las posibles mejoras y trabajos futuros identificados.

8.1 Conclusiones

Según lo realizado en el presente proyecto de fin de carrera, se concluye lo siguiente:

- Los datos son el recurso más importante de toda organización y tener este recurso con una buena calidad es una labor que debería ser prioritaria de toda organización.
- El estado del arte evidencia que existen varias técnicas y métodos para la limpieza de datos; sin embargo, el número de estudios primarios encontrados, utilizando el método de Revisión Sistemática, refleja que no es un campo de estudio muy estudiado.
- Los sistemas CRM son importantes para toda organización enfocada en los clientes. Para el caso de una institución educativa, los clientes serían los estudiantes y mantener una buena relación con estos debería ser una labor prioritaria en toda institución educativa.
- El objetivo específico 1: *Integrar y normalizar las entidades de las diversas fuentes y sistemas de información transaccionales en una única base de datos transaccional* pudo ser desarrollado y se propuso un nuevo modelo de base de datos transaccional que integre toda la información necesaria.
- El objetivo específico 2: *Diseñar e implementar un algoritmo que permita corregir los datos incorrectos* pudo ser desarrollado y se propuso un algoritmo y procesos ETL para limpiar los registros e integrar los datos en una única base de datos transaccional.
- El objetivo específico 3: *Implementar un módulo de software que permita integrar la información de la nueva base de datos transaccional en un data warehouse* pudo ser desarrollado, se propuso un nuevo modelo de data warehouse y nuevos modelos de data marts.

- El objetivo específico 4: *Implementar y mejorar los procesos de negocio relacionados a los estudiantes para mantener la calidad de los datos en el sistema transaccional* pudo ser desarrollado, se modelaron los procesos actuales de la institución educativa y se propuso nuevos procesos.

8.2 Trabajos futuros

Según lo realizado en el presente proyecto de fin de carrera, se identificaron los siguientes posibles trabajos futuros:

- La institución educativa no posee una herramienta con la cual administrar y gestionar la información relacionada a los alumnos. La implementación de un sistema de información, que utilice la base de datos transaccional propuesta, sería una posible mejora.
- Las instituciones educativas, generalmente, no poseen sistemas CRM en sus organizaciones. Un análisis de viabilidad y retorno de inversión de implementar sistemas CRM en instituciones educativas sería un posible trabajo futuro.
- Existen varias técnicas y métodos para la limpieza de datos. La implementación de un proceso de limpieza utilizando otros métodos y otro enfoque de los mostrados en el estado del arte sería un posible trabajo futuro.
- La institución educativa no posee una herramienta de inteligencia de negocios (*Business Intelligence* - BI) para la toma de decisiones. La implementación de herramientas de análisis de datos como *dashboards* y *scorecards,* que utilicen el *data warehouse* generado, sería una posible mejora.
- La institución educativa no cuenta con una arquitectura empresarial apropiada para la realización de los proyectos de implementación propuestos previamente. Una consultoría basada en arquitectura empresarial ayudaría a la institución educativa a identificar los requerimientos, a nivel organizacional, que debe cumplir para lograr cumplir sus objetivos estratégicos. La consultoría sería un posible trabajo futuro.
- Los procesos de negocio de la institución educativa no se encuentran documentados. La creación de un Manual de Procesos y Procedimientos con la totalidad de procesos utilizados sería un posible trabajo futuro.

Referencias

AACRAO. (2016). 2014-2015 State of CRM Use in Higher Education Report, 42.

Al-Mudimigh, A. S., Saleem, F., Ullah, Z., & Al-Aboud, F. N. (2009). Implementation of Data Mining Engine on CRM - Improve customer satisfaction. *2009 International Conference on Information and Communication Technologies, ICICT 2009*. https://doi.org/10.1109/ICICT.2009.5267193

Al-Mudimigh, A. S., Ullah, Z., & Saleem, F. (2009). Data mining strategies and techniques for CRM systems. *System of Systems Engineering, 2009. SoSE 2009. IEEE International Conference on*. Retrieved from http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5282344

Ali, K., & Warraich, M. A. (2010). A framework to implement Data Cleaning in Enterprise Data Warehouse for Robust Data Quality, *8491*, 1–6.

Arora, R., Pahwa, P., & Bansal, S. (2009). Alliance rules for data warehouse cleansing. *2009 International Conference on Signal Processing Systems, ICSPS 2009*, 743–747. https://doi.org/10.1109/ICSPS.2009.133

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), 1–52. https://doi.org/10.1145/1541880.1541883

Biolchini, J., Mian, P. G., Natali, A. C. C., & Travassos, G. H. (2005). Systematic review in software engineering. *Technical Report ES-679/05*, (May), 31. https://doi.org/10.1145/2372233.2372235

BizAgi. (2009). *Bizagi Process Modeler User's Guide*. Retrieved from http://help.bizagi.com/processmodeler/es/

Bleiholder, J., & Naumann, F. (2008). Data fusion. *ACM Computing Surveys*, *41*(1), 1–41. https://doi.org/10.1145/1456650.1456651

Bouman, R. (2006). Pentaho Data Integration: Kettle turns data into business.

Eckerson, W. (2011). Data Quality and the BottomLine: Achieving Business Success through a Commitment to High Quality Data. *Tech Target*, (June), 1–39. Retrieved from www.tdwi.org/research

Faed, A., Wu, C., & Chang, E. (2010). Intelligent CRM on the cloud. *Proceedings - 13th International Conference on Network-Based Information Systems, NBiS*

2010. https://doi.org/10.1109/NBiS.2010.12

Fan, W., Geerts, F., & Wijsen, J. (2011). Determining the currency of data. *Proc. 30th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, *37*(4), 71–82. https://doi.org/10.1145/2389241.2389244

KaptureCRM. (2017). ¿Por qué su colegio / universidad necesita un software de CRM Educación? Retrieved from https://www.kapturecrm.com/blog/college-university-needs-education-crm-software/

Khan, B., Rauf, A., Shah, S. H., & Khusro, S. (2011). Identification and removal of duplicated records. *World Applied Sciences Journal*, *13*(5), 1178–1184.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3. *Engineering*, *45*(4ve), 1051. https://doi.org/10.1145/1134285.1134500

Kumar, V., & Reinartz, W. (2012). *Customer Relationship Management* (2nd Editio). Springer.

Levine, S. (2000). The rise of CRM. *America's Network*, *104*(6), 34.

Linkedin.com. (2015). MySQL: Overview. Retrieved from https://www.linkedin.com/company/mysql

Lucas, J., Raja, U., & Ishfaq, R. (2014). How Clean is Clean Enough? Determining the Most Effective Use of Resources in the Data Cleansing Process, 1–10.

Melgar, H. A. (2013). Revisión Sistemática, 1–6.

Mezzanzanica, M., Boselli, R., Cesarini, M., & Mercorio, F. (2015). A Model-Based Approach for Developing Data Cleansing Solutions. *Journal of Data and Information Quality*, *5*(4), 1–28. https://doi.org/10.1145/2641575

Microsoft. (n.d.). SQL Server 2016 SP1 Express. Retrieved October 10, 2017, from https://www.microsoft.com/es-es/sql-server/sql-server-editions-express

Minedu. (n.d.). Institución Educativa. Retrieved from http://www.mineducacion.gov.co/cvn/1665/article-82752.html

MySQL. (n.d.). MySQL 5.7 Reference Manual: Chapter 6 Security. Retrieved October 10, 2017, from https://dev.mysql.com/doc/refman/5.7/en/security.html

OMG. (2012). Business Process Model and Notation (v2.0, Vol. 125, p. 538). OMG. https://doi.org/10.1007/978-3-642-33155-8

Oracle. (2013). Siebel Marketing User Guide. Retrieved October 10, 1BC, from

https://docs.oracle.com/cd/E14004_01/books/PDF/MKTG_User.pdf

Pareja, N., & Echeverría, M. (2014). La opinión pública en la era de la información. Propuesta teórico-metodológica para su análisis en México. *Revista Mexicana de Opinión Pública*, *17*, 50–68. https://doi.org/10.1016/S1870-7300(14)70899-3

Petkovic, I. (2010). CRM in the cloud. *SIISY 2010 - 8th IEEE International Symposium on Intelligent Systems and Informatics*. https://doi.org/10.1109/SISY.2010.5647402

Petticrew M, & Roberts H. (2006). *Systematic Reviews in the Social Sciences: A Practical Guide*. USA: Blackwell Publishing.

Prasad, K. H., Faruquie, T. A., Joshi, S., Chaturvedi, S., Subramaniam, L. V., & Mohania, M. (2011). Data cleansing techniques for large enterprise datasets. *Proceedings - 2011 Annual SRII Global Conference, SRII 2011*, 135–144. https://doi.org/10.1109/SRII.2011.26

RAE. (2014). *Diccionario de la Lengua Española* (23 Edition). Madrid: Planeta Publishing.

Shu-Hui, C., & Hong-Nan, L. (2013). Antecedents and consequences of customer information quality in CRM systems: Empirical evidence from financial services firms. *Service Systems and Service Management (ICSSSM), 2013 10th International Conference on*, 467–471. https://doi.org/10.1109/ICSSSM.2013.6602493

Song, J., Liu, H., Wu, J., & Bao, Y.-B. (2015). De-Duplication Scheduling Strategy in Real-Time Data Warehouse. *The Open Cybernetics & Systemics Journal*, *9*(1), 37–43.

Stair, R. M., & Reynolds, G. W. (2016). *Fundamentals Of Information Systems* (8th ed.). Boston: Cengage Learning.

Zhou, X., Zhang, Z., & Lu, Y. (2011). Review of customer segmentation method in CRM. *2011 International Conference on Computer Science and Service System, CSSS 2011 - Proceedings*, *2*(100044), 4033–4035. https://doi.org/10.1109/CSSS.2011.5974617