UNIVERSIDAD POLITÉCNICA SALESIANA
SEDE QUITO

CARRERA:
INGENIERÍA DE SISTEMAS

Trabajo de titulación previo a la obtención del título de:
Ingeniero de Sistemas

TEMA:
CREACIÓN DE UN SISTEMA DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA (ETL) PARA LA MIGRACIÓN DE DATOS AL SISTEMA DE INFORMACIÓN NACIONAL DE LA COORDINACIÓN GENERAL DE REDES COMERCIALES DEL MAGAP

AUTOR:
FAUSTO RENÉ PARDO CÁCERES

TUTOR:
WASHINGTON RAÚL PADILLA ARIAS

Quito, marzo 2017

CESIÓN DE DERECHOS DE AUTOR

Yo, Fausto René Pardo Cáceres, con documento de identificación N° 1718006487, manifestó mi voluntad y cedo a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud que soy autor del trabajo de titulación con el tema: CREACIÓN DE UN SISTEMA DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA (ETL) PARA LA MIGRACIÓN DE DATOS AL SISTEMA DE INFORMACIÓN NACIONAL DE LA COORDINACIÓN GENERAL DE REDES COMERCIALES DEL MAGAP, mismo que ha sido desarrollado para optar por el título de INGENIERO DE SISTEMAS en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En aplicación a lo determinado en la Ley de Propiedad Intelectual, en mi condición de autor me reservo los derechos morales de la obra antes citada.

En concordancia, suscribo este documento en el momento que hago entrega del trabajo final en formato impreso y digital a la Biblioteca de la Universidad Politécnica Salesiana.

FAUSTO RENE
PARDO CACERES

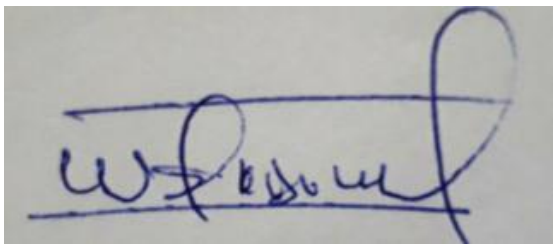
C.I:1718006487

Fecha: marzo del 2017

DECLARATORIA DE COAUTORÍA DEL DOCENTE TUTOR

Yo declaro que bajo mi dirección y asesoría fue desarrollado el Proyecto Técnico, con el tema: CREACIÓN DE UN SISTEMA DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA (ETL) PARA LA MIGRACIÓN DE DATOS AL SISTEMA DE INFORMACIÓN NACIONAL DE LA COORDINACIÓN GENERAL DE REDES COMERCIALES DEL MAGAP, realizado por Fausto Rene Pardo Cáceres, obteniendo un producto que cumple con todos los requisitos estipulados por la Universidad Politécnica Salesiana, para ser considerados como trabajo final de titulación.

Quito, marzo del 2017



WASHINGTON RAÚL PADILLA ARIAS

C.I:1707492888

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Resumen

El presente proyecto de titulación contiene el proceso tecnológico para realizar la migración de datos que están contenidos en archivos Excel hacia el Sistema de Información Nacional de la Coordinación General de Redes Comerciales del MAGAP

El contenido de este documento consta de cinco capítulos, los mismos se mencionan de forma resumida a continuación:

El capítulo uno (estado del arte), describe la situación actual de la organización a la cual se hace referencia como modelo para ejecutar la solución propuesta, además contiene toda la sustentación académica y teórica para el desarrollo del presente proyecto.

El capítulo dos (plan de migración de datos), se encarga de generar el plan de proyecto, levantamiento de los requerimientos, se evalúa la fuente de datos y asigna roles a los recursos.

El capítulo tres (Análisis y diseño de la migración de datos), donde se realiza el análisis de los datos, diseños de los procedimientos, y diseño de la arquitectura.

El capítulo cuatro (implementación y pruebas), se realiza la construcción, implementación, validación y estabilización del proyecto de migración de datos.

El capítulo cinco (cierre de la migración de datos), donde se realiza documentación, recomendaciones, conclusiones, transferencia de conocimiento y los resultados generales del proceso de migración.

Abstract

The present project has the technological process to carry out the migration of data that are contained in Excel files towards the National Information System of the General Coordination of Commercial Networks of MAGAP.

The content of this document consists of five chapters, the same that are mentioned in summary form below:

Chapter one (state of the art), describes the current situation of the organization to which reference is made as a model to execute the proposed solution, also contains all the academic and theoretical support for the development of this project.

Chapter two, is responsible for generating the project plan, lifting the requirements, evaluates the data source and assigns the roles to the resources.

Chapter Three (Analysis and Design of Data Migration), where data analysis, procedures designs, and architecture design are performed.

Chapter four (implementation and testing) performs the construction, implementation, validation and stabilization of the data migration project.

Chapter five (closing the data migration), where documentation, recommendations, conclusions, knowledge transfer and the overall results of the migration process are carried out.

INTRODUCCIÓN

La CGRC(Coordinación General de Redes Comerciales) realiza actividades para impulsar el desarrollo comercial asociado a la agricultura familiar campesina para la provisión sostenible de alimentos sanos y suficientes para la alimentación nacional y el desarrollo digno de las familias rurales. Una de las actividades que realizan es el desarrollo de CIALCO(Circuito Alternativo de Comercialización) que tiene por objeto consolidar una comercialización alternativa a los canales dominantes de distribución, a través de una política pública de precios justos para el productor y consumidor.

Toda la información de los CIALCOs se maneja de forma integral a través de un sistema informático que actualmente se está depurando para cumplir todas sus expectativas. Una necesidad importante de la CGRC es la migración de la información de CIALCOs de los años 2014, 2015 y parte del 2016 hacia el sistema informático, realizar dicha migración de forma manual conlleva un gran esfuerzo donde se requiere realizar validaciones, limpieza e integración de los datos; Por tal motivo, se justifica la realización del presente proyecto, implementando una solución tecnológica que permita migración de la información de forma eficiente y sistematizada con el fin de mitigar la carencia de recursos, información confiable, evitar información duplicada, mejorar la precisión en la recuperación de la información y reducir costos.

La construcción del proceso ETL's para la migración de la información contenida en archivos Excel de productores, organizaciones y CIALCOs hacia el sistema de la Coordinación General de Redes Comerciales permite realizar: integración con

servicios WEB, validación de los formatos, integridad e implementación de reglas de negocio, calidad de los datos y generación de scripts con la data Migrada.

Objetivo general

Crear un sistema de extracción, transformación y carga (ETL) para migrar gran parte de la información contenida en archivos Excel hacia el Sistema de Información Nacional de la Coordinación General de Redes Comerciales del MAGAP.

Objetivo Específicos

Analizar y Diseñar una solución tecnología que permita migrar la información histórica de CIALCOs contenida en archivos Excel.

Desarrollar una ETL en base a especificaciones funcionales y técnicas requeridas por la Coordinación General de Redes Comerciales.

Verificar la calidad de los datos en el proceso de migración.

Generar Scripts de inserción para PostgreSQL con la información a Migrada.

Capítulo 1

Marco Conceptual

1.1. Marco Metodológico

Department of Education Office of Federal Student Aid (2007) indica que existe 4 fases en el proceso de migración las cuales son Planificación, Análisis-Diseño, Implementación y Cierre, en algunos casos una fase pueda que requiera regresar a una fase anterior para realizar ajustes tal como se muestra en la figura 1.

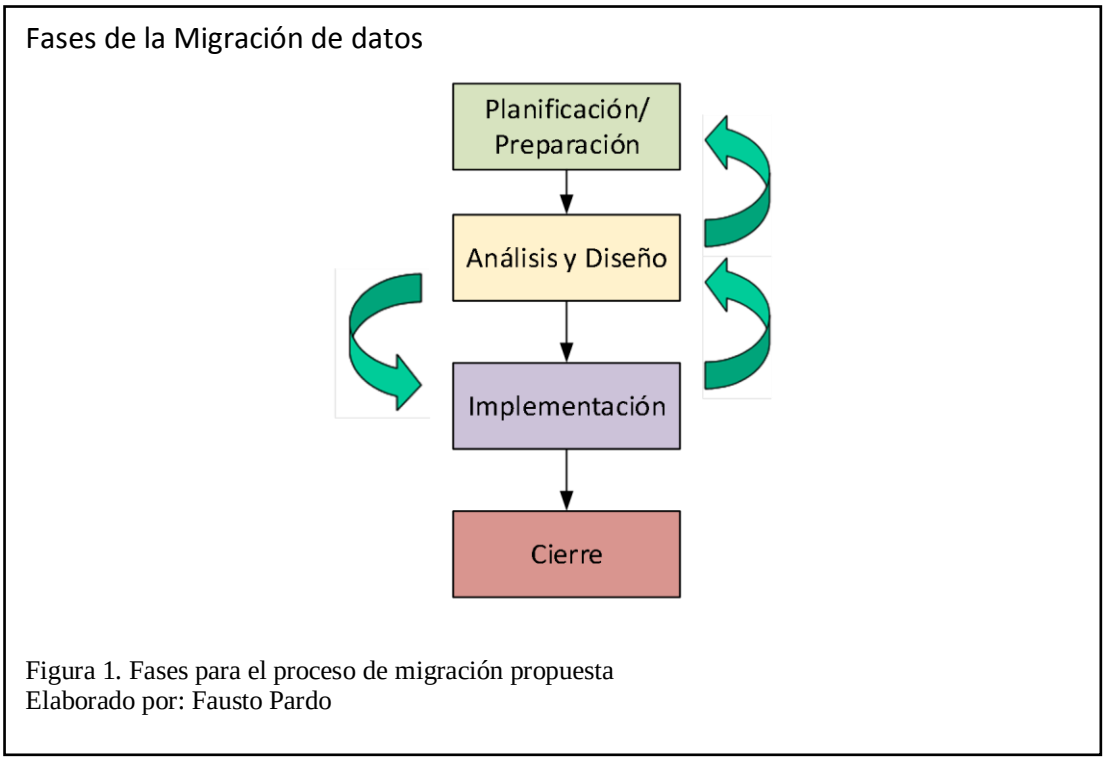


Figura 1. Fases para el proceso de migración propuesta
Elaborado por: Fausto Pardo

Estas fases se basaron en la metodología del PMBOK para alinear al proceso de gestión de proyectos.

A continuación, se mencionará en que consiste cada una de las fases.

1.1.1. Fases para la Migración de Datos

1.1.1.1. Plan de migración de datos

En esta fase se genera el plan de proyecto, levantamiento de los requerimientos, se evalúa la fuente de datos y asigna roles a los recursos.

Contempla las siguientes etapas:

Plan del Proyecto, para asegurar que tanto el proyecto de migración de datos y el proyecto de desarrollo más grandes tengan éxito, es una buena práctica desarrollar un plan de migración de datos para gestión de proyectos como un subconjunto del plan general del proyecto.

Determinar los requerimientos, requisitos para los proyectos de migración de datos son tan variados como son críticos. Delineando los requisitos de negocio para los datos ayuda a determinar qué datos a migrar. Estos requisitos pueden tomar la forma de cualquier escenario como acuerdos, expectativas, y/o objetivos de la migración. Toda la información se captura en el documento Requisitos de migración de datos.

Evaluación del entorno actual, la evaluación del entorno actual requiere la recopilación de todos los artefactos identificados para crear un plano de la actual arquitectura de datos. También se determina la tecnología a usar.

Desarrollo del plan de migración, una vez que el plan ha sido evaluado para la coherencia y el cumplimiento de las expectativas de la empresa, el plan debe ser presentado a todos los propietarios de negocios en la zona afectada y las partes interesadas para la retroalimentación y aprobación.

Definir roles y responsabilidades dentro del Equipo, los procedimientos adecuados para la fijación y documentación de las funciones y responsabilidades

para un proyecto de migración de datos no son diferentes de cualquier otro proyecto de tecnología de la información.

1.1.1.2. Análisis y diseño de una migración de datos

La tarea principal de esta fase es realizar análisis de los datos, diseños de los procedimientos, y diseño de la arquitectura.

Contempla las siguientes etapas:

Ejecutar análisis de los datos, consiste en realizar un primer análisis sobre los datos de origen, normalmente sobre tablas, con el objetivo de empezar a conocer su estructura, formato y nivel de calidad. También se evalué la tecnología de migración

Diseño del entorno de migración, se realiza el diseño de la base de ensayo (staging), arquitectura del destino, las correlaciones de data (Fuente/ensayo/Destino). Se determina la configuración de la tecnología a usar.

Diseño de los procedimientos de migración de datos, se realiza el diseño de los procedimientos para base de ensayo (staging), limpieza de datos, conversión de datos, migración de datos del destino, validación de datos y calidad de datos.

1.1.1.3. Implementación de la migración de datos

En esta fase se realiza la construcción, implementación, validación y estabilización del proyecto de migración de datos.

Contempla las siguientes etapas:

Desarrollo de procedimientos para la migración de datos, se realiza la configuración de las fuentes; Se desarrolla procedimientos para probar la limpieza, validación, limpieza y migración de los datos.

Base de Ensayo (staging), se crea, carga, integra y valida el área de ensayo.

Limpieza de datos, de acuerdo al plan de remediación de datos y validación.

Conversión y transformación de los datos, se convierte y transforma los datos. Se valida la transformación.

Migración de los datos, se ejecuta y valida una prueba de migración. Obtiene la aprobación para implementar en los ambientes de producción. Se ejecuta la implementación.

Post-Migración, se ejecuta actividades de estabilización.

1.1.1.4. Cierre de la migración de datos

En esta fase se realiza documentación, recomendaciones, conclusiones, transferencia de conocimiento y los resultados generales del proceso de migración.

1.2. Metodología de Desarrollo

La metodología de desarrollo será una combinación del PMBOK y SCRUM. El SCRUM nos permite realizar cambios no contemplados donde eventualmente sucede en los procesos de migración. La metodología de gestión de proyectos PMBOK nos permitirá llevar de forma adecuada el proyecto, para llegar en los tiempos establecidos.

1.3. Migración de Datos

Un proyecto de migración de datos se centra en el movimiento de los datos entre dos sistemas, incluyendo todos los procedimientos necesarios para la transferencia y validación de los datos desde un lugar a otro. En la figura 2 se muestra la trasferencia de datos desde las fuentes hacia el destino por medio de herramientas ETL.

Proceso de Migración de Datos con ETL

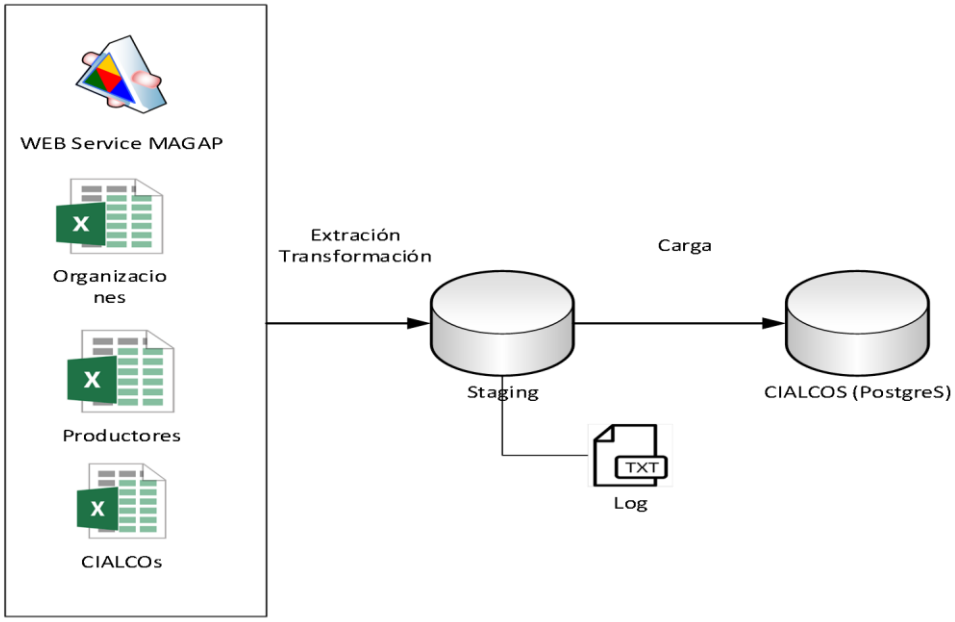


Figura 2. Proceso de migración de datos con ETL
Elaborado por: Fausto Pardo

El ciclo de vida de un proyecto de migración de datos está basado en 4 fases:

Fase 1: Plan de migración de datos

Fase 2: Análisis y diseño de la migración de datos

Fase 3: Implementación de la migración de datos

Fase 4: Cierre de la migración de datos

Fases del Proceso de Migración

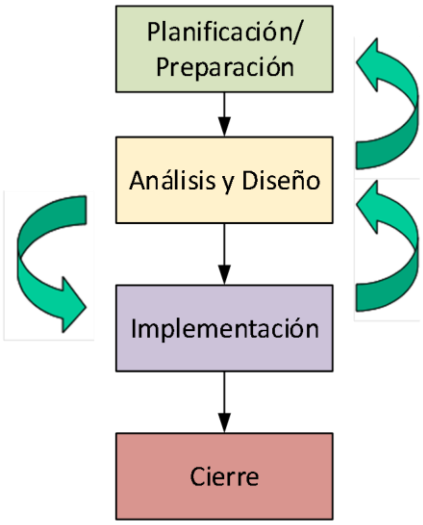


Figura 3. Propuesta de fases para el proceso de migración
Elaborado por: Fausto Pardo

1.4. Extracción, transformación y carga (ETL)

La ETL es un sistema que extrae datos de fuentes, hace cumplir las normas de calidad de los datos y la consistencia, se adecuan los datos de modo que las fuentes separadas se pueden utilizar juntos, y finalmente entrega datos en un formato definido para que los desarrolladores puedan crear las aplicaciones y los usuarios finales pueden manejar datos fiables (Kimball, 2004).

1.5. Perfilado de Datos

El procesamiento de datos es un examen sistemático de la calidad, el alcance y el contexto de una fuente de datos para permitir la construcción de un sistema ETL. En un extremo, una fuente de datos muy limpia y bien mantenida antes de que llegue al almacén de datos requiere una transformación mínima y una intervención humana para cargar directamente en tablas de dimensiones finales y tablas de hechos (Kimball, 2004).

Pero una fuente de datos sucia puede requerir: eliminar algunos campos de insumos completamente, marcar los datos perdidos, generar claves especiales de sustitución, mejorar la suposición de sustitución automática de valores dañados, la intervención humana al nivel récord y desarrollar la representación normalizada completa de los datos

1.6. Área de Ensayo

El área de preparación es la cocina del almacén de datos. Es un lugar accesible sólo para profesionales experimentados en integración de datos. Es una instalación de back-room, completamente fuera de los límites de los usuarios finales, donde los datos se colocan después de ser extraídos de los sistemas de fuentes, limpiados, manipulados y preparados para cargar en la capa de presentación. Cualquier metadata generado por el proceso ETL que sea útil para los usuarios finales debe salir de la habitación de atrás Y se ofrecerá en el área de presentación de la bodega de datos (Kimball, 2004).

1.7. Distancia de Levenshtein

Levenshtein (1965) desarrollo un algoritmo que determina el número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra, se usa ampliamente en teoría de la información y ciencias de la computación. Se entiende por operación, bien una inserción, eliminación o la sustitución de un carácter. Es útil en programas que determinan cuán similares son dos cadenas de caracteres, como es el caso de los correctores de ortografía .

El principio de Levenshtein es el mostrado en la figura 4, donde para trasformar la palabra **POMME** en **POIRE** necesita de 2 operaciones de sustitución.

Principio de Levenshtein

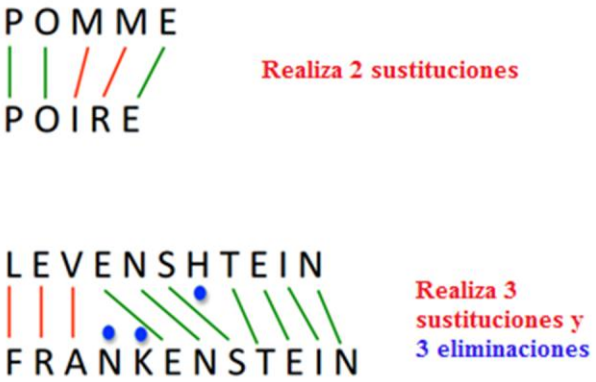


Figura 4. Principio de Levenshtein - Ejemplo
Elaborado por Fausto Pardo

El Algoritmo de Levenshtein es la mostrada en la figura 5 donde: Tenemos 2 palabras X y Y (sus tamaños son i y j respectivamente), los caracteres X1, X2….Xi corresponden a la palabra X y los caracteres Y1, Y2….Yj corresponden a la palabra Y.

Se realiza iteración de las columnas por cada fila, se compara un carácter con otro y se va encontrado el mínimo.

A la final la distancia es el número del valor di,j tal como se muestra en la figura 5.

Algoritmo de Levenshtein

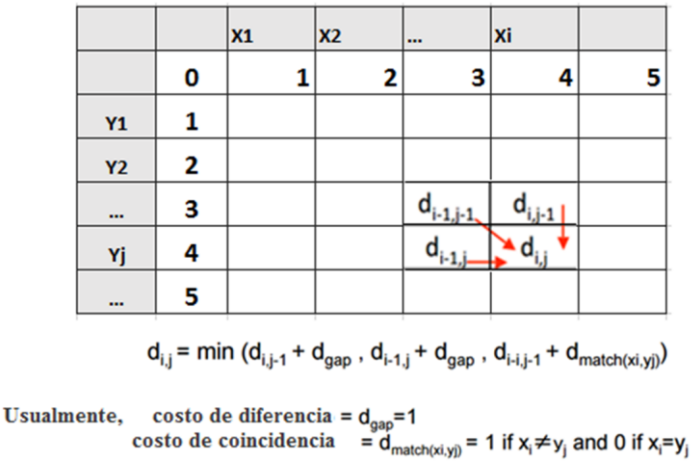


Figura 5. Algoritmo de Levenshtein para encontrar el número de saltos
Elaborado por: Fausto Pardo

Ejemplo del Algoritmo de Levenshtein

		P	O	M	M	E
	0	1	2	3	4	5
P	1	0	1	2	3	4
O	2	1	0	1	2	3
I	3	2	1	1	2	3
R	4	3	2	2	2	3
E	5	4	3	3	3	2

Distancia(POMME , POIRE) = 2

Figura 6. Ejemplo del Algoritmo de Levenshtein
Elaborado por: Fausto Pardo

Capítulo 2

Planificación de la Migración

2.1. Introducción

En base a lo especificado en el marco metodológico sobre la planificación de la migración de datos se procede a desarrollar cada uno de las etapas que se usarán a lo largo del presente capítulo.

Las etapas que se verán en este capítulo son: Plan del Proyecto, determinar los requerimientos, evaluación del entorno actual, desarrollo del plan de migración y definir roles y responsabilidades dentro del Equipo

2.2. Plan del Proyecto de Migración de datos

2.2.1. Identificación de Riesgos

Lo riesgos son los problemas más comunes que existe en todo proyecto de software, saber identificar dichos riesgos tempranamente ayuda mitigar y dimensionar de manera adecuada un proyecto de migración de datos.

En la tabla 1 se menciona los riesgos conocidos para el proyecto.

Tabla 1. Riesgos del proyecto de Migración

Cod.	Descripción	Criticidad
R1	Los archivos no tienen integridad entre sus entidades. Los campos claves han sido manipulados sin control algu	ALTA
R2	La respuesta por parte del cliente no es la esperada, se tiene que esperar las definiciones a destiempo.	ALTA
R3	La estructura de datos destino se encuentra depurándose, generando posible re-trabajo a futuro.	MEDIO
R4	Por manipular datos en Excel se tiene muchos	MEDIO

	inconvenientes que se tiene que resolver para cumplir con los objetivos. El detalle de los inconvenientes se menciona en el apartado **Evaluación del Entorno Actual**.	
R5	Si el cliente necesita validar un valor erróneo, ellos preguntan a los encargados de cada provincia demorándose en la respuesta.	CRITICO

Nota: Lista de los riesgos del proyecto de Migración
Elaborador por: Fausto Pardo

2.2.2. Plan de Mitigación de Riesgos

Los riesgos es un problema crítico que se necesita mitigar de alguna forma para prevenir desastres en el proyecto. En la tabla 2 se mencionan las diferentes alternativas para mitigar los riesgos encontrados en el proyecto.

Tabla 2. Formas de mitigación de riesgos

Código de la mitigación Riesgo	Código del Riesgo	Forma de Mitigación	Criticidad
MR1	R1	Se realizará una depuración de los archivos de carga para poder relacionar las entidades. Para lograr este propósito, se corregirá revisando los archivos matriz de respaldo de cada provincia contrastando con el archivo consolidado.	ALTA
MR2	R2	Se establecerá reuniones semanales para mencionar los avances y establecer compromisos.	ALTA
MR3	R3	Se implementará en la área de ensayo un modelo relacional homogéneo al modelo de la base de destino, para tener control de los cambios en las estructuras de la base destino.	MEDIO
MR4	R4	Con el presente proyecto se solventará los inconvenientes suscitados en este ítem.	MEDIO
MR5	R5	El cliente mencionó si sucediera eso, ellos verían la forma que no afecte en los tiempos.	MEDIO

Nota: Las diferentes formas para mitigar cada uno de los riesgos
Elaborador por: Fausto Pardo

2.3. Requerimientos de la Migración de datos

Los requerimientos funcionales se encuentran en las actas de reuniones que se han ido realizando a lo largo del proyecto, dichas actas se adjuntan al presente proyecto como anexos de documentos.

2.3.1. Los requerimientos del negocio general

Para realizar el diseño y desarrollo de la migración es importante tener claro cuáles son las expectativas del usuario con respecto al proceso que se piensa realizar. En la tabla 3 se menciona los requerimientos del negocio.

Tabla 3. Requerimientos del Negocio

Cod.	Descripción
REQ01	Que la información migrada sea **Fiable**
REQ02	Que no exista perdida de datos
REQ03	Que exista integridad de la información
REQ04	Que los procesos de migración sean controlados y organizados en caso de requerir modificar los procesos.
REQ05	Que se cumpla en los tiempos establecidos en lo posible

Nota: Los requerimientos del Negocio
Elaborador por: Fausto Pardo

2.3.2. Los requerimientos de hardware y software

La Definición de una herramienta de migración de datos es muy importante ya que debe cumplir con los objetivos de proyecto y satisfacer las necesidades que espera el cliente respecto a dicha herramienta. En la tabla 4 se muestra los requisitos que debe tener la herramienta.

Tabla 4. Requerimientos de hardware y software

Req.	Descripción
REQNF01	Que permita realizar limpieza de datos
REQNF02	Poseer un motor de datos para el procesamiento temporal de datos.
REQNF03	Que el desarrollador esté familiarizado con dicha herramienta.
REQNF04	Que tenga flexibilidad en escalar sus funcionalidades de acuerdo a las necesidades.
REQNF05	Que las herramientas sean gratuitas.

Nota: Los diferentes requerimientos de hardware y software
Elaborador por: Fausto Pardo

2.3.3. Los requerimientos de seguridad de los datos

La seguridad de los datos es un aspecto importante que se tiene que considerar en la migración de los datos, mucho más cuanto se tiene información sensible. En la tabla 5 se muestra los requerimientos de seguridad necesarios para el proyecto.

Tabla 5. Requerimientos de seguridad de datos

Req.	Descripción
REQSE01	Confidencialidad de los datos proporcionados al desarrollador del proyecto.
REQSE02	Enmascarar funcionalidades sensibles
REQSE03	Que los datos no sean compartidos con personas ajenas a la coordinación general de redes comerciales.

Nota: Los requerimientos necesarios para la seguridad de datos
Elaborador por: Fausto Pardo

2.4. Evaluación del entorno actual

2.4.1. Arquitectura del Origen de Datos

La gestión de los CIALCOs en los periodos 2014,2015 y 2016 se realiza a través de archivos Excel, toda la información es ingresada de forma manual llegando a tener los problemas que se muestra en la tabla 6.

Tabla 6. Problemas de manipular información en archivos Excel

Problema	Descripción
Redundancia de Datos	Hace referencia al almacenamiento de los mismos datos varias veces en diferentes
Carencia de integridad de datos	Ya que cada celda es única; esto hace que el documento pueda ser muy inconsistente. Un número no es necesariamente el formato establecido, así puede haber variaciones entre datos.
No existe calidad de datos	No existe procesos, técnicas, algoritmos y operaciones encaminados a mejorar la calidad de los datos de los archivos.
Límite de Espacio	Excel en su versión más reciente solo soporta 1 Millón de registros por hoja.

Cuando se trata de trabajo en Equipo	Es muy difícil saber quién modificó el documento y en qué lugar.

Nota: Problemas que se presentan al manipular información en archivos Excel
Elaborador por: Fausto Pardo

2.4.2. Tecnología que usa el origen de datos

Toda información la gestionan a través de archivos planos como es el Excel, además de dichos archivos hay que tener presente que gran parte de la información de productores se extrae de un web Service del MAGAP.

2.5. Desarrollo del plan de migración

2.5.1. Método de Migración

Proceso de extracción, transformación y carga

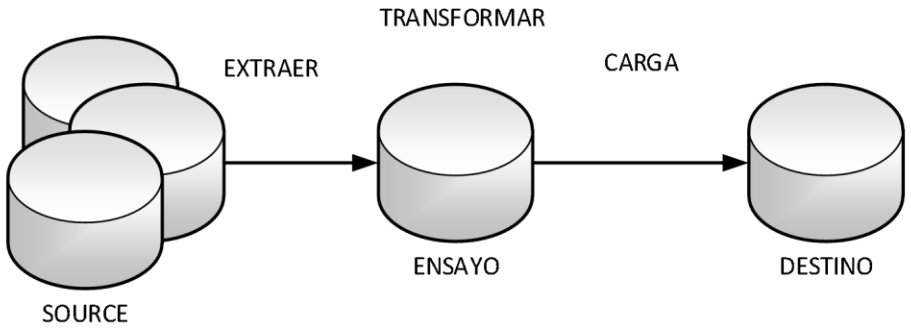


Figura 7. Proceso de extracción, transformación y carga para la limpieza de información
Elaborado por Fausto Pardo

El método de migración será basado en ETL (Extraction, Transform y Load) el cual nos permite la flexibilidad de poder agregar componentes cualquiera según como se vaya avanzando el proyecto . Además de poseer una arquitectura ya definida para la migración de datos.

El desarrollo de la migración se regirá en base a los pasos mencionados en la tabla 7 para el cumplimiento de los objetivos.

Tabla 7. Pasos del desarrollo de la migración de datos

Paso	Nombre del Paso
1	Preparación del archivo fuente GPR para corregir la integridad referencial de las entidades (organizaciones, productores y CIALCOS).
2	Validación del archivo de carga una vez depurado
3	Perfilado de Datos para saber la calidad de los datos
4	Realizar las extracciones de las fuentes de datos hacia la base de datos de ensayo.
5	Realizar las transformaciones (limpieza, formateo, homologaciones, calidad y validaciones)
6	Generar Reportes de validaciones y métricas del avance de la migración
7	Generar los scripts con la data satisfactoria
8	Ejecutar los scripts en la base de datos destino (postgreS)

Nota: Pasos del desarrollo de la migración de datos establecidos para el proyecto
Elaborador por: Fausto Pardo

2.5.2. Plan de integración de datos

El plan de integración tal como se hace referencia en el marco metodológico es la que nos permite integrar varias fuentes de datos en un solo repositorio de información. Para el presente proyecto el plan de integración contempla los pasos descritos en la tabla 8.

Tabla 8. Pasos del plan de integración de datos

Paso	Nombre del Paso
1	Llevamos la información de los archivos Excel hacia la base de datos de ensayo.
2	Por medio de las cedulas de los productores realizamos llamadas al servicio WEB del Magap para extraer la información.

3	Las estructuras de provincias, cantones y parroquias usados en postgreSQL se los carga en la base de ensayos por medio de scripts.
4	Los catálogos usados en postgreSQL se los carga en la base de ensayos por medio de scripts.

Nota: Pasos del plan de integración de datos para la extracción.
Elaborador por: Fausto Pardo

2.5.3. Plan de calidad de datos

La calidad de los datos es muy importante en el proceso de migración de datos, ya que la misma establece puntos de control en todo el proceso de migración de los datos que son traslados desde la fuente hacia el sistema desti Sin ella la migración no tendría sentido.

A continuación, se mencionada las actividades que regirán la calidad de datos.

Implementación de métricas como: Cantidad de registros procesados satisfactoriamente, cantidad de registros erróneo, porcentaje de registro procesados satisfactoriamente y segmentación por categorías y tipo de mensaje (Error o Advertencias)

Tolerancia a la perdida de datos, como la fuente y destino de datos son diferentes es importante poner reglas de validaciones que permitan identificar dichos los campos o registros erróneos.

Perfilado de Datos, el perfilado de los datos es una forma de medir la calidad de la información de la fuente de datos antes de ser procesado para tener una idea de lo que se piensa migrar.

Políticas de calidad de datos como: Todas las columnas que tenga vacío o nulos se pondrá nulo si no es obligatorio, los campos obligatorios que son enteros se pondrá

por default cero, en caso de no existir datos, los campos obligatorios que son cadenas se pondrá por default vacío ‘’, en caso de no existir datos y para evitar la duplicación de datos se agrupará los datos por los campos únicos.

2.5.4. Plan de Pruebas de la Migración de datos

Todos los procedimientos de movimiento de datos, procedimientos de transformación / conversión, los procedimientos de limpieza de datos y procedimientos de validación de datos deben estar en el contexto de la arquitectura de migración de datos. Los planes de pruebas de una migración de datos buscan la satisfacción del cliente por medio de las siguientes actividades:

Conteo de Registros como: Conteo de los registros de la extracción de datos, conteo de los registros de la transformación de datos, conteo de los registros del destino de datos.

Validaciones de los campos como: Validar campos vacíos, nulos o incorrectos, validar campos que estén atados a una regla de negocio en particular, validar integridad de los datos y visualizar en reportes las métricas de las pruebas

2.6. Definir roles y responsabilidades dentro del Equipo

Para cubrir a cabalidad el proyecto en la tabla 9 se mencionan los roles y responsabilidades de cada persona involucrada en el proceso de migración de datos.

Tabla 9. Roles y responsabilidades del equipo

Nombre	Rol	Responsabilidad
Pablo Roberto	Director de Normativa de la	Sponsor del

Izquierdo	Coordinación General de Redes Comerciales	proyecto
Julio Cabezas Giménez	Técnico de la dirección de Normativa	Usuario Técnico
Clara Delgado	Técnico de la dirección de Normativa	Usuarios de la parte interesada
Washington Padilla	Tutor	Tutorías
Fausto Pardo	Tesista	Desarrollador la solución

Nota: Se menciona los roles y responsabilidades de los miembros del equipo
Elaborador por: Fausto Pardo

Capítulo 3

Análisis y Diseño

3.1. Introducción

En base a lo especificado en el marco metodológico sobre la análisis y diseño de la migración de datos se procede a desarrollar las etapas que se usarán a lo largo del presente capítulo.

Las etapas que se utilizarán en este capítulo son: Análisis de la Migración de Datos, analizar ambiente actual, evaluar la tecnología de Migración, evaluar la calidad de datos y perfilado de datos.

En lo que respecta a Diseño de la Migración de Datos se realizará las siguientes etapas: Diseño de la arquitectura, diseño de objetos del área de ensayo, diseño de limpieza de Datos, diseño de las validaciones de los datos, mapeo de datos(Fuente/ensayo/Destino), diseño del algoritmo de coincidencias y configuración de la migración.

3.2. Análisis de la Migración de Datos

3.2.1. Analizar ambiente actual

Actualmente el ambiente de migración esta dividió en 3 partes, que es el origen de los datos, la base de ensayo y el desti Cada parte tiene tecnología diferente como se muestra en la tabla 10.

Tabla 10. Elementos del ambiente actual

Nombre	Tecnología	Tipo
Archivos del GPR	Archivos en formato Excel	FUENTE DE DATOS
Servicio WEB	Tecnología SOAP	FUENTE DE DATOS
Base de ensayo	SQL Server express	BASE DE ENSAYO

CIALCOS	Base del sistema nuevo POSTGRES	DESTINO

Nota: Los elementos de la arquitectura ambiente actual
Elaborador por: Fausto Pardo

3.2.2. Evaluar la tecnología de migración

La tecnología a usar es propietaria con ediciones gratuitas y de evaluación, para realizar la migración de los datos. Está cumple con todos los requisitos técnicos para realizar la migración.

En la tabla 11, se menciona cada una de las herramientas que se usará en el desarrollo del proyecto:

Tabla 11. Tecnología de la migración

Herramienta	Propósito
Visual Studio COMUNITY 2013	IDE para las ETLs Link de descarga: https://www.visualstudio.com/en-us/news/releasenotes/vs2013-community-vs
SQL Data Tools Business Intelligence for Visual Studio 2013	Para realizar las ETLs. Link de descarga: https://www.microsoft.com/es-co/download/details.aspx?id=42313
SQL server 2014 Express	Para la base de datos de ENSAYO. Link de descarga: https://www.microsoft.com/en-us/download/details.aspx?id=42299
Power BI Desktop y Office 365	Para mostrar los dashBoard del Proyecto
Computador core I7 con Windows 10 64bits	Computador para alojar todas las herramientas de migración.

Nota: La tecnología usada para el proceso de migración
Elaborador por: Fausto Pardo

3.2.3. Evaluar la calidad de datos

Al evaluar la calidad de datos es importante definir las métricas que sustenten y entreguen valor a la calidad de datos. En la tabla 12 se menciona las métricas de calidad de datos del proyecto.

Tabla 12. Métricas de calidad de datos

Métrica	Descripción	Fase
Cantidad de Productores	Se quita los duplicados y se cuenta los registros de los productores	EXTRACCION
Cantidad de Organizaciones	Se quita los duplicados y se cuenta los registros de los organizaciones	EXTRACCION
Cantidad de CIALCOs	Se quita los duplicados y se cuenta los registros de los CIALCOs	EXTRACCION
Cantidad de Productores satisfactorios	Conteo de registros de la estructura de trasformación de productores	TRANSFORMACION
Cantidad de Organizaciones satisfactorios	Conteo de registros de la estructura de trasformación de Organizaciones	TRANSFORMACION
Cantidad de CIALCOs satisfactorios	Porcentaje de registros de la estructura de trasformación de CIALCOs	TRANSFORMACION
Porcentaje de Productores satisfactorios	Porcentaje de registros de la estructura de trasformación de productores	TRANSFORMACION
Porcentaje de Organizaciones satisfactorios	Porcentaje de registros de la estructura de trasformación de Organizaciones	TRANSFORMACION
Porcentaje de CIALCOs satisfactorios	Porcentaje de registros de la estructura de trasformación de CIALCOs	TRANSFORMACION

Nota: Métricas de calidad de datos implementados en el proyecto
Elaborador por: Fausto Pardo

3.2.4. Perfilado de Datos (Antes de Depurar el Archivo)

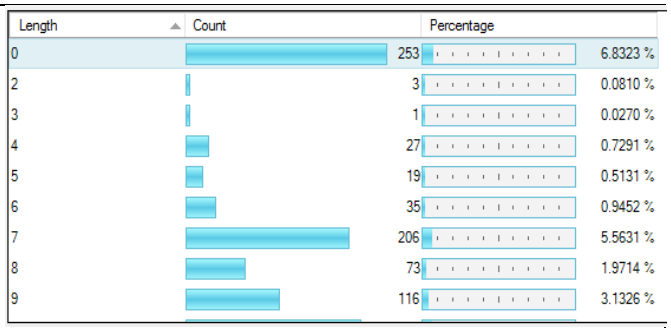
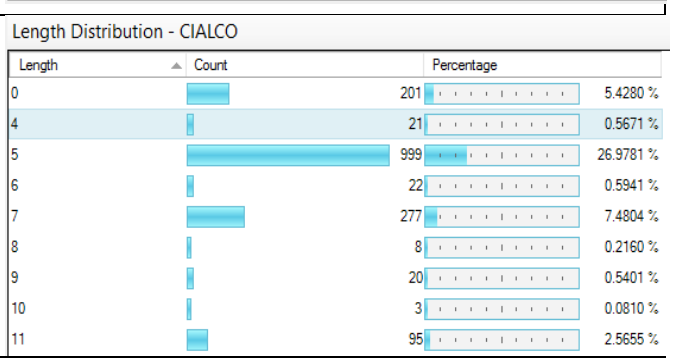
Los archivos de Excel serán modificados para aumentar la consistencia y calidad de datos de la información que se va procesar. Por tal motivo es necesario realizar un perfilado de datos antes de que se realice la depuración de los archivos excel para poder contrastar a futuro.

3.2.4.1. Productores

El reporte y análisis del perfilado de datos de la estructura de productores se muestra a continuación en la tabla 13.

Tabla 13. Perfil de Datos(antes del archivo depurado) de Productores

PERFIL DE DATOS - PRODUCTORES	
Clave Única	
Clave Única La clave candidata la combinación de ANIO, CUATRIMESTRE, CEDULA, ORGANIZACION, CIALCO se ha encontrado los siguientes registros duplicados	Key Violations ANIO | CEDULA | CIALCO | CUA | ORGANIZACION | Count 2015 | 104694690 | FERIA DE PRODUCTORES SANTA ISABEL | 3 | SAN ALFONSO | 2 2015 | 100433507 | FERIA DE PRODUCTORES AGROECOLOGICOS DE CHICAN | 3 | AGUAS BLANCAS | 2 2015 | 0908053986 | PIE DE FINCA | 1 | | 2 2015 | 0603146648 | PIE DE FINCA | 1 | | 2 2015 | 0400785531 | CANASTA RED DE PRODUCTORES SOSTENIBLES | 3 | MEDALLA MILAGROSA | 2 2015 | 0201157708 | PIE DE FINCA | 2 | PRODUCTOR INDEPENDI... | 2 2015 | 0200845212 | PIE DE FINCA | 1 | | 2 2015 | 02001418829 | PIE DE FINCA | 1 | | 2 2015 | 0105438667. | FERIA | 2 | APAY | 2 2015 | 0104030028. | FERIA | 2 | APAY | 2 2015 | 0103734687. | FERIA | 2 | APAY | 2
Tamaño de Datos	
Campo	**Gráfico**
Cédula Gran parte de la totalidad de productores tiene cédulas con diferentes tamaños a 10	Length | Count | Percentage 0 | 1 | 0.0270 % 1 | 1 | 0.0270 % 7 | 1 | 0.0270 % 8 | 12 | 0.3241 % 9 | 576 | 15.5550 % 10 | 2044 | 55.1985 % 11 | 1060 | 28.6254 % 12 | 7 | 0.1890 % 19 | 1 | 0.0270 %

Organización Existen 253 organizaciones que no tienen un valor asignado. En los casos de tamaño 2, está asignado la palabra "NO".	
CIALCO Existen 201 CIALCOs que no tienen un valor asignado.	

Nota: Perfil de Datos (antes del archivo depurado) de Productores por campos
Elaborador por: Fausto Pardo

3.2.4.2. Organizaciones

El reporte y análisis del perfilado de datos de la estructura de organizaciones se muestra a continuación en la tabla 14.

Tabla 14. Perfil de Datos (antes del archivo depurado) de Organizaciones

PERFIL DE DATOS - ORGANIZACIONES	
Clave Única	
Clave Única La clave candidata la combinación de ANIO, CANTON, RUC, CUA, CIALCO se ha encontrado los siguientes registros duplicados.	Se ha encontrado bastantes inconsistencias, esto se debe a su mayoría no existe RUC.
Tamaño de Datos	
Campo	**Gráfico**
Ruc Existen 253 registros que no tienen RUC.	

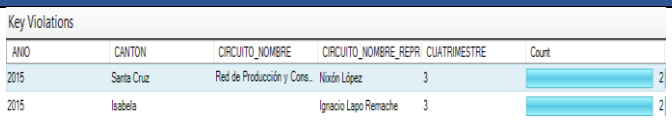
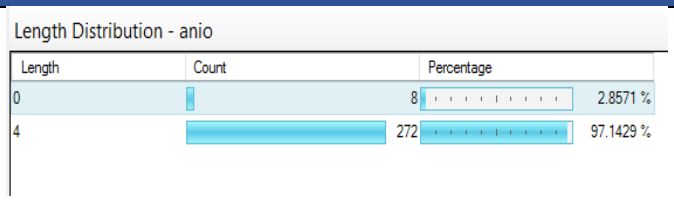
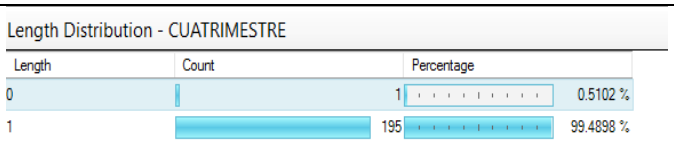
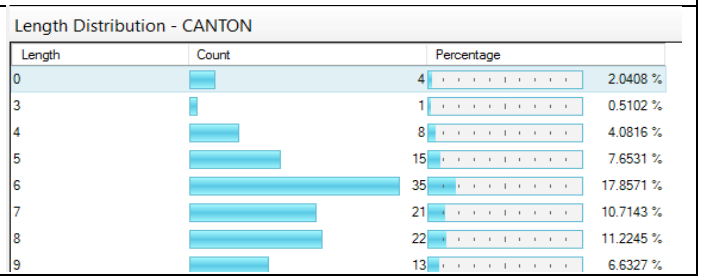
	Length Distribution - ruc Length / Count / Percentage 0 / 253 / 90.3571 % 8 / 1 / 0.3571 % 9 / 1 / 0.3571 % 10 / 2 / 0.7143 % 13 / 23 / 8.2143 %
Cua Existen 10 registros que no tienen CUATRIMESTRE.	Length Distribution - cua Length / Count / Percentage 0 / 10 / 3.5714 % 1 / 270 / 96.4286 %
Nombre del CIALCO Existen 106 registros que no tienen el nombre de CIALCO.	Length Distribution - circuitoNombre Length / Count / Percentage 0 / 106 / 37.8571 % 5 / 2 / 0.7143 % 6 / 1 / 0.3571 % 7 / 2 / 0.7143 % 8 / 4 / 1.4286 % 9 / 4 / 1.4286 % 10 / 6 / 2.1429 %
Año Existen 8 registros que no tienen el ANIO.	Length Distribution - anio Length / Count / Percentage 0 / 8 / 2.8571 % 4 / 272 / 97.1429 %
Cantón Existen 10 registros que no tienen el CANTON.	Length Distribution - canton Length / Count / Percentage 0 / 10 / 3.5714 % 4 / 15 / 5.3571 % 5 / 28 / 10.0000 % 6 / 75 / 26.7857 % 7 / 39 / 13.9286 %
Num Registro SEPs Existen 271 registros que no tienen asignado un valor.	Length Distribution - numRegistroSEPS Length / Count / Percentage 0 / 271 / 96.7857 % 9 / 2 / 0.7143 % 10 / 2 / 0.7143 % 22 / 5 / 1.7857 %
Num. Reg. MAGAP Existen 274 registros que no tienen asignado un valor.	Length Distribution - numRegistroMAGAP Length / Count / Percentage 0 / 274 / 97.8571 % 8 / 3 / 1.0714 % 10 / 3 / 1.0714 %

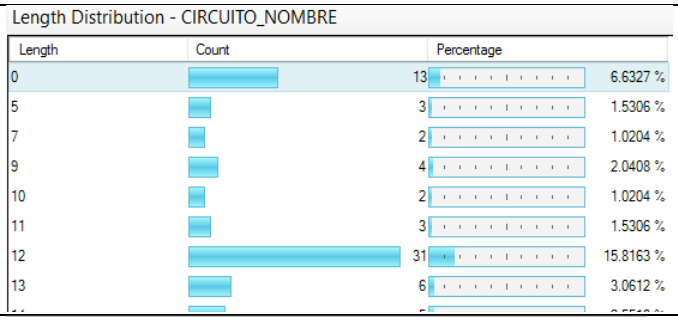
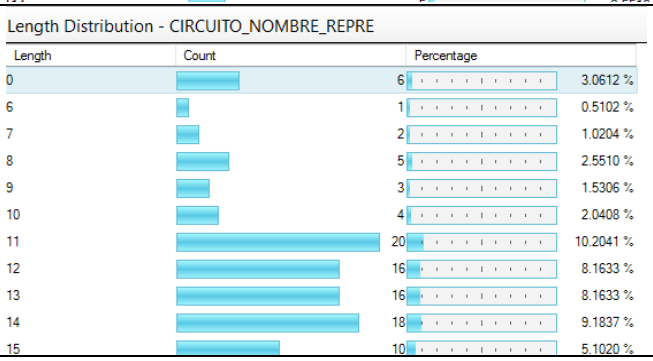
Nota: Perfil de Datos (antes del archivo depurado) de Organizaciones por campos
Elaborador por: Fausto Pardo

3.2.4.3. CIALCOS

El reporte y análisis del perfilado de datos de la estructura de CIALCOs se muestra a continuación en la tabla 15.

Tabla 15. Perfil de Datos (antes del archivo depurado) de CIALCOS

PERFIL DE DATOS - CIALCOS	
Clave Única	
Clave Única La clave candidata la combinación de ANIO, CUATRIMESTRE, CANTON, CIRCUITO, REPRESENTANTE, se ha encontrado los siguientes registros duplicados	
Tamaño de Datos	
Campo	**Gráfico**
Año Existen 8 registros que no tienen el ANIO.	
Cuatrimestre Existen 1 registros que no tienen el CUATRIMESTRE.	
Cantón Existen 4 registros que no tienen el CANTON.	

CIRCUITO Existen 13 registros que no tienen el CIRCUITO.	
Nombre del Representante Existen 6 registros que no tienen el NOMBRE DE REPRESENTANTE.	

Nota: Perfil de Datos (antes del archivo depurado) de CIALCOs por campos de la estructura.
Elaborador por: Fausto Pardo

3.3. Diseño de la Migración de Datos

3.3.1. Diseño de la arquitectura

Arquitectura de la Migración

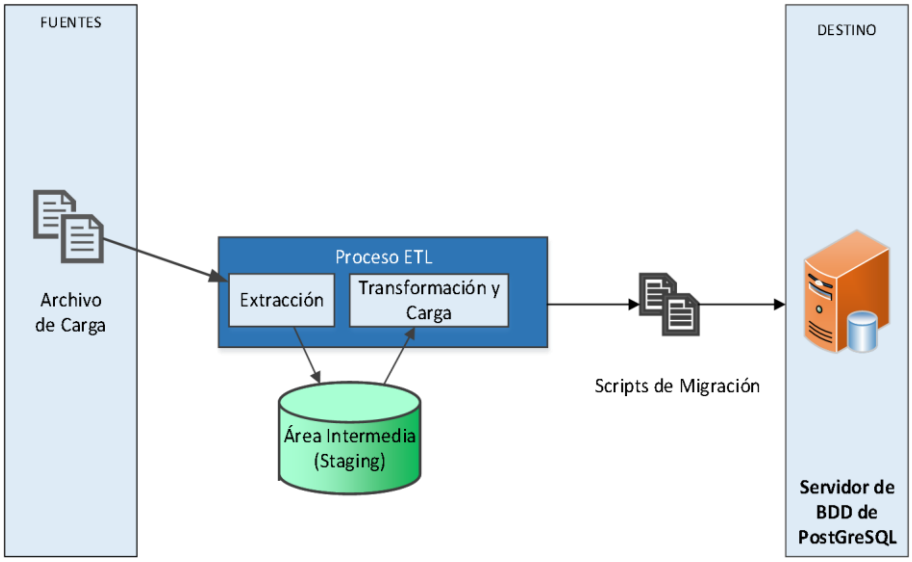


Figura 8. Arquitectura de la Migración que contempla como base un proceso ETL
Elaborado por: Fausto Pardo

En la figura 8 se ilustra la arquitectura la cual contiene los siguientes elementos:

Archivos de Carga(Fuente), Son los archivos Excel de CIALCOs de los años 2014, 201 y 2016

Proceso de ETL, Realiza la extracción de la información desde la fuente, los pone en el área de ensayo, realiza la transformación y posteriormente genera los scripts de inserción para ser ejecutados en la base de datos de PostGreSQL.

En la base de datos de ensayo, se realiza los procesos de estandarización, limpieza, transformación y homologación de datos.

Servidor de BDD de PostGreSQL, Es el destino se ejecuta lo scripts y se realiza la validación respectiva.

3.3.2. Diseño de objetos del área de ensayo

3.3.2.1. Fuentes de Datos

En la tabla 16 se muestra las fuentes de datos que se procesarán.

Tabla 16. Fuentes de datos del área de ensayo

Tipo de Fuente	Nombre	Descripción
Archivo	Consolidado GPR 2015.xlsx	Archivo consolidado de organizaciones, productores y CIALCOs del 2015
Archivo	Consolidado GPR 2016.xlsx	Archivo consolidado de organizaciones, productores y CIALCOs del 2016
Archivo	organizacion20151216.xls	Archivo del registro de Organizaciones(Supiero n manifestar que el archivo proviene de una fuente de datos confiable que gestionan las organizaciones).
Archivo	OrganizacionesMiembroMagap. xls	Archivo de Miembros que se encuentran en cada uno de las organizaciones.
Archivo	miembrosMagap.xls	Archivos de Miembros que se encuentran en cada uno de las organizaciones.
WEB Service	http://sinagap.magap.gob.ec/Enl aces /Service.asmx	Obtener información relevante de los productores por medio de la cédula.

Nota: La fuente de información para ser cargado en el área de ensayo
Elaborador por: Fausto Pardo

3.3.2.2. Estructuras

En la siguiente figura 9 se muestra del modelo lógico del base de datos de ensayo la cual se usará para insertar la información procesada satisfactoriamente.

Modeló logico de la área de ensayo

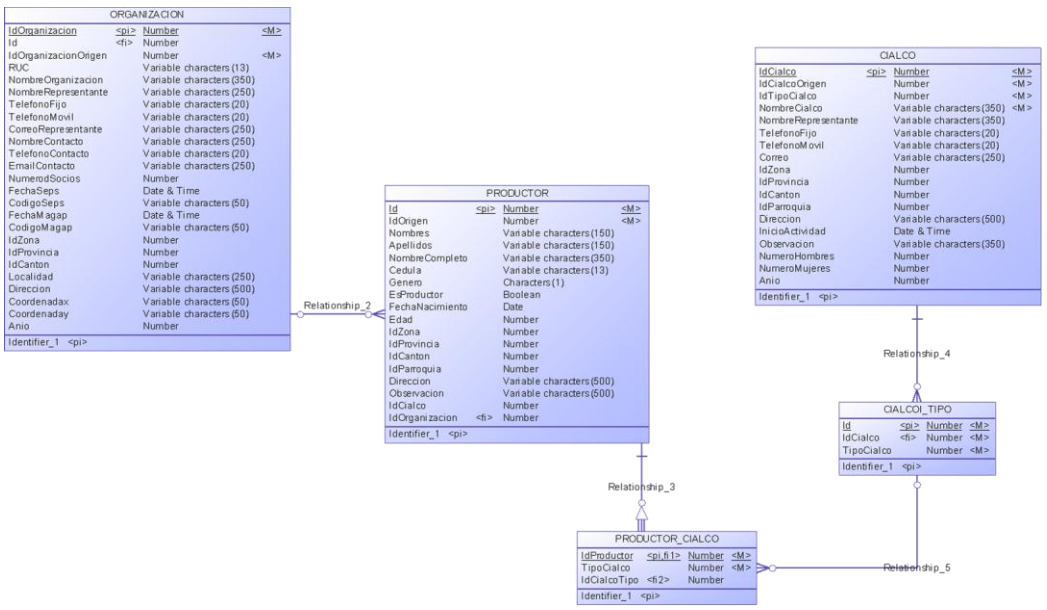


Figura 9. Modelo lógico del área de ensayo establecido para el proyecto
Elaborado por Fausto Pardo

El área de ensayo es una base de datos intermedia que nos permite realizar la extracción y transformaciones de los datos, para después pasar toda la información a la base de datos destino. En la tabla 17 se muestran las estructuras de la base de datos de ensayo.

Tabla 17. Estructuras del área de ensayo

Nombre de la Estructura	Funcionalidad
temporal. PRODUCTORES	Productores del GPR
temporal. PRODUCTORES_RC	Productores del WEB services
temporal.LISTA_ORGANIZACIONES	Organizaciones del GPR
temporal. CONSOLIDADO_GPR	CIALCOs del GPR
temporal. ORGANIZACIONES20152016	Registro de Organizaciones

temporal.COINCIDENCIAS_LEVENSHTEI_ORGANIZACION	Palabras de coincidencias entre las Organizaciones del GPR y el registro de organizaciones
temporal.COINCIDENCIAS_ORGANIZACION	Coincidencias de Nombre entre las Organizaciones del GPR y el registro de organizaciones
temporal. MAPEO_HOMOLOGACIONES	Mapeo de Organizaciones, Organizaciones del GPR y el registro de organizaciones
transformacion.CIALCO	Los CIALCOS que han sido transformados
transformacion.CIALCO_TIPO	Los TIPOS DE CIALCOS que han sido transformados
transformacion.ORGANIZACION	Las ORGANIZACIONE S que han sido transformados
transformacion.PRODUCTOR	Los PRODUCTORES que han sido transformados
transformacion.PRODUCTOR_CIALCO	Los TIPOS DE PRODUCTORES que han sido transformados

Nota: Las estructuras del área de ensayo
Elaborador por: Fausto Pardo

3.3.2.3. Paquetes ETL

Los paquetes ETL son los encargados de realizar tareas como trasformación, limpieza y carga desde fuentes de datos hacia un destino de datos. Adicional, tienen la facultad de desarrollar los traslados de información a manera de un flujo de trabajo, y se puede realizar pontos de control en caso de requerir.

En la tabla 18 se menciona los paquetes que contiene los procesos del área de ensayo.

Tabla 18. Paquetes ETL del área de ensayo

Paquete	Descripción
EX_CIALCOs.dtsx	Paquete para la extracción de CIALCOS
EX_Organizaciones.dtsx	Paquete para la extracción de Organizaciones
EX_Productores.dtsx	Paquete para la extracción de Productores
EX_Redes.dtsx	Paquete para la extracción de Redes

TR_CIALCOS_ORG_PROD.dtsx	Paquete para la transformación de CIALCOS, Organizaciones y Redes
TR_ORG_COINCIDENCIAS.dtsx	Paquete para la generación de coincidencias de Organizaciones
LD_CIALCOS_ORG_PROD.dtsx	Paquete para la generación de los SCRIPTS de carga
BATCH.dtsx	Paquete que ejecuta todos los paquetes de forma ordenada.

Nota: Paquetes ETL usados en el área de ensayo para la limpieza.
Elaborador por: Fausto Pardo

En la figura 10 se muestra el flujo de migración general que genera los scripts de inserción de la data satisfactoria para la base de datos de postgreSQL. Toda la lógica es plasmada en el paquete BATCH.dtsx.

Flujo de los paquetes ETL del área de ensayo

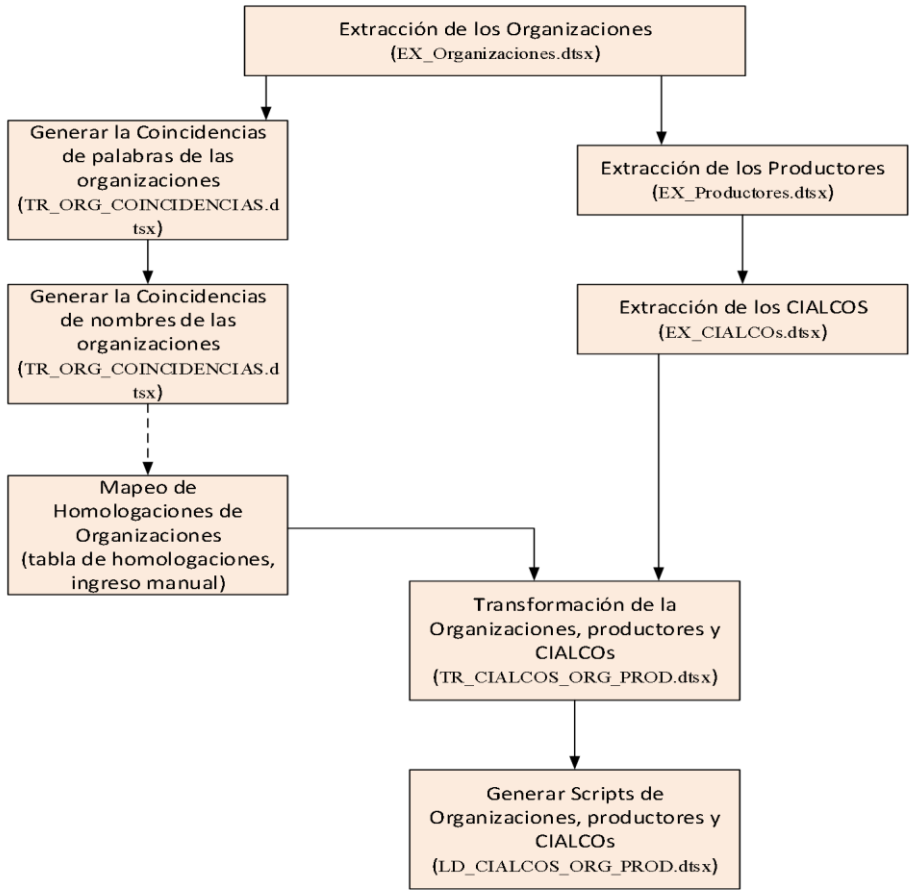


Figura 10. Flujo de los paquetes ETL desarrollado para la transformación de la informacion
Elaborado por: Fausto Pardo

3.3.2.4. Procedimientos Almacenados

En la tabla 19 se menciona los procedimientos almacenados que se deben desarrollar para realizar actividades como trasformación, limpieza y carga.

Tabla 19. Procedimientos almacenados del área de ensayo

Proceso	Descripción	Paquete
[SP_GENERAR_ COINCIDENCIAS_ PALABRAS_LEVENSHTEI_ ORGANIZACIONES]	Procedimiento Almacenado para la generación de coincidencias por palabra de las organizaciones	TR_ORG_ COINCIDENCIAS.dtsx
SP_GENERAR_ COINCIDENCIAS_ PORCENTAJES_ ORGANIZACIONES	Procedimiento Almacenado para la generación de coincidencias con porcentaje de las organizaciones	TR_ORG_ COINCIDENCIAS.dtsx
SP_TRANSFORMAR_ CIALCOS	Procedimiento Almacenado para la transformación, limpieza y validación de los CIALCOs	TR_CIALCOS_ ORG_PROD.dtsx
SP_TRANSFORMAR_ ORGANIZACIONES	Procedimiento Almacenado para la transformación, limpieza y validación de las Organizaciones	TR_CIALCOS_ ORG_PROD.dtsx
SP_TRANSFORMAR_ PRODUCTORES	Procedimiento Almacenado para la transformación, limpieza y validación de los Productores	TR_CIALCOS_ ORG_PROD.dtsx
SP_GENERAR_SCRIPT_ CIALCOS	Procedimiento Almacenado para la generación de los scripts de CIALCOs	LD_CIALCOS_ ORG_PROD.dtsx
SP_GENERAR_SCRIPT_ ORGANIZACIONES	Procedimiento Almacenado para la generación de los scripts de Organizaciones	LD_CIALCOS_ ORG_PROD.dtsx
SP_GENERAR_SCRIPT_ PRODUCTORES	Procedimiento Almacenado para la generación de los scripts de Productores	LD_CIALCOS_ ORG_PROD.dtsx

Nota: Procedimientos almacenados usados en el área de ensayo
Elaborador por: Fausto Pardo

3.3.3. Diseño de limpieza de Datos

Para lograr definir los procedimientos es necesario analizar los resultados del perfil de datos y el plan de calidad de datos.

De forma general, los procesos estarán regidos bajo las convenciones siguientes: Todos los campos vacíos colocar nulo, quitar espacios al inicio y final de los campos tipo textos y remover los caracteres especiales en los campos tipo textos.

La limpieza se realiza exclusivamente por cada entidad (productores, organizaciones y CIALCOs) las cuales se menciona en los siguientes temas.

3.3.3.1. Limpieza en Productores

La limpieza de datos de la estructura de productores se realizará en base a las especificaciones de la tabla 20.

Tabla 20. Limpieza de los productores

Productores		
Campo	**Error**	**Acciones para Corregir**
Cédula	Tiene valores con 9 caracteres	Se añade un cero al inicio
Cédula	La cantidad de caracteres es menos que 9 , Es mayor a 10 caracteres o es vacío	Se validará con los encargados de las provincias para que mencione la cédula real.
Organización	677 Registros vacíos	Pertenecen a las productores que participan en un CIALCO directamente de forma independiente .Poner le valor de nulo
CIALCO	121 registros con vacíos	Colocar el valor de nulo para identificar el CIALCO.

Nota: Limpieza de la información realizada en Productores
Elaborador por: Fausto Pardo

3.3.3.2. Limpieza en Organizaciones

La limpieza de datos de la estructura de organizaciones se realizará en base a las especificaciones de la tabla 21.

Tabla 21. Limpieza de las organizaciones

Organizaciones		
Campo	**Error**	**Acciones para Corregir**
Ruc	Tiene 150 registros que no tiene valor asignado	Hay que obtener el RUC del registro de organizaciones por medio del nombre.
CIALCO	415 registros que no tiene asignado el nombre de CIALCO.	Hay que obtenerla relación del CIALCO y Organización a través de los productores.
Cantón	327 registros sin valor asignado	Hay que poner nulo
Num. Reg. SEPS	Existen 149 registros que no tienen asignado un valor.	Hay que poner nulo
Num. Registros MAGAP	Existen 152 registros que no tienen asignado un valor.	Hay que poner nulo

Nota: Limpieza de la información realizada en las organizaciones
Elaborador por: Fausto Pardo

3.3.3.3. Limpieza en CIALCOS

La limpieza de datos de la estructura de CIACOs se realizará en base a las especificaciones de la tabla 22.

Tabla 22. Limpieza de los CIALCOS

CIALCOS		
Campo	**Error**	**Acciones para Corregir**
Cantón	Tiene 118 registros que no tiene valor asignado	Hay que poner nulo
Nombre del CIALCO	Existen 10 registros que no tienen el Nombre.	Hay que validar con las provincias para saber que nombre le corresponde.

Nombre del Representante	Existen 118 registros que no tienen valor asignado.	Hay que poner nulo

Nota: Limpieza de la información realizada en los CIALCOs
Elaborador por: Fausto Pardo

3.3.4. Diseño de las validaciones de los datos

3.3.4.1. Reglas de validación de productores

Las Reglas de Validaciones de la estructura de productores se realizará en base a las especificaciones de la tabla 23.

Tabla 23. Reglas de validación de productores

Campo	Regla de Validación	Mensaje	Tipo de mensaje
Cédula	Validar contra el servicio WEB si la cedula es válida.	El servicio WEB http://sinagap.magap.gob.ec /Enlaces/Service.asmx ?wsdl tiene un error en la cédula	Error
Cédula	Validar si la cedula es vacío o nulo y si supera el tamaño de 10 caracteres	La cédula del productor es inválida	Error
Nombre de la Organización	Si supera el tamaño de 350	El nombre de la organización supera el tamaño de '[nombre] ' caracteres	Error
Nombre Completo	Si supera el tamaño de 350	'El Nombre Completo del Productor supera el tamaño de '+ [Nombre Completo]+' caracteres'	Error
Nombre	Si supera el tamaño de 150	'El Nombre del Productor está vacío	Error

Nombre	El nombre está vacío o nulo	'El Nombre del Productor supera el tamaño de '+ [Nombre]+' caracteres'	Error
Apellido	Si supera el tamaño de 150	'El Apellido del Productor supera el tamaño de '+ [Apellido]+' caracteres'	Error
Apellido	El Apellido está vacío o nulo	El Apellido está vacío	Error
Genero	El Genero está vacío o nulo	El Genero está vacío	Error
Edad	La Edad está vacío o nulo	La Edad está vacío	Error
Nombre de la organización	Validar si el nombre de la organización ya está procesada	El productor no tiene asignado una organización	Error
Nombre del CIALCO	El nombre del CIALCO está vacío o nulo	El productor no tiene asignado un Nombre de Cialco	Advertencia
Tipo de CIALCO	Validar si el productor está asignado a un CIALCO que es diferente a FERIA,CANASTA, COMPRA PUBLICA,PIE DE FINCA Y TIENDA	El productor no tiene algún tipo de cialco (como feria , canasta..etc)	Advertencia
Tipo de Cialco y Nombre	Validar si el productor no tiene coincidencia con el CIALCO por medio de Nombre y Tipo de cialco	El productor no tiene coincidencia con el CIALCO por medio de Nombre y Tipo de cialco	Advertencia

Nota: Reglas de validación implementado en la limpieza de productores
Elaborador por: Fausto Pardo

3.3.4.2. Reglas de validación de organización

Las Reglas de Validaciones de la estructura de organizaciones se realizará en base a las especificaciones de la tabla 24.

Tabla 24. Reglas de validación de organizaciones

Campo	Regla de Validación	Mensaje	Tipo de mensaje
Nombre de la organización	El nombre de la organización supera el tamaño de 350	El nombre de la organización supera el tamaño de 350 caracteres	Error
Ruc	El Ruc supera el tamaño de 13	El Ruc supera el tamaño de 13 caracteres	Error
Ruc	El Ruc está vacío o tiene menos de 13 caracteres	'El RUC de la organización tiene'+[tam de Ruc] +' caracteres'	Error
Nombre de Representante	El Nombre de Representante supera el tamaño de 250	El Nombre de Representante supera el tamaño de 250 caracteres	Error
Nombre de Representante	El Nombre de Representante supera el tamaño de 250	El Nombre de Representante supera el tamaño de 250 caracteres	Error
Teléfono Fijo	El Teléfono Fijo supera el tamaño de 20	El Teléfono Fijo supera el tamaño de 20	Error
Teléfono Móvil	El Teléfono Móvil	El Teléfono Móvil	Error

	supera el tamaño de 20	supera el tamaño de 20	
Correo del Representante	El Correo del Representante supera el tamaño de 250	El Correo del Representante supera el tamaño de 250	Error
Nombre del Contacto	El Nombre del Contacto supera el tamaño de 250	El Nombre del Contacto supera el tamaño de 250	Error
Teléfono del Contacto	El Teléfono del Contacto supera el tamaño de 20	El Teléfono del Contacto supera el tamaño de 20	Error
Correo del Contacto	El Correo del Contacto supera el tamaño de 250	El Correo del Contacto supera el tamaño de 250	Error
Num. De Socios	Validar si viene Vacío	El número de socios tiene un valor inválido	Error
Fecha Seps	Validar si es tipo fecha	La fecha Seps de la organización tiene caracteres inválidos	Error
Código SEPS	El Código SEPS supera el tamaño de 50	El Código SEPS supera el tamaño de 50	Error
Fecha MAGAPs	Validar si es tipo fecha	La fecha MAGAPs de la organización tiene caracteres inválidos	Error
Código MAGAPs	El Código MAGAPS supera	El Código MAGAPS supera el	Error

	el tamaño de 50	tamaño de 50	
Localidad	La Localidad supera el tamaño de 250	La Localidad supera el tamaño de 250	Error
Dirección	La Dirección supera el tamaño de 500	La Dirección supera el tamaño de 500	Error
Cantón	Validar si el cantón está vacío o nulo	La organización no tiene cantón	Error
Provincia	Validar si la Provincia está vacío o nulo	La organización no tiene provincia	Error
Cantón	Validar si el cantón pertenece a la provincia	'El cantón '+[Canton]+' no pertenece a la provincia '+[Provincia]	Error

Nota: Reglas de validación implementado en la limpieza de las organizaciones
Elaborador por: Fausto Pardo

3.3.4.3. Reglas de validación de CIALCOs

Las Reglas de Validaciones de la estructura de CIALCOs se realizará en base a las especificaciones de la tabla 25.

Tabla 25. Reglas de validación de CIALCOS

Campo	Regla de Validación	Mensaje	Tipo de mensaje
Cantón	Validar si está vacío o nulo	El CIALCO no tiene cantón	Error
Provincia	Validar si está vacío o nulo	El CIALCO no tiene provincia	Error
Parroquia	Validar si está vacío o nulo	El CIALCO no tiene parroquia	Error
Nombre del	El Nombre del	El Nombre del	Error

CIALCO	CIALCO supera el tamaño de 350	CIALCO supera el tamaño de 350	
Nombre del CIALCO	Validar si está vacío	El CIALCO no tiene nombre	Error
Nombre del Representante	El Nombre del Representante supera el tamaño de 350	El Nombre del Representante supera el tamaño de 350	Error
Teléfono Fijo	El Teléfono Fijo supera el tamaño de 20	El Teléfono Fijo supera el tamaño de 20	Error
Teléfono Móvil	El Teléfono Móvil supera el tamaño de 20	El Teléfono Móvil supera el tamaño de 20	Error
Correo del Representante	El Correo del Representante supera el tamaño de 250	El Correo del Representante supera el tamaño de 250	Error
Dirección	La Dirección supera el tamaño de 500	La Dirección supera el tamaño de 500	Error
Observación	La Observación supera el tamaño de 350	La Observación supera el tamaño de 350	Error
Número de Mujeres	Validar si es numérico	El número de mujeres no es numérico	Error
Número de Hombres	Validar si es numérico	El número de Hombres no es numérico	Error
Anio	Validar si es numérico	El Año no es numérico	Error
Cantón	Validar si el cantón pertenece a la provincia	'El cantón '+[Canton]+' no pertenece a la provincia '+[Provincia]	Error
Parroquia	Validar si la Parroquia pertenece a la provincia	'La Parroquia '+[Parroquia]+' no pertenece al cantón '+[canton]	Error
Tipo de CIALCO	Validar si el productor está asignado a un CIALCO que es diferente a FERIA,CANASTA, COMPRA PUBLICA,PIE DE	El productor no tiene algún tipo de cialco (como feria, canasta..etc)	Error

	FINCA Y TIENDA		

Nota: Reglas de validación implementado en la limpieza de los CIALCOs
Elaborador por: Fausto Pardo

3.3.5. Mapeo de Datos(Fuente/ensayo/Destino)

3.3.5.1. Mapeo de Productores

El Mapeo de los datos de la estructura de productores se realizará en base a las especificaciones de la tabla 26.

Tabla 26. Mapeo de campos de productores

Destino		Origen	
Estructura	Campo	Estructura	Campo
persona_tbl	id_persona	AUTONUMERICO	
persona_tbl	nombrepersona	PRODUCTORES GPR Y SERVICIO WEB REGISTRO CIVIL	NOMBRES
persona_tbl	apellidopersona	PRODUCTORES GPR Y SERVICIO WEB REGISTRO CIVIL	APELLIDOS
persona_tbl	nombrecompleto	PRODUCTORES GPR Y SERVICIO WEB REGISTRO CIVIL	NOMBRES + APELLIDOS
persona_tbl	cedulapersona	PRODUCTORES GPR	CEDULA
persona_tbl	edadpersona	SERVICIO WEB REGISTRO CIVIL	FECHA DE NACIMIENTO
persona_tbl	fechanacimiento	SERVICIO WEB REGISTRO CIVIL	FECHA DE NACIMIENTO
persona_tbl	genero	SERVICIO WEB REGISTRO CIVIL	GENERO
persona_tbl	esproductor		VERDADERO
persona_tbl	estado		1 - ACTIVO
productor_tbl	id_productor		AUTONUMERICO
productor_tbl	id_persona	PRODUCTORES GPR	ID DE PRODUCTOR
productor_tbl	id_zona	PRODUCTORES GPR	ZONA
productor_tbl	id_provincia	PRODUCTORES GPR	PROVINCIA
productor_tbl	id_canton	SERVICIO WEB REGISTRO CIVIL	CANTON

productor_tbl	id_parroquia	SERVICIO WEB REGISTRO CIVIL	PARROQUIA
productor_tbl	direccion	SERVICIO WEB REGISTRO CIVIL	DIRECCION
productor_tbl	observacionfuentes	PRODUCTORES GPR	FUENTES
productor_tbl	estado		1 - ACTIVO
productorxcialco_tbl	id_productorxcialco		AUTONUMERICO
productorxcialco_tbl	id_productor	PRODUCTORES GPR	ID DE PRODUCTOR
productorxcialco_tbl	id_cialco	PRODUCTORES GPR	CIALCO
productorxcialco_tbl	fechainicioencialco	PRODUCTORES GPR	ANIO
productorxorganizacion_tbl	id_productororg		AUTONUMERICO
productorxorganizacion_tbl	id_productor	PRODUCTORES GPR	ID DE PRODUCTOR
productorxorganizacion_tbl	id_organizacion	PRODUCTORES GPR	ORGANIZACIÓN

Nota: Mapeo de los campos del origen y destino de productores
Elaborador por: Fausto Pardo

3.3.5.2. Mapeo de Organizaciones

El Mapeo de los datos de la estructura de organizaciones se realizará en base a las especificaciones de la tabla 27.

Tabla 27. Mapeo de campos de organizaciones

Destino		Origen	
Estructura	Campo	Estructura	Campo
organizacion_tbl	id_organizacion		AUTONUMERICO
organizacion_tbl	ruc character	Organización GPR o Registro de Organizaciones MAGAP	ruc
organizacion_tbl	nombreorganizacion	Organización GPR o Registro de Organizaciones MAGAP	nombre
organizacion_tbl	Nombrerepresen	Organización GPR	nombreRepresentante

	tante		
organizacion_tbl	Telefonofijo	Registro de Organizaciones MAGAP	telefono1
organizacion_tbl	Telefonomovil	Registro de Organizaciones MAGAP	celular1
organizacion_tbl	Correorepresentante	Registro de Organizaciones MAGAP	correo
organizacion_tbl	Nombrecontacto	Registro de Organizaciones MAGAP	propietario_nombres propietario_APELLIDOS
organizacion_tbl	Telefonocontacto	Registro de Organizaciones MAGAP	propietario_telefono1
organizacion_tbl	emailcontacto	Registro de Organizaciones MAGAP	propietario_correo
organizacion_tbl	numerodsocios	Registro de Organizaciones MAGAP	total_socios
organizacion_tbl	fechaseps	Organización GPR	fechaRegistroSEPS
organizacion_tbl	Codigoseps	Organización GPR	numRegistrosSEP
organizacion_tbl	Fechamagap	Organización GPR	fechaAcreditacion Magap
organizacion_tbl	Codigomagap	Organización GPR	numRegistrosMAGAP
organizacion_tbl	id_zona	Organización GPR	zona
organizacion_tbl	id_provincia	Organización GPR	provincia
organizacion_tbl	id_canton	Organización GPR	canton
organizacion_tbl	Localidad	Organización GPR	localidad
organizacion_tbl	Dirección	Registro de Organizaciones MAGAP	calle1,calle2, numero, lote
organizacion_tbl	Coordenadax	Registro de Organizaciones MAGAP	latitud
organizacion_tbl	Coordenaday	Registro de Organizaciones MAGAP	longitud
organizacion_tbl	estado		

Nota: Mapeo de los campos del origen y destino de organizaciones
Elaborador por: Fausto Pardo

3.3.5.3. Mapeo de CIALCOS

El Mapeo de los datos de la estructura de CIALCOs se realizará en base a las especificaciones de la tabla 28.

Tabla 28. Mapeo de campos de CIALCOS

Destino		Origen	
Estructura	Campo	Estructura	Campo
cialco_tbl	id_cialco		
cialco_tbl	id_catalogocialco	CIALCOGPR	MODALIDAD
cialco_tbl	nombrecialco	CIALCOGPR	DATOS CIRUCITO - NOMBRE
cialco_tbl	Nombrerepresentantecialco	CIALCOGPR	NOMBRE DEL REPRESENTANTE DEL CIRCUITO
cialco_tbl	telefonofijo	CIALCOGPR	INFORMACIÓN DE CONTACTO DEL REPRESENTANTE
cialco_tbl	telefonomovil	CIALCOGPR	INFORMACIÓN DE CONTACTO DEL REPRESENTANTE
cialco_tbl	Correo	CIALCOGPR	INFORMACIÓN DE CONTACTO DEL REPRESENTANTE
cialco_tbl	id_zona	CIALCOGPR	ZONA
cialco_tbl	id_provincia	CIALCOGPR	PROVINCIA
cialco_tbl	id_canton	CIALCOGPR	CANTON
cialco_tbl	id_parroquia	CIALCOGPR	PARROQUIA
cialco_tbl	direccion	CIALCOGPR	UBICACIÓN
cialco_tbl	anioinicioactividad	CIALCOGPR	ANIO
cialco_tbl	observacion	CIALCOGPR	OBSERVACION
cialcoproductor_tbl	numerohombres	CIALCO	NUM

		GPR	PRODUCTORES HOMBRES
cialcoproductor_tbl	numeromujeres	CIALCO GPR	NUM PRODUCTORES MUJERES

Nota: Mapeo de los campos del origen y destino de CIALCOs
Elaborador por: Fausto Pardo

3.3.6. Diseño del algoritmo de coincidencias

El algoritmo surge de la necesidad de encontrar correspondencia entre dos estructuras por medio del nombre de las Organizaciones, dicha búsqueda es importante ya que se requiere información de un lado que no se encuentra en el otro para poder insertar en la estructura del destino.

La comparación exacta solo cubre un 2% de la totalidad de organizaciones, por tal motivo tenemos que usar el algoritmo planteado para poder encontrar similitudes.

El algoritmo realiza los siguientes pasos para encontrar las similitudes:

Paso 1: Obtener las palabras del nombre por cada organización tanto del GPR y del registro de organizaciones. En la figura 11 se muestra las organizaciones con sus correspondientes palabras que serán usados en los siguientes pasos.

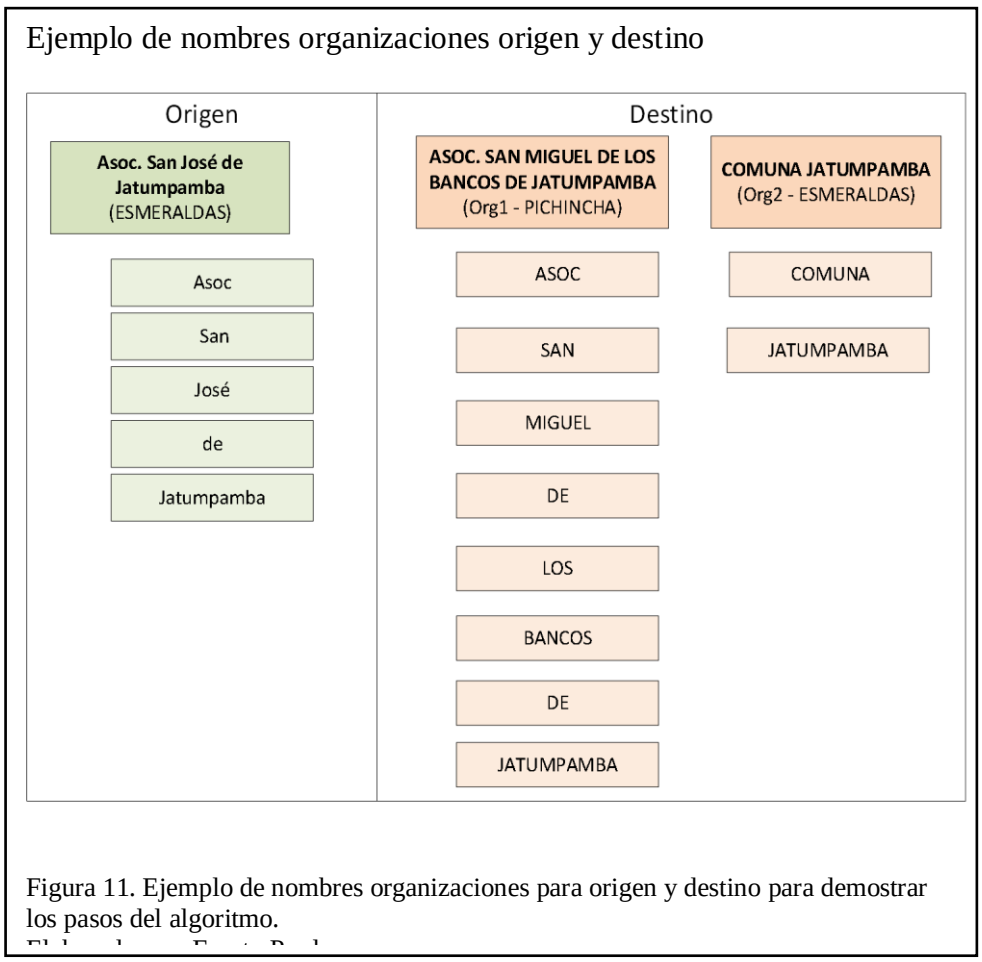


Figura 11. Ejemplo de nombres organizaciones para origen y destino para demostrar los pasos del algoritmo.

Paso 2: En la tabla 29 se muestra las palabras excluidas en el proceso de coincidencias, es decir, las palabras que no serán consideradas en la búsqueda de palabras del GPR y registro organizaciones.

Tabla 29. Palabras excluidas por el proceso de coincidencias

PALABRAS EXCLUIDAS			
DE	O	ASOC	GANADERO
DEL	EL	PRODUCCION	GANADEROS
LA	AL	PRODUCCIÓN	COOPERATIVA
LOS	ASOCIACION	PRODUCTORES	COMUNA
LAS	ASOCIACIÓN	AGROPECUARIO	SAN
Y	ASOC.	AGROPECUARIOS	SANTA

Nota: Palabras excluidas en el proceso de coincidencias, necesarias para mejorar el rendimiento
Elaborador por: Fausto Pardo

En la figura 12 se realiza una eliminación de las palabras que se excluyen en el proceso.

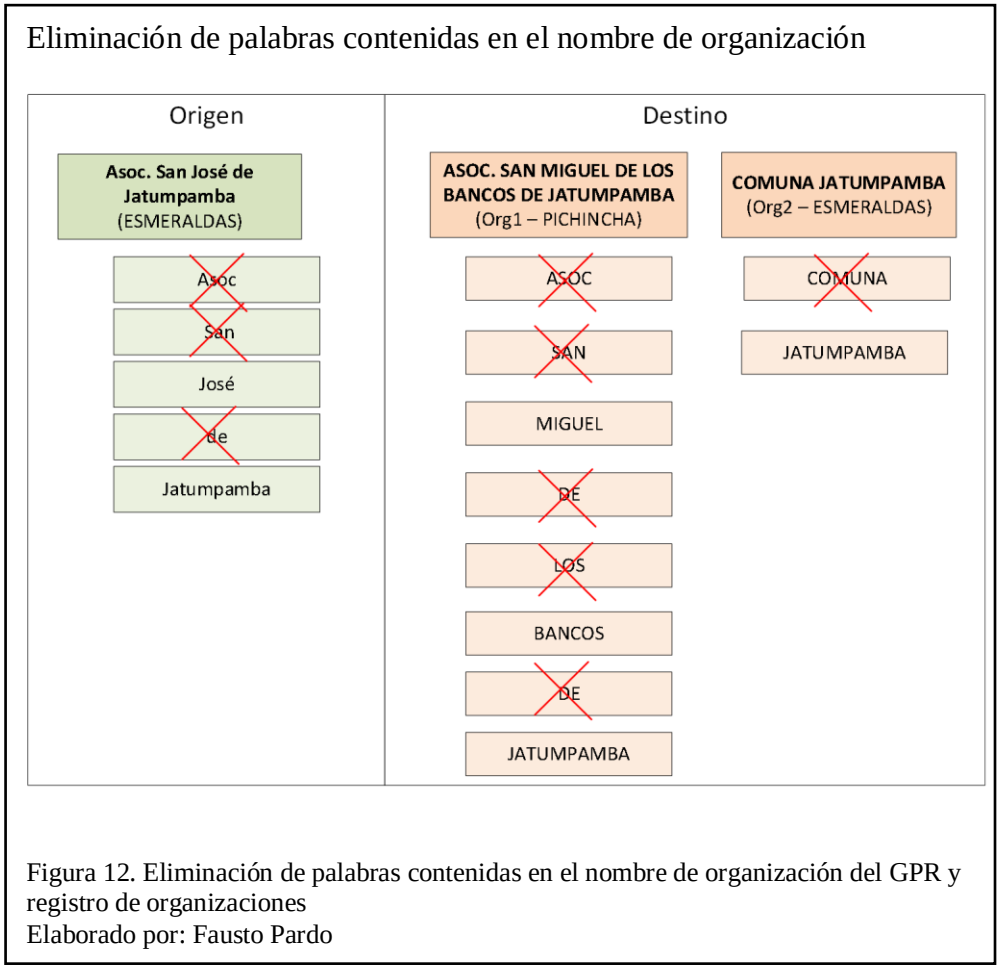


Figura 12. Eliminación de palabras contenidas en el nombre de organización del GPR y registro de organizaciones
Elaborado por: Fausto Pardo

Paso 3: Implementar el algoritmo de distancia de Levenshtein en cada una de las combinaciones. El algoritmo de distancia de Levenshtein en resumen lo que trata de hacer es saber cuántos cambios de caracteres se necesita para que una palabra sea igual otra. En la tabla 30 se muestra la distancia de las posibles combinaciones de las palabras del GPR y registro de organizaciones.

Tabla 30. Distancia de las combinaciones de palabras del GPR y reg. de org.

Ord.	Origen	Destino	Org. Destino	Distancia
C1	JOSE	MIGUEL	Org1	5
C2	JOSE	BANCOS	Org1	5
C3	JOSE	JATUMPAMBA	Org1	9
C4	JOSE	JATUMPAMBA	Org2	9
C5	JATUMPAMBA	MIGUEL	Org1	9
C6	JATUMPAMBA	BANCOS	Org1	9
C7	JATUMPAMBA	JATUMPAMBA	Org1	0

C8	JATUMPAMBA	JATUMPAMBA	Org2	0

Nota: Distancia de las combinaciones de palabras del GPR y reg. de org.
Elaborador por: Fausto Pardo

Paso 4: Agregamos nuevos cálculos como el tamaño del origen y también el cálculo del porcentaje de coincidencias. En la figura 13 se muestra el algoritmo de porcentaje de coincidencias.

Porcentaje de Coincidencias

```
SI [tamaño del origen]<[distancia] entonces
  % de coincidencias=  0%
Caso contrario,
  % de coincidencias=
  (([tamaño del origen]- [distancial)/ [tamaño del
origen])*100%
```

Figura 13. Porcentaje de coincidencias de las palabras del origen y destino
Elaborado por: Fausto Pardo

En la tabla 31 se muestra el porcentaje de coincidencias entre el GPR y el registro de organizaciones.

Tabla 31. Porcentaje de coincidencias de palabras entre el GPR y el registro de organizaciones.

Origen	Destino	Org. Destino	Distancia	Tamaño del Origen	Porcentaje de Coincidencia
JOSE	MIGUEL	Org1	5	4	0%
JOSE	BANCOS	Org1	5	4	0%
JOSE	JATUMPAMBA	Org1	9	4	0%
JOSE	JATUMPAMBA	Org2	9	4	0%
JATUMPAMBA	MIGUEL	Org1	9	10	10%
JATUMPAMBA	BANCOS	Org1	9	10	10%
JATUMPAMBA	JATUMPAMBA	Org1	0	10	100%
JATUMPAMBA	JATUMPAMBA	Org2	0	10	100%

Nota: Porcentaje de coincidencias de palabras entre el GPR y el registro de organizaciones.
Elaborador por: Fausto Pardo

Paso 5: Eliminamos los registros que tienen un porcentaje de coincidencia menos del 25%. En la tabla 32 se muestra un ejemplo.

Tabla 32. Eliminación de registros que tienen un % de coincidencias menor al 25%

Origen	Destino	Org. Destino	Porcentaje de Coincidencia
JOSE	MIGUEL	Org1	0%
JOSE	BANCOS	Org1	0%
JOSE	JATUMPAMBA	Org1	0%
JOSE	JATUMPAMBA	Org2	0%
JATUMPAMBA	MIGUEL	Org1	10%
JATUMPAMBA	BANCOS	Org1	10%
JATUMPAMBA	JATUMPAMBA	Org1	100%
JATUMPAMBA	JATUMPAMBA	Org2	100%

Nota: Eliminación de registros que tienen un porcentaje de coincidencias menor al 25%
Elaborador por: Fausto Pardo

Paso 6: Realizamos un agrupamiento por el nombre de la organización GPR y el conteo de registro para obtener el porcentaje por medio de un cálculo.

La Organización Origen (GPR) tiene 2 palabras (JOSE y JATUMPAMBA)

La Org1 de Destino (registro de organizaciones) tiene 1 palabra con 100% de coincidencia

La Org2 de Destino (registro de organizaciones) tiene 1 palabra con 100% de coincidencia

$$\%\ de\ Concidencia = \frac{Cantidad\ de\ Palabras\ OK}{Cantidad\ de\ Palabras\ Organización\ del\ Origen\ (GPR)}$$

En la tabla 33 se muestra un agrupamiento por el nombre de la organización GPR y el conteo de registro para obtener el porcentaje por medio de un cálculo.

Tabla 33. Porcentaje de coincidencias entre los nombres de organizaciones del GPR y registro de organizaciones.

Org. Origen	Org. Destino	Provincia	Cant. De Palabras OK	Cant. De Palabras del Origen (GPR)	Porcentaje
JOSE J.	Org1	PICHINCHA	1	2	50%
JOSE J.	Org2	ESMERALDAS	1	2	50%

Nota: Porcentaje de coincidencias entre los nombres de organizaciones del GPR y registro de organizaciones.
Elaborador por Fausto Pardo

Paso 7: Eliminamos las organizaciones que no corresponden a las Provincia de la organización Origen. En la tabla 34 se muestra la eliminación del registro que contiene la provincia PICHINCHA.

Tabla 34. Eliminación de los registros que no corresponden a la provincia.

Org. Destino	Provincia. Destino	Cant. De Palabras OK	Cant. De Palabras de la Organización Origen	Porcentaje
Org1	PICHINCHA	1	2	50%
Org2	ESMERALDAS	1	2	50%

Nota: Eliminación de los registros que no corresponden a la provincia.
Elaborador por: Fausto Pardo

Paso 8: Se almacena en una estructura de coincidencias las organizaciones origen (GPR) y su organización destino (registro de organizaciones) que pasaron todos los filtros. En la tabla 35 se muestra los registros OK después de aplicar el algoritmo de coincidencias.

Tabla 35. Registros OK después de aplicar el algoritmo de coincidencias.

Org. Origen	Org. Destino	Porcentaje
Asoc. San José de Jatumpamba	COMUNA JATUMPAMBA (Org2)	50%

Nota: Registros OK después de aplicar el algoritmo de coincidencias.
Elaborador por: Fausto Pardo

Paso 9: Se valida con el cliente cuál de las combinaciones es correcta para poder insertar en la estructura de homologaciones.

3.3.7. Configuración de la Migración

3.3.7.1. Rutas de carpetas y archivos

Para poder realizar la carga y salidas de los archivos de la migración es importante crear las carpetas que se muestran en la tabla 36 dentro de la ruta que crea conveniente.

Tabla 36. Rutas de carpetas de carga y salida

Rutas de carpetas	
Carpetas	Descripción
.\ Archivos de Carga	Carpeta de los archivos de carga
.\ Archivos de Carga\Productores	Archivos de carga de productores
.\ Archivos de Carga\Organizaciones	Archivos de carga de organizaciones
.\ Archivos de Carga\Redes	Archivos de carga de redes
.\ Archivos de Carga\CIALCOs	Archivos de carga de CIALCOs
.\ Archivos de Salida	Archivos de salida
.\ Archivos de Salida\Logs	Archivos de salida de logs
.\ Archivos de Salida\Scripts	Archivos de salida de los scripts de generación para ejecutar en postgreS

Nota: Rutas de carpetas de carga y salida de la solución
Elaborador por: Fausto Pardo

Los archivos que se deben depositar en cada una de las carpetas de carga de productores, organizaciones y CIALCOs son del tipo txt., y la estructura es igual al archivo consolidado xls. En la tabla 37 se muestran los nombres de los archivos de carga.

Tabla 37. Rutas de archivos

Archivos	
Carpetas	Archivos
.\ Archivos de Carga\Productores	Consolidado GPR 2015_PRODUCTORES.txt
.\ Archivos de Carga\Productores	Consolidado GPR 2016_PRODUCTORES.txt

.\ Archivos de Carga\Organizaciones	Consolidado GPR 2015_ORGANIZACIONES.txt
.\ Archivos de Carga\Organizaciones	Consolidado GPR 2016_ORGANIZACIONES.txt
.\ Archivos de Carga\Redes	Consolidado GPR 2015_REDES.txt
.\ Archivos de Carga\Redes	Consolidado GPR 2016_REDES.txt
.\ Archivos de Carga\CIALCOs	Consolidado GPR 2015_CIALCOS.txt
.\ Archivos de Carga\CIALCOs	Consolidado GPR 2016_CIALCOS.txt

Nota: Rutas de archivos para la carga
Elaborador por: Fausto Pardo

3.3.7.2. Configuración del proyecto de paquetes ETL

Para poder ejecutar los paquetes ETL es necesario realizar la configuración del proyecto de ETLs en base la tabla 38.

Tabla 38. Configuración del proyecto de ETLs

Parámetros de configuración	
Parámetro	Descripción
IdUsuarioMigracion	Id del usuario de migración de la base de PostgreSQL
IdZonaUsuario	Id Zona del usuario de migración de la base de PostgreSQL por lo general es 1
PathArchivosCarga	Ruta de los archivos de Entrada. .\ Archivos de Carga
PathArchivosSalida	Ruta de los archivos de salida. .\ Archivos de Salida
UltimoIdCIALCO	Id máximo de la estructura cialco_tbl PostgreSQL
UltimoIdOrg	Id máximo de la estructura organizacion_tbl PostgreSQL
UltimoIdPersona	Id máximo de la estructura persona_tbl PostgreSQL
UltimoIdProductor	Id máximo de la estructura productor_tbl PostgreSQL
UltimoIdProductorXCialco	Id máximo de la estructura productorxcialco_tbl PostgreSQL
UltimoIdProductorXOrg	Id máximo de la estructura productorxorganizacion_tbl PostgreSQL

Nota: Configuración del proyecto de ETLs necesarios para iniciar el proceso.
Elaborador por: Fausto Pardo

3.3.7.3. Mapeo de Homologaciones

La estructura de homologaciones nos permite registrar equivalencias entre dos estructuras, es muy usado en las transformaciones cuando el origen viene un dato que tiene que ser otro.

En la tabla 39 se muestra las homologaciones que se registran en esta estructura (MAPEO_HOMOLOGACIONES).

Tabla 39. Mapeo de homologaciones

Homologaciones		
Código	Tabla Origen	Tabla Destino
ORG001	LISTA_ORGANIZACIONES	ORGANIZACIONES 20152016
ORG002	LISTA_ORGANIZACIONES	PROVINCIA_TBL
ORG003	LISTA_ORGANIZACIONES	CANTON_TBL
CIA001	CONSOLIDADO_GPR	PROVINCIA_TBL
CIA002	CONSOLIDADO_GPR	CANTON_TBL
CIA003	CONSOLIDADO_GPR	CANTON_TBL

Nota: Mapeo de homologaciones de las listas de organizaciones
Elaborador por: Fausto Pardo

3.3.7.4. Log de los Errores

Para poder auditar los sucesos que están sucediendo internamente dentro de los procesos, es indispensable crear una estructura genérica que pueda servir para poder registrar eventos como errores y advertencias de las validaciones que se realiza.

En la tabla 40 se muestra la estructura de la tabla se llama **monitero.Log**.

Tabla 40. Estructura de LOG

Campo	Tipo	Descripción
Id	Numérico	Id interno
IdRegistro	Numérico	Id del Registro del registro que se está analizando
Mensaje	Cadena	Mensaje del error/advertencia
Valor	Cadena	Valor del campo analizado
Estado	Numérico	2 – Error 1 – Advertencia
SubCategoria	Cadena	Subcategoria del campo

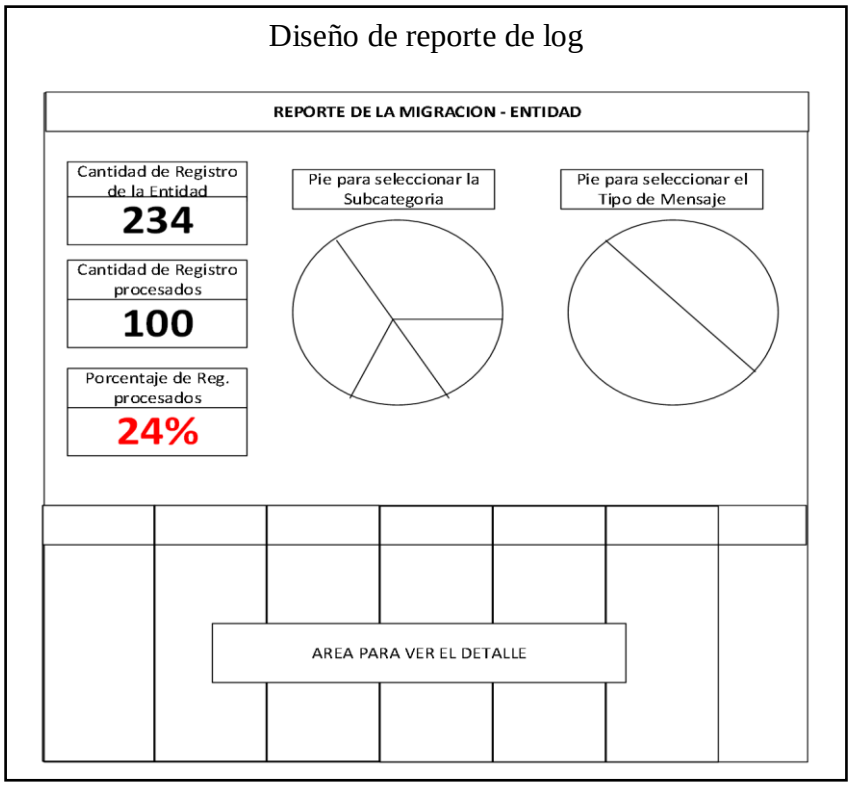
		analizado
Categoria	Cadena	Categoría del campo analizado
Funcionalidad	Cadena	Organización, Productores o CIALCOS
Fecha de Registro	Date and Time	Fecha del Registro

Nota: Estructura del archivo de log cuando se realiza validaciones.
Elaborador por: Fausto Pardo

3.3.8. Diseño de los Reportes

Todos los reportes de indicadores tendrán la misma estructura donde se bosqueja en la figura 14.

Diseño de reporte de log



Capítulo 4

Implementación y Pruebas

4.1. Introducción

En esta fase se dará conocer la implementación y pruebas realizadas en la migración de los datos en base al diseño y requerimientos mencionados en los capítulos 2 y 3.

Las pruebas se han desarrollado en base al "Plan de Pruebas de la Migración de datos" especificado en el capítulo 2 de planificación.

4.2. Desarrollo de las Pruebas

Las pruebas de la migración de datos serán validadas a través de las métricas de datos procesados y erróneos. La visualización de las métricas será a través del tablero de control de log de errores que fue realizado de forma independiente para productores, organizaciones y CIALCOs pero con las mismas medidas.

En las siguientes sub-secciones se mostrará las diferentes pruebas realizadas a cada una de las entidades que son productores, organizaciones y CIALCOs. En resumen, existe un alto porcentaje de registros que no pasaron las reglas de validación sobre todo productores y organizaciones, esto se debe a la mala calidad de los datos como se evidencia en la sección de "Perfilado de Datos del capítulo 3 de diseño y análisis". En CIALCOs no existe muchas novedades, ya que tiene un gran porcentaje de registros procesados satisfactoriamente.

4.2.1. Pruebas en la estructura de Productores

En productores se logró procesar satisfactoriamente un 28.55% de registros, existiendo un gran porcentaje de registros que no pasaron la validación de datos porque la mayoría de los registros tienen una mala calidad de datos. Por ejemplo, la Cédula (campo clave único) tiene 500 registros que tiene errónea la cédula, ya sea porque su tamaño no es adecuado o incumple la regla de validación con el servicio del MAGAP.

En la tabla 41 se muestra los diferentes motivos porque gran porcentaje de los registros no fueron procesados.

Tabla 41. Reporte de pruebas de productores

Productores			
Cantidad de Registros Total	Cantidad de Registros Procesados	Porcentaje de Registros Procesados	Cantidad de Registro erróneos
4154	1186	28.55%	**2968**
Sub Categoría	**Cant. Registros erróneos**	**Motivo**	
Cédula	500	La cédula es inválida, se debe a que la cédula incumple alguna regla de validación.	
Edad	435	No tiene la edad, porque no tiene cédula.	
Genero	435	No tiene Genero, porque no tiene cédula.	
Apellido	142	No tiene apellido, porque no tiene cédula.	
Tipo de CIALCO y Nombre	892	No existe relación con la estructura de CIALCO por medio del nombre y tipo	
Tipo de CIALCO	529	El productor no es tipo de CIALCO que se está procesando. TIENDA, COMPRA PUBLICA, PIE DE FINCA, CANASTA o FERIA	
Tipo de CIALCO Vació	363	No tiene ningún tipo de CIALCO asignado	
Organización Vació	644	Registros que tiene vacío su organización, esto es normal ya que	

		existe productores que aplican a un CIALCO de forma directa.
Organización Homologado	2751	Falta por homologar las organizaciones, pero si tiene asignado un valor

Nota: Reporte de pruebas de productores
Elaborador por: Fausto Pardo

4.2.2. Pruebas en la estructura de Organizaciones

En organizaciones se logró procesar satisfactoriamente un 19.40% de registros, existiendo un gran porcentaje de registros que no pasaron la validación de datos porque la mayoría de los registros tienen una mala calidad de datos. Por ejemplo, el Nombre de la Organización falta homologar contra el registro de organizaciones para obtener los datos faltantes del registro.

En la tabla 42 se muestra los diferentes motivos porque gran porcentaje de los registros no fueron procesados.

Tabla 42. Reporte de pruebas de organizaciones

Organizaciones			
Cantidad de Registros Total	Cantidad de Registros Procesados	Porcentaje de Registros Procesados	Cantidad de Registro erróneos
634	123	19.40%	**511**
Sub Categoría	Cant. Registros erróneos	Motivo	
Nombre de la Organizaciones	444	Nombre tiene homologación con el registro de organizaciones	
Cantón	15	Registro que no tienen cantón	
RUC	39	Existe 39 registros que no tienen RUC	
Provincia	7	Registros que no tiene Provincia	
Fecha Magap	21	Registros que tienen fechas inválidas	

Fecha SEPS	28	Registros que tienen fechas inválidas

Nota: Reporte de pruebas realizadas en organizaciones
Elaborador por: Fausto Pardo

4.2.3. Pruebas en la estructura de CIALCO

En CIALCOs se logró procesar satisfactoriamente un 75% de registros, existiendo un gran porcentaje de registros satisfactorios porque la calidad de la información es mejor.

En la tabla 43 se muestra los diferentes motivos porque algunos registros no pasaron las reglas de validación.

Tabla 43. Reporte de pruebas de CIALCOs

CIALCO			
Cantidad de Registros Total	Cantidad de Registros Procesados	Porcentaje de Registros Procesados	Cantidad de Registro erróneos
283	220	75%	**63**

Sub Categoría	Cant. Registros erróneos	Motivo
Tipo de CIALCO	53	El productor no es del tipo de CIALCO que se está procesando. TIENDA, COMPRA PUBLICA, PIE DE FINCA, CANASTA o FERIA
Provincias Vació	2	Registros con provincias vacías
Cantón	2	Registros que tienen cantones que no pertenecen a la provincia
RUC	39	Existe 39 registros que no tienen RUC
Nombre de CIALCO vació	1	Registros con nombre de CIALCO vació

Nota: Reporte de pruebas realizadas en CIALCOs

Elaborador por: Fausto Pardo

Capítulo 5

Cierre

5.1. Introducción

En esta fase se desarrolla las recomendaciones, conclusiones, transferencia de conocimiento y los resultados generales del proceso de migración.

5.2. Resultados

En productores y organizaciones existe gran porcentaje de registros que no pasaron las validaciones debido a la mala calidad de los datos, mientras que el CIALCOs el panorama es muy diferente, ahí se logró procesar satisfactoriamente un gran porcentaje de registros. Para poder tener una claridad un poco más al detalle de los registros que no fueron procesados satisfactoriamente hay que dirigirse al capítulo 4 - implementación y prueba.

En la tabla 44 se muestran los porcentajes de registros que han sido procesados satisfactoriamente.

Tabla 44. Resultados de la migración porcentajes por entidad

Entidad	Porcentaje
Productores	28.55%
Organizaciones	19.40%
CIALCOs	75%

Nota: Resultados de la migración entidad con sus porcentajes
Elaborador por: Fausto Pardo

En las siguientes secciones se abordará los resultados obtenidos por cada entidad

5.2.1. Resultado de productores

En productores de 4.154 registros se logró procesar satisfactoriamente 1.186 registros que corresponde al 28.55%, existe un gran porcentaje de registros que no pasaron la validación de datos debido a que la información de productores es de mala calidad. En la figura 14 se muestra con gráfico circular donde cada parte corresponde a un campo de la estructura de productores que fue validado por medio de las reglas definidas en el capítulo 3 de diseño y análisis.

Como se evidencia en la figura 15, gran porcentaje de los registros erróneos corresponde a la validación de la homologación de organización, esto se debe a que todavía la organización del productor no ha sido homologada contra el registro de organizaciones.

En que lo respecta a la Cédula, Edad, Apellido y Genero, existe registros erróneos debido a que estos campos provienen del servicio WEB del MAGAP por medio de la cédula, y como la cédula no es válida entonces los campos descritos en este párrafo también heredan el error.

En que lo respecta al Tipo y Nombre de CIALCO, existe registros erróneos debido a que estos campos no tiene correspondencia con la estructura de CIALCOs.

Reporte de resultados de productores

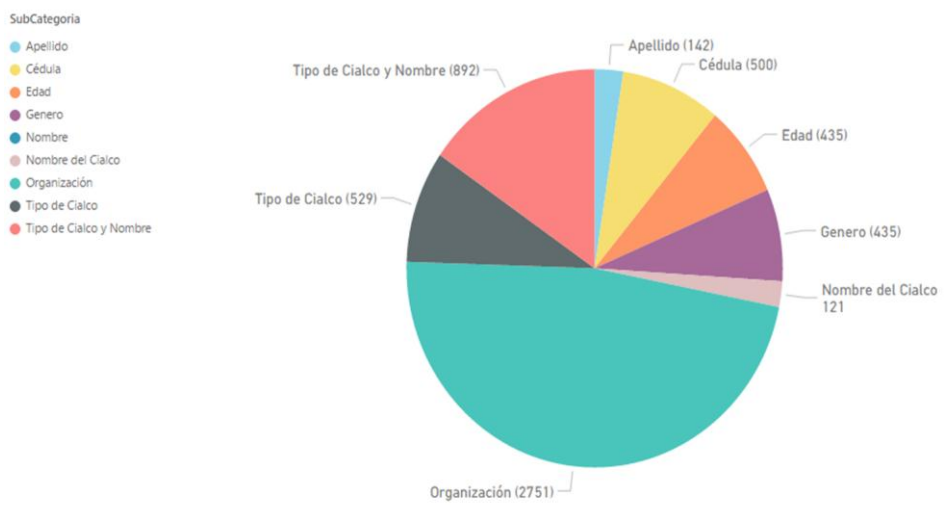


Figura 15. Reporte de resultados de la migración de los productores
Elaborado por: Fausto Pardo

5.2.2. Resultado de organizaciones

En organizaciones de 634 registros se logró procesar satisfactoriamente 123 registros que corresponde al 19.40%, existe un gran porcentaje de registros que no pasaron la validación de datos debido a que la información de organizaciones es de mala calidad. En la figura 15 se muestra con gráfico circular donde cada parte corresponde a un campo de la estructura de organizaciones que fue validado por medio de las reglas definidas en el capítulo 3 de diseño y análisis.

Como se evidencia en la figura 16, gran porcentaje de los registros erróneos corresponde a la validación del nombre de la organización que todavía no ha sido homologado en el registro de organizaciones, esto se debe a que todavía el nombre de la organización no ha sido homologado contra el registro de organizaciones.

En que lo respecta al RUC, Cantón y Provincia, existe registros erróneos debido a que el campo viene vacío o es invalido.

En que lo respecta a la FechaSeps y FechaMagap, existe registros erróneos debido a que el campo es invalido.

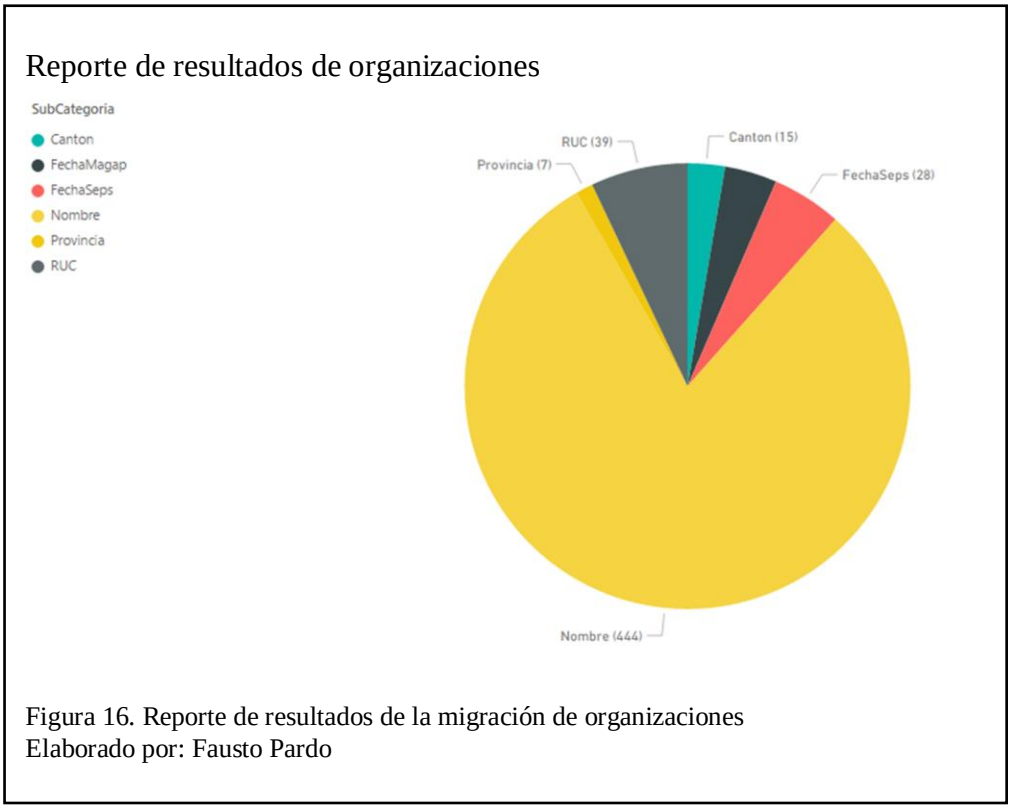


Figura 16. Reporte de resultados de la migración de organizaciones
Elaborado por: Fausto Pardo

5.2.3. Resultado de CIALCOs

En CIALCOs de 283 registros se logró procesar satisfactoriamente 220 registros que corresponde al 77.74%, existe un pequeño porcentaje de registros que no pasaron la validación. En la figura 16 se muestra con gráfico circular donde cada parte corresponde a un campo de la estructura de organizaciones que fue validado por medio de las reglas definidas en el capítulo 3 de diseño y análisis.

Como se evidencia en la figura 17, gran porcentaje de registros corresponde a la validación de la parroquia, esto se debe a que la parroquia tiene registros vacíos. Cabe mencionar que los registros pasaron la validación ya que corresponden a una advertencia.

En que lo respecta al Cantón y Provincia, existe registros erróneos debido a que el campo viene vacíos.

En que lo respecta al Tipo y Nombre de CIALCO, existe registros erróneos debido a que estos campos no tiene un valor o no corresponde a los tipos de CIALCOs que se están procesando.

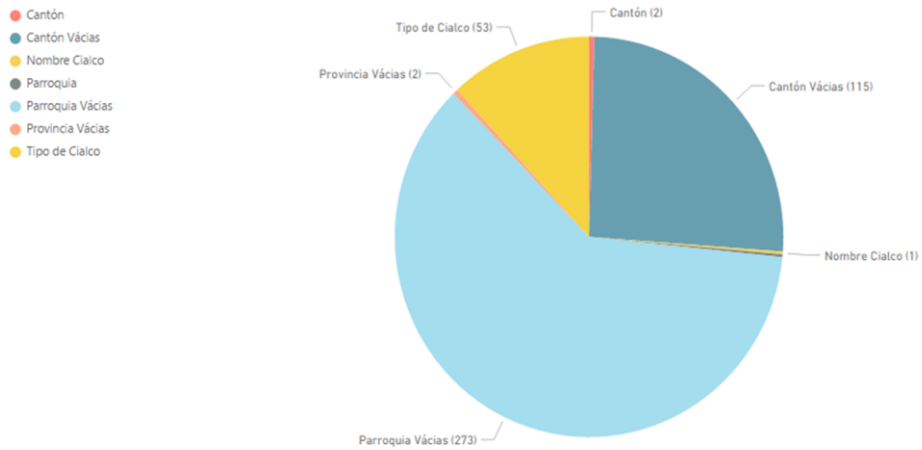


Figura 17. Reporte de resultados de la migración de CIALCOs
Elaborado por: Fausto Pardo

CONCLUSIONES

- Se ha cumplido con el objetivo general del proyecto realizando la migración de los archivos Excel hacia el sistema de la coordinación general de redes comerciales.
- El algoritmo de Levenshtein es una excelente técnica para encontrar la cantidad de operaciones necesarias para que una cadena se parezca a otra, la misma nos ayudado con la homologación de las organizaciones y el registro de organizaciones.
- La metodología de migración implementada es un conjunto de buenas prácticas para este tipo de proyectos, la misma ayudo a cubrir todas las necesidades que se han presentado a lo largo del proyecto.
- Saber identificar los riesgos y mitigarlos tempranamente es buena práctica en proyectos donde la información a migrar tiene una mala calidad.
- El uso de herramientas gratuitas como es SQL Server Express ayudo en proyecto para realizar actividades de limpieza y calidad de datos.

RECOMENDACIONES

- Si se requiere sistematizar un proceso de migración una excelente alternativa es la metodología propuesta, la misma abarca todas las etapas para un proyecto exitoso.
- Se recomienda usar estructura de homologaciones para realizar actividades de transformaciones.
- Para proyectos muy cambiantes y con gran incertidumbre es recomendable usar una metodología de proyectos agiles como es SCRUM
- Definir bien las actividades y responsabilidades con todas las personas que intervienen en la migración, con esto se evita confrontaciones y malos entendidos.

LISTA DE REFERENCIAS

Department of Education Office (2007). *Data Migration Roadmap*. Washington D. C., United States of America: Federal Student Aid. Retrieved from https://studentaid.ed.gov/sa/sites/default/files/fsawg/static/gw/docs/ciolibrary/ECONOPS_Docs/DataMigrationRoadmap.pdf

Kimball, R. (2004). *The Data Warehouse ETL Toolkit*. Indianápolis, United States of America: Wiley.

Kimball, R. (2010). *The Data WareHouse Toolkit*. Indianápolis, United States of America: Wiley.

Levenshtein, V. (1966). *Binary codes capable of correcting deletions, insertions, and reversals*. Rusia: Soviet Physics Doklady.

IBM (2006). *Best Practices for Data Migration (White Paper)*. New York, United States of America: Softek

IBM (2006). *Simplifying Technology Refresh with Data Migration Software (White Paper)*. New York, United States of America: Softek

Rainardi, V. (2007). *Building a Data Warehouse: With Examples in SQL Server*. New York, United States of America: Apress

Ponniah, P. (2001). *Data Warehousing Fundamentals*. Indianápolis, United States of America: Wiley

GLOSARIO DE TÉRMINOS

MAGAP: Ministerio de Agricultura, Ganadería, Acuacultura y Pesca

CGRC: Coordinación General de Redes Comerciales del MAGAP.

CIALCO: Circuitos Alternativos de Comercialización