ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE INFORMÁTICA Y ELECTRÓNICA

ESCUELA DE INGENIERÍA EN SISTEMAS

"ANÁLISIS COMPARATIVO DE HERRAMIENTAS DATA QUALITY PROPIETARIO FRENTE A LAS HERRAMIENTAS DE SOFTWARE LIBRE DISPONIBLES EN EL MERCADO. APLICADO A LA BASE DE DATOS OASIS"

TESIS DE GRADO

Previo a la obtención del título de:

INGENIERO EN SISTEMAS INFORMÁTICOS

Presentado por:

CARLOS JAVIER MEDINA BENALCÁZAR

Riobamba – Ecuador

2014

AGRADECIMIENTO

Expreso un agradecimiento muy especial a la Escuela Superior Politécnica de Chimborazo especialmente a la Facultad de Ingeniería en Sistemas, por su responsabilidad y alto nivel académico desarrollado en los años de estudios, de la misma manera mi gratitud a todos y todas las personas que directa e indirectamente han contribuido para desarrollar este trabajo.

DEDICATORIA

Mi tesis se la dedico con todo amor y cariño principalmente a mis padres que me dieron la vida y han estado con migo en todo momento. Gracias por todo papá y mamá por darme una carrera para mi futuro y por creer en mí, aunque hemos pasado momentos difíciles siempre han estado apoyándome y brindándome todo su amor, por todo esto les agradezco de todo corazón de que estén conmigo.

FIRMAS DE RESPONSABLES

NOMBRE	**FIRMA**	**FECHA**
Ing. Iván Menes **DECANO DE LA FACULTAD DE** **INFORMÁTICA Y ELECTRÓNICA**	…….	…….
Ing. Jorge Huilca **DIRECTOR DE LAS ESCUELA DE** **INGENIERÍA EN SISTEMAS**	…….	…….
Ing. Paúl Paguay **DIRECTOR DE TESIS**	…….	…….
Ing. Patricio Moreno **MIEMBRO DEL TRIBUNAL**	…….	…….
DIRECTOR DEL CENTRO **DE DOCUMENTACIÓN**	…….	…….

NOTA DE LA TESIS…….

"Yo, Carlos Javier Medina Benalcázar, soy responsable de las ideas, doctrinas y resultados expuestos en esta tesis, y el patrimonio intelectual de la misma pertenece a la Escuela Superior Politécnica De Chimborazo".

Carlos Javier Medina Benalcázar

ÍNDICE DE ABREVIATURAS

BI: Business Intelligence.

CRM: Customer Ralationship Management.

DBA: Data Base Administrator.

DQ: Data Quality.

DSS: Sistema de soporte de decisiones.

DWH: Datawarehouse.

ERP: Enterprise Resource Planning.

ESPOCH: Escuela Superior Politécnica Chimborazo.

ETL: Extracción, Transformación, Carga.

JDBC: Java Database Connectivity.

MDM: Master Data Management.

ODS: Object Data Store.

SCM: Software Configuration Management

SII: Sistema de Información Institucional.

SOA: Arquitectura Orientada a Servicios.

XML: Extensible Markup Language.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

La calidad de datos en muchas empresas se ha convertido en el activo más importante que estas poseen es por esta razón se han diseñado muchas herramientas para analizar, limpiar y mejorar estos datos dándole a las empresas datos de relevancia y confiables para la toma de decisiones.

Dentro del ambiente de calidad de datos existen muchas herramientas que nos brindan una gran ayuda para mejorar una o varias dimensiones de calidad con el fin de obtener datos más exactos y de confianza para todos los niveles de la empresa sea está a nivel operacional o de toma de decisiones.

En esta tesis se presentan los resultados de la investigación realizada para un análisis comparativo de las herramientas Data Quality propietario frente a las herramientas de software libre disponibles en el mercado.

El objetivo principal de esta tesis es realizar un análisis comparativo sobre las herramientas Data Quality, para garantizar la calidad de datos en la base de datos OASIS.

En adelante se mostrará el resultado del análisis comparativo de las herramientas, además del análisis de las dimensiones de calidad para garantizar la calidad de datos dentro de un almacén de datos.

Al analizar los indicadores a utilizar se mostrará de manera detallada los parámetros a utilizar para lograr la demostración de la hipótesis planteada.

En el desarrollo del proyecto de calidad se analizarán las dimensiones de calidad como la integridad, la consistencia y la confiabilidad. La metodología empleada para el desarrollo del proyecto de calidad de datos para la base de datos OASIS es SII-ESPOCH. La metodología se encuentra detallada en la documentación entregada en los anexos.

El primer capítulo de esta tesis trata sobre el marco referencial en el cual se encuentra de manera general la justificación del proyecto de tesis además de los objetivos a alcanzar con el desarrollo de la tesis.

En el segundo capítulo se describe los conceptos básicos de calidad de datos y las herramientas que se van a analizar.

Para el tercer capítulo se ha previsto analizar cada una de las herramientas elegidas planteando un escenario de pruebas para conocer el beneficio que nos presenta cada una de estas herramientas y determinar las ventajas que ostentan cada una de ellas.

El desarrollo del proyecto de calidad de datos para la base de datos OASIS se presenta en el capítulo cuarto, el análisis y aplicación de la metodología se resumen en este capítulo así como la lógica de la aplicación, además de en este capítulo se detallará los parámetros utilizados para la comprobación de la hipótesis planteada.

CAPÍTULO I

MARCO REFERENCIAL

1.1. Antecedentes

Debido a la incorporación del software de fortalecimiento de las dimensiones de calidad de datos en los sistemas informáticos algunas de las labores en las empresas han evitado serios problemas como en los envíos publicitarios a miles de direcciones erróneas o duplicadas, oportunidades de negocio desaprovechadas, toma de decisiones equivocadas porque los datos no estaban disponibles a tiempo o eran incorrectos. La calidad de los datos debería tenerse en cuenta por cualquier empresa que desee ahorrar dinero, multiplicar las respuestas a sus propuestas, evitar los malos entendidos y causar una excelente impresión a sus clientes actuales y futuros.

Los errores en los datos tienen que ver con la falta de flexibilidad de algunos sistemas informáticos y con las propias equivocaciones y negligencias humanas. [1]

[1] Tomada de: http://www.marketingcomunidad.com/el-valor-de-la-calidad-de-los-datos-en-marketing

Lo que habría que hacer es asumir precisamente que la falta de calidad de los datos ha de atacarse desde su raíz, esto es, desde los procesos de entrada y carga de datos hasta los procesos de salida.

Existen varios estándares de calidad tal como ISO / IEC 25012:2008 el cual se puede utilizar para establecer los requisitos, medidas de calidad de datos, o el plan y llevar a cabo evaluaciones de calidad de datos. Se podría utilizar, para definir y evaluar las necesidades de datos de calidad en los procesos de producción, la adquisición y la integración, para identificar los criterios de garantía de calidad, también es útil para la re-estructuración, la evaluación y mejora , para evaluar la conformidad de los datos con legislación y / o requisitos (1).

En la ESPOCH se ha realizado estudios de calidad de datos para el desarrollo de una propuesta metodológica en la gestión de la calidad de datos en proyectos de integración para ser aplicado en el SII-ESPOCH (Sistema de Información Institucional de la Escuela Superior Politécnica De Chimborazo).

La propuesta consiste en cumplir siete fases: Estudio y preparación del proyecto, Análisis de la información, Evaluación y análisis de los datos, Limpieza de datos, Evaluación y análisis final, Mejoramiento y prevención y Control de la calidad.

Debido a que se cuenta con la metodología para el desarrollo del sistema pero los datos de las fuentes como el “OASIS” donde se debe conocer su estructura, se debe tomar medidas de seguridad y confiabilidad de la fuente por esta razón se debe analizar los datos mediante la utilización de una herramienta capaz de limpiar y consolidar esta información la cual es relevante para alcanzar el fin requerido.

1.2. JUSTIFICACIÓN DEL PROYECTO DE TESIS

Se lo realizó en función a una justificación Teórica, Metodológica y Práctica.

1.2.1. Justificación Teórica.

La necesidad de calidad y gestión de datos en los almacenes de datos es uno de los factores clave de éxito en cualquier proyecto del mundo real, de los datos se puede conocer que son imperfectos, víctimas de diversas formas de defectos tales como la variabilidad del sensor, los errores de estimación, la incertidumbre, los errores humanos en los datos de entrada, etc.

Comprender la existencia del problema, identificarlo plenamente, diseñar un plan de trabajo y dedicar los recursos necesarios, tanto para su corrección inicial como para la corrección continuada a lo largo del tiempo. Por lo tanto, primero es asumir que la calidad de los datos es un asunto que hay que incorporar en la agenda empresarial, y situarlo en el mismo nivel que cualquier otro servicio crítico.

Además que una buena práctica en la utilización de las herramientas para calidad de datos nos lleva a obtener grandes ventajas como:

- Incremento de la productividad de TI.
- Ahorro de costes y recursos eliminando los registros duplicados.
- Aumento de su competitividad incorporando sólo información relevante y necesaria a sus almacenes de datos.

1.2.2. Justificación Metodológica

Para el presente documento, se adoptará el método científico, mediante el cual se detallarán situaciones y eventos, midiendo y evaluando diversos marcos de trabajo para el desarrollo de la presente investigación, la cual guiará en el proceso de implementación de las mismas.

La aplicación del proceso metodológico utilizará varias técnicas como la observación, la experimentación y el análisis de cada una de las herramientas a ser evaluadas.

1.2.3. Justificación Práctica

Debido a la necesidad de lograr un conjunto de datos con alta consistencia y exactitud se analizarán los datos de la base de datos "OASIS" de la ESPOCH para ayudar con la propuesta del SII-ESPOCH (Sistema de información Institucional) a lograr una solución que permita gestionar la información institucional de relevancia, de forma ágil, confiable, oportuna, que sirva de soporte para una adecuada toma de decisiones dentro de la institución y lograr una administración moderna y eficiente en el ámbito académico y administrativo.[2]

1.3. OBJETIVOS

1.3.1. Objetivo General

Realizar un análisis comparativo de las herramientas Data Quality propietario frente a las herramientas de software libre disponibles en el mercado.

[2] Tomado de:
http://www.espoch.edu.ec/Descargas/rectoradopub/b8242f_TALLER_EQUIPO_DE_GOBIERNO-PEDI.pdf

1.3.2. Objetivos Específicos

- Analizar las herramientas y procesos metodológicos dedicados a Data Quality.
- Determinar los criterios y parámetros de comparación de las herramientas propietarias: Informática Data Quality, Data Integrator y de software libre: DataCleaner, SQL Power DQGuru.
- Determinar las herramientas Data Quality más adecuadas para aplicarlas en la fuente "OASIS".
- Aplicar el proceso de Data Quality con las herramientas escogidas para la limpieza de los datos.

1.4. HIPÓTESIS

El uso de las herramientas para Data Quality permite mejorar la confiabilidad y consistencia de los datos en la base de datos "OASIS".

CAPÍTULO II

CONCEPTOS BÁSICOS DE CALIDAD DE DATOS

En este capítulo se definirá los conceptos básicos y términos que rodean el medio ambiente dentro de la calidad de datos.

Aunque estos términos no tienen definiciones estrictas, se puede utilizar esta sección como una referencia, por lo menos para el ámbito de cómo usar y qué esperar de las herramientas para la optimización de la calidad de datos, en relación con los temas que se analizarán.

2.1. INTRODUCCIÓN Y CONCEPTOS

2.1.1. Calidad de datos (DQ)

Calidad de datos se refiere a los procesos, técnicas, algoritmos y operaciones encaminados a mejorar la calidad de los datos existentes en empresas y organismos (2).

Trabajar con calidad de los datos por lo general varía mucho de un proyecto a otro, así como los problemas en la calidad de los datos varían mucho. Ejemplos de problemas de calidad de datos incluyen:

1. Integridad de los datos
2. Exactitud de los datos
3. La duplicación de datos
4. Uniformidad / normalización de los datos (4).

Una definición menos técnica de datos de alta calidad es: que los datos son de alta calidad "si son aptos para los usos previstos en las operaciones, la toma de decisiones y la planificación" (JM Juran) (3).

Sin embargo, la calidad de datos generalmente se refiere al mejoramiento de la calidad de los datos de personas físicas y jurídicas, pues son éstos probablemente los datos que más tienden a degradarse y cuya falta de calidad más impacta en la productividad de las organizaciones (2).

2.1.2. Beneficios de la calidad de datos

- **Ahorrar costes directos**: evitando tener información duplicada y por lo tanto evitar el envío duplicado de cartas a un mismo cliente (2).
- **Potenciar las acciones de marketing y la gestión**: la normalización de ficheros mejora el análisis de datos y permite segmentaciones precisas para que sus acciones de marketing y su gestión ganen en precisión y eficacia (2).
- **Optimizar la captación y la fidelización de clientes:** con los datos correctos, se mejoran los ratios de respuestas y el cliente se siente plenamente identificado con la empresa (2).
- **Mejorar la imagen corporativa:** el cliente sólo recibe el envío que le corresponde, una sola vez y con sus datos correctos (2).
- **Mejorar el servicio:** identificación más rápidamente del cliente que llama a un Call Center, reduciendo los tiempos de espera y, dejando tiempo al operador para centrarse en el mensaje de negocio (2).

2.1.3. Gestión de Datos Maestro (MDM)

La gestión de datos maestros describe un conjunto de disciplinas, tecnologías y soluciones utilizadas para crear y mantener datos coherentes, completos, contextuales y precisos de todas las partes (usuarios, aplicaciones, almacenes de datos, procesos y socios) (4).

El concepto de un repositorio que albergue todos los datos de referencia (o datos maestros) de la empresa no es nuevo. Tras años de informática distribuida, proliferación e imbricación de sistemas heterogéneos que dan soporte a las actividades de negocio y a la

gestión de la organización, todos los responsables de los sistemas de información sueñan con disponer de una base centralizada y depositaria exclusiva de "la verdad". Un sueño que ha sido elevado al rango de necesidad urgente debido a las nuevas exigencias de respuesta en contextos de fusiones y adquisiciones, gestión de rendimiento, conformidad normativa, etc. La falta de gestión unificada de los datos de referencia se traduce en pérdidas cotidianas de eficacia operativa que ejercen un impacto directo sobre el rendimiento global de la organización (5).

Existen varias normas que especifican un control para la calidad de los datos, la norma ISO 8000-110:2009 específica los requisitos que puedan ser controlados por ordenador para el intercambio, entre las organizaciones y sistemas, de los datos maestros que consta de los datos característicos. El objetivo de la norma ISO 8000-110:2009 es en los requisitos que puedan ser controlados por ordenador (6).

2.1.4. Análisis de la calidad de datos (DQA)

El análisis de la base de datos nos permite conocer por un lado la calidad de la misma y por otro el perfil de las personas/empresas que componen la misma (4).

El análisis de la base de datos puede ser de dos tipos:

- **Análisis de la estructura de la base de datos:** Mediante este análisis se investiga la estructura con la que cuenta la base de datos: Registros que componen la BD, campos a analizar, calidad de los datos en la misma. Este tipo de análisis nos da los resultados

necesarios para diseñar acciones concretas de actualización y enriquecimiento de la base de datos (4).

- **Análisis del contenido de la base de datos:** Este tipo de análisis se realiza cuando se supone que la BD tiene una estructura correcta y los campos y registros de la misma están actualizados (4).

2.1.5. Perfiles de datos / Data Profiling

Perfiles de datos es el proceso de analizar y explorar sus datos, para tener un mejor conocimiento y de entender si hay inconsistencias o entradas problemáticas en sus datos (4).

Perfiles de datos es la actividad de investigación de un almacén de datos para crear un "perfil" de la misma. Con un perfil un almacén de datos estará mucho mejor equipados para realmente mejorar (4).

La forma de hacer perfiles a menudo depende de si ya tiene algunas ideas acerca de la calidad de los datos o si no tiene experiencia en el almacén de datos. De cualquier manera se recomienda un acercamiento exploratorio, a pesar que hay una cierta cantidad de problemas las cuales hay que buscar, es muy importante verificar los datos en la información que se piensa que son correctas. Por lo general es barato incluir un conjunto de datos en su análisis y los resultados podrían sorprender y ahorrar tiempo (4).

2.1.6. Limpieza de datos / Data Cleasing.

El data cleansing, data scrubbing o limpieza de datos, es el acto de descubrimiento, corrección o eliminación de datos erróneos de una base de datos. El proceso de data cleansing permite identificar datos incompletos, incorrectos, inexactos, no pertinentes, etc. y luego substituir, modificar o eliminar estos datos sucios ("data duty"). Después de la limpieza, la base de datos podrá ser compatible con otras bases de datos similares en el sistema (7).

2.1.7. Auditoria de datos / Data Auditing

Auditoría de datos es el proceso de garantizar la calidad de datos desde el comienzo de un proceso para el destino final de una manera repetible y medida. Los datos de auditoría es el proceso de llevar a cabo una auditoría de datos para evaluar la cantidad de datos de la empresa que es apta para el fin determinado. Se trata de perfiles de los datos y evaluar el impacto de los datos de mala calidad en el desempeño de la organización y los beneficios (8).

2.1.8. Monitoreo de calidad de datos

Se ha argumentado que de perfiles de datos es una actividad ideal de exploración. El monitoreo de Calidad de los datos no lo es, las mediciones que se hacen al perfilar muchas veces necesita ser continuamente revisadas a fin de que las mejoras se aplican a través del tiempo. Esto es de lo que el monitoreo calidad de los datos se trata

Monitoreo Calidad de los datos se presenta en diferentes formas y tamaños. Se puede configurar su propio volumen de trabajos programados que se ejecutan todas las noches.

Construir alertas a su alrededor que envíe mensajes de correo electrónico, si una medida en particular va más allá de lo que está permitido, o en algunos casos se puede tratar de descartar el problema en su totalidad por el método del primer tiempo-derecha (FTR) principios que validan los datos en tiempo de entrada. Por ejemplo, en los formularios de registro de datos y más (4).

2.2. HERRAMIENTAS PARA CALIDAD DE DATOS

2.2.1. Características de las herramientas para calidad de datos

Los productos de calidad de datos permiten definir y automatizar estándares para asegurar que los datos dentro de la organización son capturados y almacenados de tal manera que puedan ser compartidos por otras aplicaciones y sistemas en cualquier momento. Proporcionan mecanismos que permiten auditar el nivel de la calidad de los datos, identificar problemas, desarrollar procesos que automatizan la resolución de esos problemas y, a continuación, monitorizar si se mantienen los niveles de calidad de la información requeridos por el negocio (9).

Las herramientas para la calidad de datos son soluciones dedicadas a detectar y corregir problemas en los datos dentro de un almacén de datos, aquellos datos que afectan a la precisión y la eficiencia de las aplicaciones de análisis de datos. Una gama de productos que es independiente de plataformas, en la medida más grande posible, deben ofrecer las condiciones óptimas para el análisis, limpieza y la reducción de problemas con la máxima

flexibilidad a través del apoyo de los fabricantes de ordenadores, sistemas operativos y aplicaciones que se pueden encontrar comúnmente en el mercado (9).

2.2.2. Ventajas del uso de las herramientas para calidad de datos

El uso de productos de Calidad de Datos trasciende en numerosos beneficios:

- **Operaciones Eficientes:** Se obtienen menores tasas de error en aplicaciones ERP Y CRM, páginas de comercio electrónico y otros sistemas transaccionales (9).
- **Reporting Preciso:** Desde los departamentos que se ocupan de la operación del día a día hasta las áreas directivas de planificación estratégica reciben informes que reflejan fielmente la realidad del negocio (9).
- **Análisis Intuitivo**: Resultados mejores y más fiables de las herramientas predictivas, presupuestarías, de campañas de marketing y de Inteligencia de Negocio (9).
- **Mejor Servicio al Cliente**: Una visión unificada de la información de contacto, incluyendo el historial de compras y preferencias, proporciona a los empleados que trabajan de cara al cliente la información que necesitan para ofrecerles el mejor servicio para incrementar su satisfacción (9).
- **Aumento de Ingresos**: Se pueden identificar nuevas oportunidades de aumentar ventas y ventas cruzadas, gestionar las cuentas de forma eficaz, entender y anticipar los patrones de compras del cliente a través de en una visión uniforme del cliente (9).
- **Cumplimiento Regulatorio Fiable:** Disponer de informes precisos y establecer procesos de negocio relacionados con la gestión de datos fiables y replicables permite que se cumplan normativas como Basilea II, LOPD, Sarbanes Oxley, ISO, blanqueo de capitales, etc. de marketing y de Inteligencia de Negocio (9).

2.2.3. Aplicabilidad de las herramientas para calidad de datos

Las organizaciones aplican productos de calidad de datos en muchos entornos diferentes, porque los datos cambian de uno a otro rápidamente y a menudo se reutilizan o son capturados a través de sistemas diferentes. Los procesos de calidad de datos se incluyen habitualmente en:

- **Sistemas CRM:** Como fundamento de una visión unificada de los clientes.
- **Sistemas ERP:** Para eliminar los errores en las transacciones, especialmente antes de que la información se replique en otros sistemas.
- **Data Warehouses:** Para hacer cumplir los estándares e integrar los datos de las aplicaciones fuente.
- **Comercio Electrónico:** Para garantizar que la información que entra en los sistemas empresariales sea exacta y completa desde el principio.
- **Arquitecturas Orientadas a los Servicios (SOA):** Para ofrecer estándares reutilizables, basados en reglas, que puedan incorporarse en procesos de negocios complejos en tiempo real.
- **Integración de Datos:** Reduciendo riesgos involucrados en la consolidación de grandes volúmenes de datos y migraciones de un sistema a otro.

Las organizaciones a menudo comienzan abordando el problema de la Calidad de Datos dentro de un sistema o aplicación conocido por tener mal los datos o como parte de otro proyecto de sistemas. Los beneficios más altos se obtienen cuando los mismos procesos de

calidad de datos se van expandiendo posteriormente a todos los sistemas fuentes que reciben información nueva (9).

2.3. TIPOS DE HERRAMIENTAS PARA CALIDAD DE DATOS

Se ha logrado identificar algunas de las herramientas más conocidas en el mercado y se lo ha dividido en dos grupos los cuales son los siguientes:

- **Propietarias:**
    - Ascential Software.
    - DataFlux.
    - Evoke Software.
    - Informatica.
    - Trillium Software.
    - IBM InfoSphere Quality Stage.
    - Sap Data Integrator.
    - Oracle Data Integrator
- **Open Sourse:**
    - Talend Data Quality.
    - Data Cleaner.
    - Infosolve.
    - SQL Power.

Todas estas herramientas nos permiten realizar análisis de datos y utilizar una o varias técnicas de Data Quality, para este caso de estudio serán analizadas las herramientas propietarias Oracle Data Integrator e Informatica y las herramientas libres DataCleaner y

SQL Power debido a su facilidad de obtención y por ser las más populares dentro de las herramientas para calidad de datos, estas herramientas serán analizadas en instancias posteriores (10).

2.3.1. INFORMATICA

Informatica Corporation proporciona software de integración de datos y servicios que permiten a las organizaciones obtener una ventaja competitiva en la economía global de la información mediante la potenciación de los datos oportunos, relevantes y de confianza para sus principales procesos de negocio (11).

Al permitir a los consumidores de la información asumir la responsabilidad de la calidad de la información, Informatica Data Quality permite a las personas que están en mejor posición para entender las necesidades de información de calidad de la organización. Permite a los trabajadores del conocimiento implementar perfiles de calidad de los datos, la limpieza, la estandarización, la adecuación y seguimiento de los procesos en toda la empresa. Informatica Data Quality permite a los analistas de negocio y la gente de TI de su organización para implementar estrategias de calidad de datos eficaces y duraderos (11).

2.3.1.1. Informatica Data Quality: Es una suite de aplicaciones y componentes que se pueden integrar con Informatica PowerCenter para entregar la empresa la capacidad de resistencia a la calidad de los datos en una amplia gama de escenarios (12).

Los componentes básicos son:

- **Data Quality Workbench:** Se utiliza para diseñar, probar y desplegar procesos de calidad de datos, llamados planes. Workbench le permite probar y ejecutar planes, según sea necesario, lo que permite la investigación rápida de datos y prueba de metodologías de calidad de datos. También puede implementar los planes, así como los datos asociados y archivos de referencia, a otras máquinas de Calidad de Datos. Los planes se almacenan en un repositorio de Calidad de Datos.

  Workbench proporciona acceso a múltiples bases de datos, basado en archivos, y algorítmicos componentes de calidad de datos que puede utilizar para construir los planes (12).

- **Data Quality Server:** Se utiliza para habilitar el uso compartido de archivos y ejecutar los planes en un entorno de red. El servidor de Datos de Calidad apoya a la creación de redes a través de servicios de dominios y se comunica con workbench a través de TCP / IP. El servidor de calidad de los datos permite a múltiples usuarios colaborar en proyectos de datos, acelerar el desarrollo e implementación de la solución de calidad de datos. (12).

  Puede instalar los siguientes componentes, junto con workbench y el servidor.

- **Integration Plug-In:** Este plug-in de Informatica habilita a PowerCenter para ejecutar los planes de calidad de datos para la normalización, la limpieza, y las operaciones correspondientes. La integración de plug-in está incluido en el Informatica Data Quality al instalar el conjunto de archivos (12).

- **Free Reference Data:** Es un diccionarios de texto básico empresarial y términos de clientes.
- **Subscription-Based Reference Data:** Bases de datos, procedente de terceros, de entrega de direcciones postales en un país o región (12).
- **Pre-Built Data Quality Plans:** Planes de Calidad de datos construida por Informatica, para llevar a cabo limpieza, estandarización, y las operaciones correspondientes. Informatica proporciona planes de demostración. Donde Usted puede comprar planes pre construidos para el uso comercial (12).
- **Association Plug-In:** Este plug-in de Informatica permite a PowerCenter identificar las coincidencias de campos de datos a partir de las transformaciones de integración múltiples y asociar estos registros con fines de consolidación de datos (12).
- **Consolidation Plug-In:** Este plug-in de Informatica permite a PowerCenter comparar los registros vinculados enviados como salida de una transformación de la Asociación y crear un solo registro maestro de estos registros (12).

2.3.1.2. Características principales de Informatica Data Quality

- **Negocios enfocados en calidad de datos** - Activación de la empresa a su "propia" calidad de los datos mediante el establecimiento de objetivos de calidad de datos. En estas primeras etapas, los analistas de negocio identifican, clasifican y cuantifican los problemas de calidad de datos, y con ello se apropian (12).

- **Todos los tipos de datos maestros** - Ampliación de la calidad de datos a través de todas las unidades de negocio para abarcar todos los datos de la empresa-cliente, producto, financieros, materiales, precios, pedidos y datos de los activos (12).
- **Métricas La calidad de datos e informes** - permite la medición y por lo tanto gestión efectiva de la calidad de los datos para impulsar el cambio dentro de una organización con dimensiones significativas, tales como la integridad, la conformidad, consistencia (12).
- **Implementación empresarial** - Apoyo a todos los requisitos de TI para los procesos de calidad de datos proactiva, incluyendo la escalabilidad y la reutilización de una infraestructura de alto rendimiento para la integración de datos de calidad y de datos (12).

2.3.1.3. Componentes de Infromatica Data Quality

Los componentes operacionales de Informatica Data Quality proporcionan una amplia gama de análisis de datos y la funcionalidad de mejora que abarcan todos los aspectos de un proyecto de datos.

Informatica Data Quality organiza sus planes y proyectos de una manera modular. De manera predeterminada, un nuevo proyecto de calidad de datos en workbench contendrá las subcarpetas de cuatro módulos: creación de perfiles, la normalización, de comparación, y la consolidación (12).

- **Perfiles de datos** permiten realizar un análisis FODA de los datos - la identificación de sus fortalezas y debilidades y las oportunidades y amenazas que pueden

representar para su proyecto y de la empresa. Las herramientas clave de información en la etapa de creación de perfiles son cuadros de mando, que presentan información estadística clara y gráfica sobre la calidad de sus datos y que pueden activar los mensajes de alarma cuando los umbrales estadísticos cumplen o es violada (12).

- **Estandarización de datos** tanto estandarizar la forma de sus datos como validar y mejorar los datos mediante la comparación de los valores en una serie de filtros y los diccionarios de referencia. Normalización de las operaciones pueden ser elemental "eliminación de ruido" o múltiples reglas de negocio a medida, en todos los casos, los componentes gráficos de calidad de los datos hacen definiciones de las operaciones necesarias lo suficientemente simples para todos los niveles de usuarios. (12).
- **Coincidencia de datos** identifica los registros de datos equivalentes, duplicados, o afines, dentro de un conjunto de datos (archivos o tablas de bases de datos) o entre conjuntos de datos. En consecuencia se puede identificar mediante la comparación de los datos inexactos con los datos de los diccionarios de referencia. Los planes de coincidencias se utilizan también para prepararse para la consolidación de bases de datos mejoradas (12).

La Calidad de los Datos emplea poderosos algoritmos de comparación que, como todas las funciones de Calidad de Datos, operar "bajo cubierta" del software, de modo que el usuario puede definir y ajustar los procesos de comparación en forma rápida y sencilla en el escritorio. Los resultados de comparación se pueden ponderar

en el proceso de comparación de manera que las salidas de datos resultantes pueden reflejar las prioridades del negocio (12).

- **La consolidación de datos** es a menudo el paso final en un ciclo de calidad de los datos, se permite a la empresa crear, por ejemplo, los nuevos conjuntos de datos que se utilizarán en unos ambientes comerciales, o para crear el conjunto de datos maestros para un proyecto de migración de datos (12).

2.3.1.4. Beneficios

- Incrementa ventas y servicios de clientes utilizando datos de gran calidad
- Mejorar la toma de decisiones con mayor confianza en los datos
- Reducir los riesgos asociados con los informes de cumplimiento
- Mejorar la eficiencia operativa con un proceso de gestión de la calidad de datos en curso
- Reducir los costos y riesgos del proyecto (12).

2.4. ORACLE DATA INTEGRATOR

La Familia de Productos Oracle para Calidad de Datos Empresariales ayuda a las organizaciones a alcanzar el máximo valor de sus aplicaciones críticas de negocio mediante la entrega de datos adecuadas a su propósito. Estos productos también permiten a los individuos y los equipos de colaboración identificar y resolver rápida y fácilmente cualquier problema en los datos subyacentes. Con los productos empresariales Oracle para calidad de datos, los clientes pueden identificar nuevas oportunidades, mejorar la eficiencia

operativa, y cumplir de manera más eficiente con la industria o la regulación gubernamental (13).

Rápido de instalar y fácil de usar, los productos de Oracle Enterprise Data Quality traer la capacidad de mejorar la calidad de los datos para todos los interesados en cualquier iniciativa de gestión de datos. Los productos Oracle Enterprise de calidad de datos son:

- Oracle Enterprise Data Quality Profile and Audit
- Oracle Enterprise Data Quality Parsing and Standardization
- Oracle Enterprise Data Quality Match and Merge
- Oracle Enterprise Data Quality Address Verification Server
- Oracle Enterprise Data Quality Product Data Parsing and Standardization
- Oracle Enterprise Data Quality Product Data Match and Merge

Cada uno de estos productos se describe en las siguientes secciones (13).

2.4.1. Oracle Enterprise Data Quality Profile and Audit.

Oracle Enterprise Data Quality Profile and Audit proporciona una base para la comprensión de los problemas de calidad de datos y una base para la creación de reglas de calidad de datos de la gestión correctora y de prevención. Proporciona la capacidad de entender los datos, destacando las áreas clave de discrepancia de datos, lo que ayuda a analizar el impacto en el negocio de estos problemas y aprender del análisis histórico, y definir reglas de negocio directamente de los datos. Esto evita las ideas preconcebidas de cómo los campos de datos que se relacionan entre sí y se identifica rápidamente las deficiencias en los procesos de negocio existentes y las implementaciones de tecnología (13).

Oracle Enterprise Data Quality Profile and Audit permite a los equipos de negocio definir el perfil de grandes volúmenes de datos de bases de datos, hojas de cálculo y archivos planos con facilidad. El enfoque único de Oracle de perfiles de frases para la comprensión de textos de datos ayuda a identificar la información clave oculta en los campos de datos de texto de formato libre. Proporcionar un área de ensayo que contiene las estadísticas recogidas, Oracle Enterprise Data Quality Profile and Audit deja la fuente de datos sin alterar. Los exámenes sistematizados de auditoría detectan indicadores claves de calidad, los datos que faltan, los valores incorrectos, los registros duplicados e inconsistencias. Cuando se utiliza en conjunto con Oracle Enterprise Data Quality Parsing and Standardization, puede ofrecer comprensión sin precedentes de sus datos (13).

Los resultados de estos procesos de creación de perfiles y la auditoría se presentan en cuadros de mando ejecutivos fáciles de entender. Utilizando un navegador web, los trabajadores y los gerentes pueden supervisar y revisar la calidad de datos en curso contra las métricas definidas. Los cuadros de mando de calidad de datos permiten que los problemas sean identificados y se traten con rapidez antes de que comiencen a causar un impacto comercial significativo. Las vistas gráficas muestran las tendencias de la calidad de datos a través del tiempo, lo que ayuda a su organización a proteger su inversión en la calidad de los datos, dando visibilidad a las personas adecuadas (13).

2.4.2. Oracle Enterprise Data Quality Parsing and Standardization

Oracle Enterprise Data Quality Parsing and Standardization proporcionan una paleta rica de funciones para transformar y normalizar los datos mediante los datos de referencia de fácil

administración y configuración gráfica simple. Además de las funciones base para valores numéricos, cadenas y los campos de tipo fecha, se proporcionan funciones para datos contextuales, como los nombres de direcciones y números de teléfono. Los usuarios también pueden configurar rápidamente, empaquetar, compartir y desplegar nuevas funciones que encapsulan reglas específicas a sus datos y a la industria sin ningún tipo de codificación (13).

Los datos de texto están muy rara vez disponibles de una manera completamente limpia y ordenada. Los problemas más comunes incluyen:

- Campos construidos, en los que un ID de cliente puede estar formado por un código de ubicación, una referencia al cliente, y un código de administrador de cuentas.
- Datos atrapados, como nombres, comentarios, o números de teléfono en los bloques de direcciones.
- Datos mal estructurados, como direcciones donde los datos pueden fluir de un campo a otro.
- Campos de notas que almacenan la información de la estructura de datos no admitibles, pero que contienen datos semi-estructurados útiles que normalmente no se pueden analizar o extraer (13).

Todos estos problemas se pueden resolver utilizando Oracle Enterprise Data Quality Parsing and Standardization. Utilizando un enfoque basado en datos para etiquetar o describir rápidamente los datos, se puede manipular un único registro por que el análisis en múltiples elementos de estructura (y, si es necesario, registros) y normalizar los resultados

de acuerdo con reglas predefinidas. La tecnología innovadora de análisis y análisis de expresión permite únicamente encontrar conocimiento oculto dentro de cualquier campo de texto y crear reglas para estandarizarlos en datos estructurados (13).

Oracle Enterprise Data Quality Parsing and Standardization también se pueden utilizar para auditar las reglas definidas del negocio y transformar los datos sobre la marcha contra esas reglas, proporcionando un servidor de seguridad flexible y adaptable de calidad de datos. Además, permite reunir todo el proceso de calidad de datos para ser llamado como un servicio Web en tiempo real. Los resultados de los procesos de análisis y estandarización se pueden ver en los cuadros de mando gráficos que proporcionan una visión completa, exacta y accesible del mundo de los negocios (13).

2.4.3. Oracle Enterprise Data Quality Match and Merge

El emparejamiento es un componente clave de muchos de los proyectos de calidad de datos, y se puede utilizar para apoyar las diferentes actividades, como la de-duplicación, consolidación, integración de datos de clientes (CDI) y la gestión de datos maestros (MDM). Oracle Enterprise Data Quality Match and Merge proporciona capacidades de emparejamiento de gran alcance que le permiten identificar los registros coincidentes y, opcionalmente, enlazar o combinar registros coincidentes en base a reglas de supervivencia. La flexible e intuitiva configuración permite ajustar las reglas para adaptarse a la tarea y apoyar en un enfoque iterativo. Una capacidad independiente y de sólo una simple revisión significa que puede exponer a los resultados de los emparejamientos para su revisión, sin acceder a la configuración de reglas subyacente. Utilizada en conjunto con los otros

productos miembros de la familia, Oracle Enterprise Data Quality Match and Merge se convierte en una solución muy potente y flexible que se puede adaptar para producir resultados impresionantes en diferentes proyectos (13).

Oracle Enterprise Data Quality partido y combinar también incluye un conector que le permite acceder fácilmente a los datos de Siebel CRM de Oracle. Las funciones de auditoría le permiten ejecutar las reglas de calidad de datos y de control de flujo dentro de sus procesos de calidad de datos. La funcionalidad cuadros de mando presenta los resultados de los procesos de auditoría en formato gráfico, mientras que la funcionalidad servicio Web en tiempo real permite a todo el proceso de calidad de datos agruparse para ser llamado como un servicio en tiempo real (13).

2.4.4. Oracle Enterprise Data Quality Address Verification Server

Muchos de los problemas de calidad de datos incluyen nombres y direcciones. Una cosa para certificar una dirección es validar el formato, pero otra muy distinta es comprobar que en realidad existe y es una dirección real, donde se puede entregar. Oracle Enterprise Data Quality Address Verification Server llena este vacío y utiliza la información de referencia de diversas autoridades postales de todo el mundo para verificar que las direcciones son "reales". Además, para las direcciones verificadas, también puede devolver un geo-código para aplicaciones de mapeo. El sistema puede analizar, estandarizar, transcribir y procesar las direcciones de más de 240 países - básicamente todos los territorios poblados del planeta y puede manejar en forma estructurada o no estructurada, y en cualquier juego de caracteres (13).

2.4.5. Oracle Enterprise Data Quality Product Data Parsing and Standardization

En el mundo o en calidad de datos, los datos del producto proporcionan algunos retos específicos. Las normas que rigen los datos del producto son específicas de la categoría del producto que se describe. Por ejemplo, las normas de calidad de datos para resistencias son diferentes de condensadores, que también son diferentes de interruptores, sujetadores, y cualquier otra categoría de productos. Cada categoría de producto tendrá un vocabulario diferente, términos, abreviaturas, los valores válidos y estandarizaciones. Además, la información del producto se comunica normalmente a través de campos de descripción no estándares que deben ser reconocidos y analizados. Agravando este problema, la mayoría de los escenarios de calidad de datos que involucran datos de productos no cubren sólo una categoría, sino cientos o miles de categorías de productos (13).

Para hacer frente a este nivel de Oracle Enterprise Data Quality Product Data Parsing and Standardization utiliza el reconocimiento semántico para reconocer rápidamente la categoría de producto y aplicar las normas adecuadas en función del contexto. En función del contexto, también puede hacer inferencias sobre el significado de una palabra o frase en particular y "aprender" las nuevas normas y el contexto a medida que avanza. Una vez reconocido correctamente la información del producto, este puede ser transformado y estandarizado incluida la clasificación (es), atributos, y las descripciones que se pueden generar en cualquier idioma para el consumo en los sistemas posteriores (13).

2.4.6. Oracle Enterprise Data Quality Product Data Match and Merge

Los datos del producto también presentan retos específicos para emparejar y fusionar los registros de productos. Oracle Enterprise Data Quality Product Data Parsing and Standardization se utiliza normalmente para crear un registro de productos estandarizados, mientras que Oracle Enterprise Data Quality Product Data Match and Merge es capaz de identificar los registros exactos, similares y afines, opcionalmente, los combina en base a reglas de supervivencia definidas (13).

Oracle Enterprise Data Quality Product Data Match and Merge puede funcionar en cualquier idioma e incluye un conector para Oracle Data Product Data Hub, permitiendo limpiar, estandarizar, de-duplicar la información de un producto para ser cargado en el concentrador MDM (13).

2.5. DATA CLEANER

2.5.1. Datos de Perfiles

Perfiles de datos es el proceso de analizar y explorar sus datos, para tener un mejor conocimiento y entender si existen inconsistencias o entradas de otro modo problemáticos en sus datos (4).

DataCleaner proporciona un rico conjunto de características de perfiles de datos, incluyendo la búsqueda de patrones, que muestra la distribución y frecuencia de los valores, el recuento y factores de ponderación, la investigación de juegos de caracteres, y trazar líneas de tiempo de las entradas de datos periódicos. Además hay una gran cantidad de

parámetros "normales" que le dan la base para juzgar si los datos están en una forma utilizable y dónde cuidado (4).

2.5.2. Limpieza de Datos

DataCleaner, como su nombre sugiere, se puede utilizar como una plataforma para la limpieza de los datos. A través de un amplio conjunto de transformadores y analizadores, he aquí algunos ejemplos de lo que puede hacer:

- Estandarizar los valores utilizando sinónimos.
- Conformar fechas a un formato de fecha única.
- Analizar, extraer y normalizar la información, por ejemplo el uso de expresiones regulares o scripts (4).

Además DataCleaner se integra con EasyDataQuality, que ofrece a la carta funciones de calidad de datos como un servicio. A través de EasyDataQuality puede realizar la limpieza avanzada y operaciones de enriquecimiento como:

- Dirección de validación y depuración.
- Validación de números telefónicos y la categorización tipo de línea.
- Validación y corrección de direcciones de Email (4).

2.5.3. Easy DQ y DataCleaner

Obtenga información inmediata en la calidad de sus datos. DataCleaner le permite generar rápidamente perfiles útiles y métricas sobre sus datos. Ahora puedes disfrutar de limpieza

EasyDQ en tiempo real y funciones de correspondencia, totalmente integrado en su herramienta favorita de código abierto de calidad de datos (4).

- **Rápido y Flexible:** DataCleaner maneja millones de discos, fuentes múltiples y una multitud de datos de perfiles de indicadores.
- **Potente y probada:** Miles de descargas, una gran comunidad de usuarios y probado en varias empresas internacionales.
- **Fácil de usar:** Comienza en cuestión de minutos. Fomenta y amplía su análisis de forma interactiva.
- **Limpieza y Coincidencia por EasyDQ**: Permite a DataCleaner ser cliente de pleno derecho de herramientas de calidad de datos.

2.5.4. Apoyo a otras Herramientas

- **Perfiles pasos ETL utilizando DataCleaner**

Cuando se trabaja con ETL's que a menudo se encuentra preguntando qué tipos de valores puede esperar de una transformación particular. Con el paquete de calidad de los datos de Pentaho, ofrecemos una integración única de perfiles y ETL: Simplemente haga clic en cualquier fase de su transformación, seleccione "Perfil", y se iniciará DataCleaner con los datos disponibles para el perfil (4).

- **Ejecutar trabajo DataCleaner**

Otra gran característica en el paquete de calidad de datos Pentaho es que ahora puede incrustar y ejecutar trabajos DataCleaner con la integración de datos Pentaho. Esto hace que sea mucho más fácil de manejar ejecuciones programadas, control de calidad de datos e

incrustar múltiples trabajos DataCleaner. Mezclar y combinar DQ DataCleaner de puestos de trabajo con las transformaciones y que se tiene lo mejor de ambos mundos (4).

- **Integración EasyDQ**

Además, el paquete de calidad de datos para Pentaho contiene las funciones de limpieza EasyDQ como pasos ETL, de forma similar a lo que sabe de sus contrapartes DataCleaner. De-duplicación y la fusión a través de DataCleaner Además de incrustar DataCleaner para el perfil de pasos, también se puede poner en marcha DataCleaner al navegar por bases de datos en la integración de datos Pentaho. Esto creará una conexión de base de datos que sea adecuada para más profundidad en las interacciones con la base de datos. Por ejemplo, se puede usar para encontrar duplicados en el origen o las bases de datos de destino (4).

2.6. SQL POWER

Fundada en 1988, SQL Power Group es un líder en Business Intelligence y la firma de migración de datos de software. Nuestro modelado de datos probados, limpieza de datos y herramientas de BI de informes han ayudado a ofrecer soluciones de Business Intelligence y CRM de la más alta calidad a los clientes orientados al valor. SQL Power nos ha consolidado como el primer ministro de Business Intelligence y Data Migration proveedor de soluciones en Canadá y alrededor del mundo (14).

2.6.1. Ventaja de SQL Power

A diferencia de otros desarrolladores de software de BI, SQL Power se especializa en brindarle las herramientas de productividad que sólo trabajan. Nuestras herramientas son fáciles de instalar, fácil de aprender y fácil de usar porque los diseñamos desde cero para trabajar de forma intuitiva. Trabajando mano a mano con los consultores y usuarios finales en todo el proceso de desarrollo, ofrecemos herramientas para satisfacer las necesidades de todas las facetas de su entorno BI desde el diseño de modelo de datos para limpieza de datos para informes de análisis (14).

2.6.2. SQL Power DQguru

El SQL Power DQguru ayuda a limpiar los datos, validar y corregir las direcciones, identificar y eliminar los duplicados y crear referencias cruzadas entre tablas de origen y de destino. Esto proporciona a los usuarios de negocio con datos completos y precisos, y una sola visión de 360 grados de todas las entidades empresariales, como cliente, producto, representativas, unidad de empleado, proveedor o negocio (15).

- **Limpieza de DQguru:** Ideal para la limpieza de cualquier almacén de datos o base de datos CRM, SQL Power DQguru permite limpiar prácticamente cualquier base de datos mediante una interfaz intuitiva, verificación de correspondencia lo que es más fácil que nunca para ofrecer a los usuarios una visión completa y exacta de 360 grados de los datos de sus negocios (15).
- **Interfaz de DQguru:** Desde la limpieza de datos puede ser inherentemente complejo, hemos desarrollado DQguru SQL Power para que el proceso sea lo más intuitivo

posible el uso de interfaces innovadoras e integrales para la definición de medidas de transformación de datos y validación de correspondencia visual (15).

- **DQguru hace corrección de dirección:** Utilizando una función de consulta Google Geocode, SQL Power DQguru le permite validar y corregir automáticamente los datos de dirección para su uso con aplicaciones de Google Maps (15).
- **DQguru trabaja con cualquier base de datos:** SQL Power DQguru se puede utilizar en prácticamente cualquier plataforma de base de datos. Así que si los datos de su cliente, producto o negocio residen en Oracle, MySQL, PostgreSQL o cualquier otra plataforma, DQguru puede limpiar y eliminar duplicados de sus dimensiones esenciales (15).
- **DQguru se ejecuta en todas las plataformas:** SQL Power DQguru está basado en Java y funciona en cualquier entorno que soporte JRE 5.0 o mejor por lo que si se está utilizando PC, Mac, Unix, otra cosa o todo lo anterior, SQL Power DQguru simplemente funciona (15).
- **DQguru Community Edition es libre de implementar:** Construido como una herramienta de limpieza de datos mejor, y creemos que cuando lo intente, usted estará de acuerdo. Es por eso que ofrecen descargas gratuitas de la herramienta SQL Edition DQguru en la comunidad (15).

2.6.2.1. Características

- Interfaz gráfica de usuario intuitiva permite una rápida adopción y utilización por los analistas de datos.
- Interfaz intuitiva "transformar" este proceso le permite crear y desplegar rápidamente los flujos de trabajo de conversión de datos.
- Los usuarios pueden definir sus propios criterios de coincidencia de datos.
- Duplicar la verificación a través de la innovadora interfaz SQL Power.
- Combinar duplicados y los datos relacionados.
- Puede ser utilizado para los datos iniciales o periódicos de limpieza.
- Genera tablas de referencias cruzadas para vincular los identificadores de origen del sistema de identificadores de la base de datos de destino.
- Amplio soporte para la transformación y funciones coincidentes:
- Concatenación
- Doble Metaphone, Metaphone, refinados codificación Soundex, Soundex fonéticos.
- Alphabet conversión de mayúsculas.
- Cadena de sustitución.
- Subcadena y substring por palabra.
- Palabra de sustitución a través de la traducción de las palabras o grupos de palabras.
- Combinar reglas para fusiones columna de la tabla, tabla y afines.
- Distintos niveles de transformación de datos también son compatibles para ayudar a controlar el desarrollo y ejecución de los procesos de limpieza de datos:

- El motor de coincidencias identifica duplicados, almacenando los resultados en una tabla de destino, sin modificar los datos de origen.
- El motor de fusión elimina los registros duplicados de los datos de origen de acuerdo con las reglas que se han definido.
- El motor de limpieza sustituye a los registros de los datos de origen de datos reformateados según sus reglas.
- Amplio soporte para varias bases de datos para los datos de origen y destino.

Si usted está construyendo un almacén de datos, Data Mart o CRM, el Poder SQL DQguru da un largo camino para asegurar la integridad de los datos de su entorno de apoyo a las decisiones o base de datos CRM (16).

2.7. METODOLOGÍAS

Las metodologías que se detallarán a continuación son: INFORMATICA, ADASTRA, DATACTICS, SII-ESPOCH.

2.7.1. INFORMÁTICA

La calidad de datos se gestiona mejor como parte de una arquitectura de integración de datos empresariales y, como resultado, el control y la gestión de la calidad de datos se complementa con el ciclo de vida de acceso, integración, transformación y entrega de los datos.

Como parte del programa de calidad de datos, las organizaciones necesitan establecer o restablecer procesos de calidad de datos como se muestra a continuación (17).

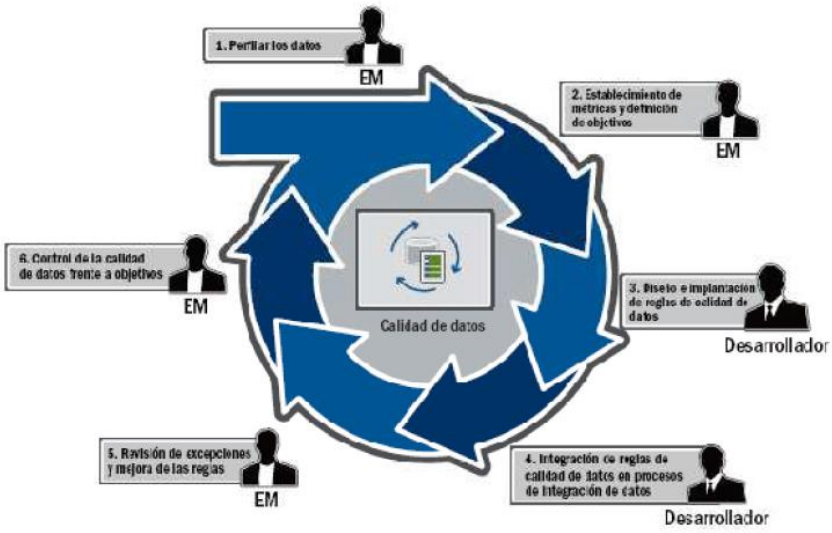


Figura II. 1. Fases Informática.

Fuente: Propuesta Metodológica SII-ESPOCH

- **Perfilado de Datos:** Es un elemento clave en la planificación de las iniciativas de calidad de datos, ya que permite determinar el contenido, la estructura y la calidad de estructuras de datos muy complejas, así como descubrir incoherencias ocultas e incompatibilidades entre las fuentes de datos y las aplicaciones de destino (17).
- **Establecer métricas y definir los objetivos:** Ayuda a los equipos de IT y a las empresas a medir los resultados obtenidos gracias a los esfuerzos realizados para garantizar la calidad de datos como parte de la iniciativa de BI (17).
- **Diseño e implementación de reglas de calidad de datos: Ayudan** a definir y medir los objetivos y los criterios de la calidad de datos (17).

- **Integración de reglas y actividades de calidad de datos:** Creación de perfiles, limpieza/correspondencia, solución automatizada y gestión con los procesos de integración de datos es fundamental para mejorar la precisión y el valor de los activos de datos (17).
- **Revisión de excepciones y la mejora de las reglas:** Se realizan de forma más eficaz como un esfuerzo conjunto que implica a miembros del equipo principal y a interesados de BI. En muchos casos, éstos últimos tienen un control limitado sobre los procesos empresariales y los sistemas operativos y esto hace que los datos sean de mala calidad (17).

  Por este motivo, es importante que los principales interesados y los ejecutivos de una organización participen en la documentación de los problemas de calidad de datos y en la ejecución de un programa de calidad de datos formal (17).
- **Control proactivo de calidad de datos:** Este control en cuadros de mando y notificaciones en tiempo real también se está convirtiendo en una de las mejores prácticas estándar. Los propios interesados de BI que participan activamente en el proceso de calidad de datos, pueden contar con las herramientas necesarias para ejercer esta tarea, ya que son los que mejor conocen cuál es el nivel de calidad que deben tener los datos (17).

2.7.2. ADASTRA

La calidad de datos no es un problema de TI, es un problema de toda la empresa y un activo fundamental que depende en gran medida en los procesos de negocio. También está

claro que la calidad de datos no es una simple tarea de una sola vez, es un complejo proceso iterativo y cíclico que emplea personal, herramientas y conocimiento. Usando este enfoque permite la identificación de problemas de calidad de datos en sus fuentes y conduce a eliminar las causas de los problemas en lugar de sus consecuencias (17).

Según la empresa ADASTRA hay cuatro fases básicas en la gestión del ciclo de calidad de los datos:

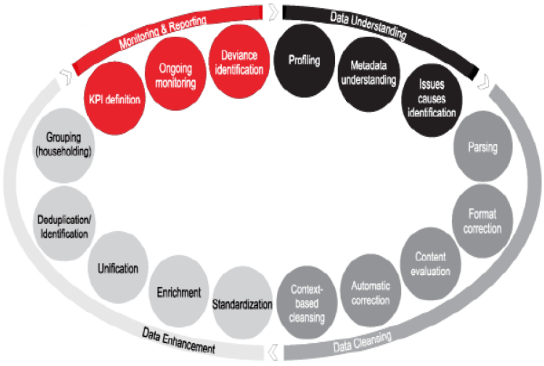


Figura II. 2. Fases y Etapas ADASTRA

Fuente: Propuesta Metodológica SII-ESPOCH

- **Comprensión de datos:** Es un conjunto de actividades entre ellas se encuentran los procesos de perfilado de datos, metadatos e identificación de posibles causas (17).
- **Limpieza de datos:** Es el proceso mediante el que la calidad de datos se ha mejorado y los problemas de datos junto con sus causas se resuelven y los procesos relativos son corregidos o mejoran (17).
- **Mejora de los datos:** Este proceso le permite agregar valor a los datos existentes (mediante la adición de la información disponible de otros-fuentes de datos externas

como la codificación geográfica, la dirección detallada y-o información sobre productos, etc.) (17).

- **Monitoreo de los datos y elaboración de informes:** Comprueba constantemente los datos y se identifican problemas nuevos datos o pérdida de calidad proporciona a los equipos de aplicación y los clientes de datos con la información sobre el éxito de las mejoras anteriores actividades (17).

2.7.3. DATACTICS

Este proceso de gestión de calidad de datos combina la experiencia de los analistas de datos y proporciona tecnología de generación de excelencia en calidad de datos en cada proyecto (17).

Ha sido desarrollada a partir de experiencias en situaciones de la vida real de calidad de datos, proporcionando un marco coherente (17).

Según DATACTICS el proceso de gestión de calidad de datos está compuesto de las siguientes fases:

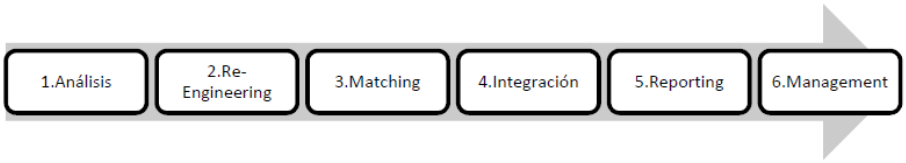


Figura II. 3. Fases de DATACTICS
Fuente: Propuesta Metodológica SII-ESPOCH

- **Análisis:** El paso inicial de cualquier estrategia de gestión de datos integral de calidad tiene que ser el descubrimiento de lo que exactamente contiene los datos que es importante y que no (17).
- **Re-Engineering:** Una vez determinado el contenido de los datos, la fase de re-ingeniería permite corregir errores, transformar los datos a las normas requeridas; mejorar los datos con información adicional y si es necesario, extraer elementos clave y de valor de la misma (17).
- **Matching:** Tener datos reestructurados a un formato y nivel adecuado se puede llevar acabo en su nivel más eficaz (17).
- **Integración:** La calidad de los datos no puede ser visto como un proceso aislado, por lo que la capacidad de integrar una metodología de gestión de calidad de los datos en los procesos empresariales existentes es fundamental (17).
- **Reporting:** Después de que las fases anteriores se han terminado, una gran cantidad de conocimientos e información sobre los datos han sido recogidos. La posibilidad de revisión, auditoría y compartir esta información es vital para una verdadera cultura de calidad de datos va a crecer y mejorar iterativamente en toda la organización (17).
- **Management:** La gestión de todos los procesos anteriores dentro de un único y simple ambiente proporciona un gran beneficio. Facilita el aumento de la productividad mediante la racionalización del flujo de trabajo y también proporciona una capa totalmente transparente desde el que profesionales de la calidad de los datos puede tanto monitorear y ejecutar los procesos de calidad de datos (17).

2.7.4. SII-ESPOCH

- **Fase 1. Estudio Y Preparación:** Esta fase es de vital importancia ya que es donde las metas del estrategias deben unir todas las acciones y decisiones una gestión siempre debe comenzar con la pregunta ¿Por qué es importante para empresa/organización?".

  Todo lo hecho con la información debe apoyar objetivos, y este paso asegura que se está trabajando en situaciones de importancia para negocio (17).

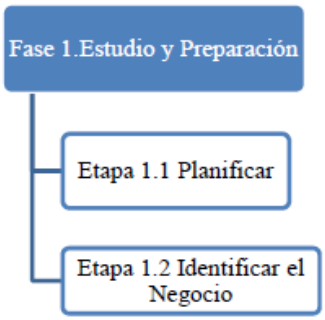


Figura II. 4. Etapas de la Fase 1.

Fuente: Propuesta Metodológica SII-ESPOCH

  - **Etapa 1.1 Planificar:** Iniciar el proyecto para hacer frente a los problemas elegidos ya sea como un equipo o individual (dependiendo del tamaño del negocio) esta etapa es fundamental ya que más de un proyecto fracasa debido a la incomprensión entre los implicados (gerencia, miembros del equipo, empresas, TI, etc.) (17).

    Para evitar tener problemas a lo largo del proyecto se debe evitar la falta de claridad en cuanto a lo que se llevara a cabo. Una planificación efectiva es esencial para el éxito y ejecución de cualquier proyecto de gestión.

Tomar suficiente tiempo de planificación asegura estar observando los problemas o las oportunidades en los que vale la pena invertir una buena gestión (17).

- **Etapa 1.2 Identificar el Negocio:** Esta etapa se centra en la compresión de los objetivos del negocio y los requisitos desde una perspectiva del mismo, a continuación, se convierte este conocimiento en una definición de una solución y un plan preliminar para lograr los objetivos de calidad de datos del negocio.

  En esta etapa se describe todo lo que se conoce sobre el negocio su misión, visión, objetivos etc. (17).

- **FASE 2. ANÁLISIS DE LA INFORMACIÓN.**

Analizar la información del negocio proporciona una base de entendimiento que se utiliza en todo el proyecto asegurando que se están evaluando los datos de relevancia asociados a los problemas de negocios.

Proporciona una comprensión de los requisitos y especificaciones contra el que la calidad de datos se compara.

Permite obtener un contexto para entender los resultados de las evaluaciones de datos y ayuda al análisis de las causas origen. Cuanto más se entienda el contexto y entorno que afectan a los datos, mejor se entiende lo que se va a evaluar de los datos (17).

Además esta fase proporciona una comprensión de los procesos, personas, y tecnología que afecta la calidad de los datos.

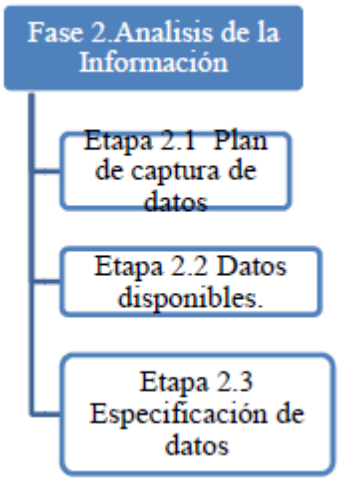


Figura II. 5. Etapas de la Fase 2

Fuente: Propuesta Metodológica SII-ESPOCH

- **Etapa 2.1 Plan de captura de datos:** La captura de los datos se refiere a cualquier extracción de ellos o acceder a ellos de alguna manera (por ejemplo, a través de una conexión directa a una base de datos) (17).

  Desarrollar el plan de captura de datos incluye:

  - Datos, método de acceso y las herramientas.

  - El formato de salida (por ejemplo, extraer a un archivo plano, copiar tablas de un servidor de pruebas.).

- **Etapa 2.2 Datos disponibles:** Esta etapa nos permite conocer el proceso de manipulación de datos que tiene el negocio y las personas que son responsables de los mismos (17).

- **Etapa 2.3 Especificación de datos:** La especificación de los datos mide la existencia y documentación de estándares de datos, modelos de datos, metadatos, datos de referencias (17).

- **FASE 3. Evaluación y Análisis Inicial de la Calidad De Datos**

Se han introducido a la calidad de datos dimensiones, aspectos o características de la información y una forma de clasificar la calidad de la información y necesidades de datos.

Las dimensiones se utilizan para definir, medir, y gestionar la calidad de los datos y de la información el beneficio más gratificante de la evaluación cualitativa de los datos se concreta en la evidencia de los problemas que subyacen en el negocio, problemas identificados en la primera fase. Los resultados de la evaluación también proporcionan información de referencia necesaria para investigar las causas de origen, corregir errores de datos, y evitar los errores de datos en el futuro (17).

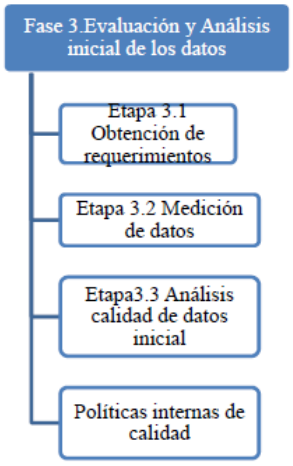


Figura II. 6. Etapas de la Fase 3

Fuente: Propuesta Metodológica SII-ESPOCH

- **Etapa 3.1 Obtención de Requerimientos:** Antes de realizar la medición de datos del negocio es necesario contar con los requerimientos claros y definidos para que en la medición centrarse en los requerimientos planteados (17).
- **Etapa 3.2 Medición de datos:** El proceso de perfilado de datos examina los datos existentes en el negocio y recopila información sobre los mismos (17).
- **Etapa 3.3 Análisis de la calidad de datos Inicial:** Una evaluación inicial es el primer conjunto de prueba realizadas a los que las evaluaciones posteriores se pueden comparar (17).

- **FASE 4. LIMPIEZA DE DATOS**

Después de tener una visión clara de la calidad actual de los datos podemos empezar ya con la limpieza de datos. La corrección en la información y en el proceso de mejora de la calidad de datos (17).

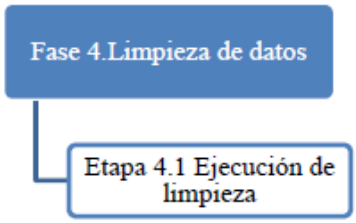


Figura II. 7. Etapas de la Fase 4

Fuente: Propuesta Metodológica SII-ESPOCH

- **Etapa 4.1 Ejecutar Limpieza:** El proceso de ejecución de limpieza de datos permite obtener finalmente datos útiles y actualizados lo que permitirá entender mejor el entorno de negocio, permitiendo maximizar los beneficios (17).

- **FASE 5. EVALUACIÓN Y ANÁLISIS FINAL DE LA CALIDAD DE DATOS**

La evaluación de la calidad de datos final proporciona información vital para indicar el estado final de calidad en los datos analizados y corregidos (17).

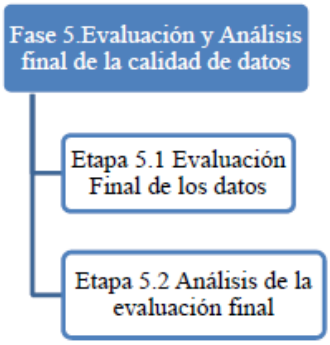


Figura II. 8. Etapas de la fase 5

Fuente: Propuesta Metodológica SII-ESPOCH

- **Etapa 5.1 Evaluación Final de los datos:** Al haber desarrollado una limpieza de los datos, esta etapa permite realizar un análisis de los mismos, lo que servirá para determinar los beneficios para el negocio y saber si se cumplieron con las necesidades del mismo (17).
- **Etapa 5.2Análisis de la evaluación final:** Esta etapa define el estado final en los cuales se entregan los datos después de todos los procesos realizados

determinando con esto como se aporta con el negocio y cuál es el beneficio obtenido (17).

- **FASE 6. MEJORAMIENTO Y PREVENCION.**

La corrección de errores en los datos es un importante avance en el proceso de mejora de la calidad de información y los datos. Sin embargo, para la mejora continua es importante no sólo corregir los datos actuales sino también prevenir los futuros (17).

Este es un punto crítico en la gestión de calidad de datos en el que la comunicación es clave para asegurar que las recomendaciones finales se apliquen y para prevenir cometer futuros errores (17).

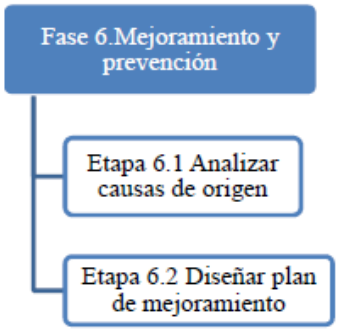


Figura II. 9. Etapas de la Fase 6

Fuente: Propuesta Metodológica SII-ESPOCH

- **Etapa 6.1 Analizar causas de origen:** Todos los problemas que surgen a partir de calidad de datos requieren diferentes niveles de tiempo, dinero, y recursos humanos. Hay una tendencia a saltar a una solución que parece ser más

conveniente para hacer frente con rapidez a una situación es el análisis de las causas de origen el cual ve en todo lo posible las causas de un problema, asunto o condición para determinar su causa real. A menudo tiempo y esfuerzo se gastan en el tratamiento de los síntomas de un problema sin la determinación de sus causas reales, y evitar que el problema vuelva a ocurrir, es por esta razón que esta etapa tiene este objetivo (17).

- **Etapa 6.2 Diseñar plan de mejoramiento:** Esta etapa se enfoca en poner en práctica soluciones adecuadas que aborden las causas fundamentales de los problemas de calidad de los datos (17).

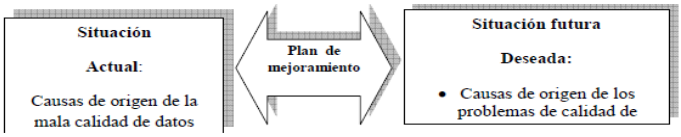


Figura II. 10. Plan de Mejoramiento.

Fuente: Propuesta Metodológica SII-ESPOCH

- **FASE 7. SEGUIMIENTO Y CONTROL.**

Realizar un seguimiento y control es fundamental para la utilización eficaz de los datos.

Esto permite comprobar constantemente la calidad de los datos y se identifican problemas nuevos o pérdidas (17).

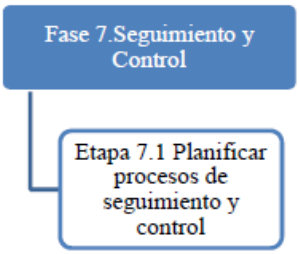


Figura II. 11. Etapas de la Fase 7.

Fuente: Propuesta Metodológica SII-ESPOCH

- **Etapa 7.1 Diseñar plan de seguimiento y control:** En esta etapa se debe implementar controles constantes de calidad (17).

2.8. DIMENSIONES DE LA CALIDAD DE DATOS

El concepto de una dimensión evoca pensamientos de medición, y eso es exactamente lo que se quiere decir cuando se utiliza el término en el contexto de la calidad de los datos. Una dimensión de la calidad de datos describe un contexto y un marco de referencia para la medición junto con las unidades de medida que sugerimos (18).

Las dimensiones aquí propuestas se basan en los principios fundamentales de calidad de datos. Diferentes dimensiones están destinadas a representar diferentes aspectos medibles de la calidad de los datos y se utilizan en la caracterización de relevancia a través de un conjunto de dominios de aplicación para vigilar contra el estándar de organización especificado de calidad de los datos. Una vez que se establece un método para la medición de una dimensión, el analista puede utilizar las mediciones para evaluar el rendimiento de calidad de datos en los diferentes niveles de la jerarquía operacional (18).

Las mediciones recogidas pueden rellenar un tablero que indica la línea de negocio en general y después del rendimiento empresarial enrollado, con respecto a las expectativas de los usuarios de negocios. Esto le da a cada grupo dentro de la organización la libertad de introducir sus propias dimensiones con características personalizadas (18).

Aunque es posible describir un conjunto diverso de las mediciones, el valor de uso de las métricas es ser capaz de sopesar los aspectos más críticos de un proceso que necesita ser medido contra aquellos aspectos que son susceptibles de ser medidos. En otras palabras, es bueno para definir las métricas basadas en mediciones cuantificables, pero no dejes que eso te desanime a las evaluaciones cualitativas. Pero recuerde que cuando las mediciones son subjetivas, existe el riesgo de que puedan ser mal interpretadas, por lo que esos tipos de métricas deben utilizarse con moderación (18).

2.8.1. Categorización de las Dimensiones

Dimensiones de la calidad de los datos pueden ser lógicamente clasificadas y ordenadas en una jerarquía para facilitar la gobernanza, la implementación de la tecnología, la definición de los procesos operativos, cumplimiento y presentación de informes. Reglas asociadas a diferentes dimensiones se pueden aplicar a diferentes aspectos de los datos de la organización. Al nivel más simple, hay dimensiones que son intrínsecos a los valores que componen un conjunto de datos. Reglas más complejas son el resultado de relaciones esperadas que se producen a nivel de registro, a continuación, en el conjunto de datos, seguido por el nivel de aplicación. Desde un punto de vista diferente, hay otros tipos de

reglas que se pueden utilizar para gobernar el cumplimiento con la política de información de negocios (18).

Las dimensiones que se describen en este capítulo abarcan los aspectos intrínsecos de la calidad de los datos que son de aplicación general dentro de una empresa. En la Figura II. 13 se destacan los diferentes niveles en los que se pueden especificar las dimensiones de calidad de datos y normas de calidad de datos. La imagen muestra los niveles de una jerarquía para indicar la dependencia de cada nivel en el manejo de información de calidad asociada a los niveles inferiores (18).

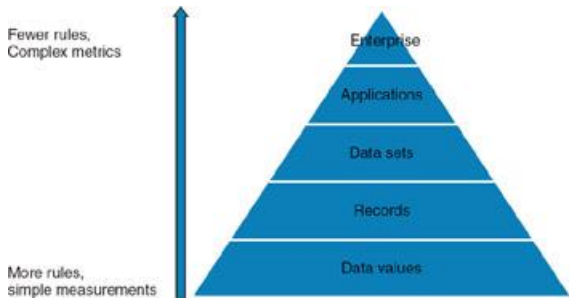


Figura II. 12. Jerarquía de las dimensiones de calidad de datos.

Fuente: http://common.books24x7.com/toc.aspx?bookid=40139.

Aunque la literatura de calidad de datos propone diferentes tipos de dimensiones el practicante se limita pragmáticamente a las dimensiones que se miden factible y razonablemente socializados en toda la empresa. Sólo nos centraremos en los detalles de las dimensiones de calidad de datos en los niveles físicos y operacionales, y las dimensiones discretamente cuantificables se presentan (18).

2.8.1.1.Dimensiones intrínsecas

Podemos considerar las medidas asociadas a los datos de los propios valores fuera de cualquier forma de asociación con un elemento de datos o un registro como " intrínseca". Las dimensiones intrínsecas se refieren a los datos de los valores propios de un conjunto de datos específicos o el contexto del modelo. Por ejemplo, especificar un rango válido de temperaturas (por ejemplo, -50 a 110 grados Fahrenheit) es intrínseco al valor, no importa donde se utiliza. Como otro ejemplo, insistiendo en que todos los valores de datos que representan los números de teléfono se ajustan a las normas definidas en el formato de Plan de Numeración de América del Norte define una regla de calidad de los datos asociados a la sintaxis del valor (18).

2.8.1.2.**Dimensiones contextuales**

Las medidas que se centran en la consistencia o validez de un elemento de datos en relación con otros elementos de datos o de un registro a otro registro pueden ser referidas como “contextuales", ya que dependen del contexto. Dimensiones contextuales dependen de diversas políticas de negocio que se implementan como reglas de negocio dentro de los sistemas y procesos. Sin embargo, a pesar de las políticas de información (tales como las relativas a la seguridad o la privacidad) son una de las principales fuentes de las afirmaciones de calidad de datos. Sin embargo, algunas de las dimensiones de calidad de datos tienen implicaciones de gobernabilidad. Como un ejemplo, un requisito para la asignación de identificadores que hacen referencia de forma única una entidad individual es una política de información; esto se traduce a las normas de calidad de datos respecto a la identificación única, el anonimato identificador, no identificable, y así sucesivamente (18).

2.8.1.3. Dimensiones cualitativas

Una tercera categoría incluye las dimensiones que podrían ser considerados "cualitativa", y estos pueden reflejar la síntesis de las medidas asociadas con las dimensiones intrínsecas y contextuales. En última instancia, se trata de la combinación de las medidas asociadas a la conformidad con el máximo nivel de calidad de los datos según lo previsto por la especificación de las políticas de información (18).

2.8.1.4. Clasificación de Dimensiones

Las clasificaciones de las dimensiones prácticas de calidad de datos son los siguientes:

- Precisión
- Linaje
- Consistencia estructural
- Coherencia semántica
- Completitud
- Consistencia
- Oportunidad
- Sensatez
- Identificabilidad

Las relaciones entre las dimensiones se muestran en la Figura II. 13.

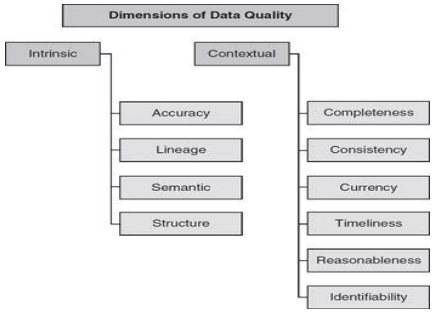


Figura II. 13. Relación entre dimensiones.

Fuente: Propuesta Metodológica SII-ESPOCH

2.8.1.5. Dimensiones prácticas de la calidad de los datos.

La consideración principal de aislar dimensiones críticas de la calidad de datos es proporcionar indicadores universales para evaluar el nivel de calidad de los datos en diferentes contextos operacionales o analíticos. Con este enfoque, los administradores de datos pueden trabajar con los gerentes de línea de negocio a:

Definir reglas de calidad de datos que representan las expectativas de validez.

Determinar los umbrales mínimos de aceptabilidad (18).

2.8.1.6. Medir los umbrales de aceptabilidad.

En otras palabras, las afirmaciones que corresponden a estos umbrales se pueden transformar en reglas que se utilizan para el seguimiento del grado en que mide los niveles de calidad se encuentran definidos y consensuados con las expectativas de negocio. A su vez, las métricas que corresponden a esas medidas de conformidad proporcionan

información sobre el examen de las causas fundamentales que impiden que los niveles de calidad de cumplir esas expectativas (18).

2.8.1.7. Descripción de las dimensiones de calidad de datos

Estas áreas generales de la concentración de enmarcar la discusión para la identificación de los aspectos medibles de la calidad de datos, pero los detalles tienen que basarse en las necesidades de negocio encuestados. Como una manera de organizar las medidas previstas, al ensamblar la lista de las dimensiones relevantes de calidad de datos para su organización, vale la pena considerar:

- La característica de la dimensión, que es el aspecto de la dimensión de alto nivel que se está midiendo.
- Los criterios que se enumeran los aspectos específicos de la dimensión que se desea medir.
- La métrica, que es la medida cuantificable - efectivamente la forma en que se mide cada criterio.
- El umbral de la conformidad, que es el nivel de medida que indica la conformación (18).

- **Dimensiones intrínsecas**

Las dimensiones intrínsecas se centran en los valores de los propios datos, sin necesidad de evaluar el contexto de esos valores. Estas dimensiones se caracteriza por la estructura, formatos, significados, y la enumeración de los dominios de datos esencialmente la calidad de los metadatos de la organización y cómo se utiliza (18).

- **Precisión:** La precisión es una de las dimensiones más difíciles de evaluar debido a que se refiere al grado en el que los valores de datos están de acuerdo con una fuente identificada de la información correcta. Puede haber muchas fuentes potenciales de información correcta, y algunos ejemplos incluyen una base de datos de registro, un conjunto similar, corroboración de los valores de datos de otra tabla, los valores calculados de forma dinámica, o tal vez los resultados de un proceso manual. En muchos casos no hay una fuente definitiva de información correcta (18).
- **Linaje:** La confiabilidad de los datos es de vital importancia para todos los participantes en la empresa. Un aspecto de la medida de la confiabilidad es la capacidad de identificar el origen de cualquier dato nuevo o actualizado. Además, la documentación de los flujos de información permite un mejor análisis de la causa raíz. Por lo tanto, una dimensión que mide las fuentes históricas de datos, llamado linaje, es valiosa en la evaluación general (18).

- **Consistencia estructural**

Consistencia estructural se refiere a la coherencia en la representación de valores de atributos similares, tanto dentro de un mismo conjunto de datos ya través de los modelos de datos asociados con tablas relacionadas. Coherencia estructural que caracteriza a la diligencia de los modeladores de base de datos, administradores y administradores en asegurarse de que los atributos similares se escriben encarecidamente utilizar paradigmas de representación bien definidos (18).

- **La coherencia semántica**

En general, todas las empresas se componen de un número de participantes, cada uno de los cuales rige sus propios entornos de datos. Este gobierno interno refleja cómo se utiliza la información dentro de sus procesos de negocio internos. Pero en todas las organizaciones, los participantes a menudo tienen que intercambiar o compartir datos, especialmente a medida que más organizaciones adoptan los repositorios de datos maestros "única fuente" Sin embargo, como el desarrollo de arquitecturas de información en gran medida ha sido impulsada a nivel de aplicación, no es probable que haya variaciones en cómo los diferentes individuos (y sus aplicaciones administradas) entender los significados de los términos comerciales de uso común. Así que a los efectos del intercambio de información es necesario para un acuerdo sobre el significado de los términos de negocios como los datos se mueven entre las aplicaciones (18).

Consistencia semántica se refiere a la consistencia de las definiciones entre los atributos dentro de un modelo de datos, así como los atributos de nombre similar en los diferentes conjuntos de datos de empresas, y que caracteriza el grado en que los objetos de datos similares comparten nombres coherentes y significados. Un aspecto de la coherencia semántica implica el significado de los atributos con nombres similares en diferentes conjuntos de datos. Los significados de estos nombres de atributo deben distinguirse, o los atributos se deben asignar nombres diferentes. La conformidad con las normas de datos definidos externamente proporciona un cierto nivel de la política para esta dimensión (18).

- **Dimensiones contextuales**

Las dimensiones contextuales proporcionan una forma para que el analista revisa la conformidad con las expectativas de calidad de los datos asociados con el número de elementos de datos están relacionados entre sí (18).

- **Integridad**

  Un modelo de datos no debe incluir información extra, ni debe su despliegue ser la falta de valores de datos pertinentes. Los datos no usados, por la naturaleza de su ser ignorados, tenderán hacia la entropía, lo que lleva a los niveles bajos de calidad de datos, junto con la introducción de un problema de mantener la coherencia de los datos a través de cualquier copia o réplica. Integridad se refiere a la expectativa de que se espera que ciertos atributos que han asignado valores en un conjunto de datos. Reglas de integridad pueden ser asignadas a un sistema en tres niveles de restricciones de datos:

  - Los atributos obligatorios requieren un valor.
  - Atributos opcionales, que pueden tener un valor (posiblemente en circunstancias específicas).
  - Atributos que no se aplican (como el nombre de soltero de un solo hombre), que no pueden tener un valor. (18).

- **Consistencia:**

En cualquier entorno empresarial, la consistencia es relevante para los diferentes niveles de la jerarquía de datos dentro de las tablas, bases de datos, a través de diferentes aplicaciones, así como con los datos suministrados externamente. En virtud de la creciente tendencia a la consolidación y el intercambio de datos a través de las líneas de negocio, puede ser descubierto inconsistencias en diferentes conjuntos de datos que pueden haber eludido el escrutinio en el pasado. Políticas y procedimientos para reportar inconsistencias a los propietarios de las fuentes de datos que contribuyen deben ser definidos para asegurar la consistencia medible entre los participantes (18).

- **Puntualidad**

La puntualidad se refiere a la expectativa de tiempo para la accesibilidad de la información. La puntualidad puede ser medido como el tiempo entre cuando se espera que la información y cuando está fácilmente disponible para su uso (18).

- **Sensatez**

Afirmaciones generales asociados con las expectativas de consistencia o razonabilidad de los valores, ya sea en el contexto de los datos existentes o más de una serie de tiempo, se incluyen en esta dimensión (18).

- **Identificabilidad**

  Identificabilidad se refiere a la asignación única de nombres y la representación de objetos conceptuales básicos, así como la capacidad de vincular instancias de datos que contienen datos de entidad juntos basados en la identificación de los valores del atributo (18).

- **Confiabilidad**

  Confiabilidad es la probabilidad de que un componente o sistema desempeñe satisfactoriamente la función para la que fue creado durante un periodo establecido y bajo condiciones de operación establecidos. La confiabilidad es calidad en el tiempo (19).

- **Dimensiones cualitativas**

Podríamos caracterizar dimensiones adicionales para los que la capacidad de obtener mediciones cuantitativas es menos clara. Sin embargo, proporcionar dimensiones cualitativas permite evaluar un orden superior de supervisión - la revisión de lo bien que la información cumple con las expectativas y necesidades definidas (18).

Uno de los retos más importantes en la implementación de un programa de calidad de los datos es la transición desde el desarrollo del programa a nivel conceptual y poniendo los procesos establecidos que guían la conformidad con el programa. Una razón es que la transición de una organización reactiva a una proactiva requiere un cierto grado de disciplina en lo que respecta a la identificación de los objetivos de desempeño y monitoreo

permanente. La definición y la especificación de la métrica en el inicio de un programa en un documento de referencia y haciendo caso omiso de ellos en la práctica conducirá al fracaso del programa (18).

Con el fin de integrar la calidad de datos de los entornos operativos, las calificaciones se pueden definir que el informe sobre la calidad del dato. Estas calificaciones de calidad pueden ser almacenadas como atributos retrospectivos de sus elementos de datos asociados, por ejemplo, en el nivel más bajo, una dimensión calidad de los datos de precisión se puede especificar con un atributo "precisión" (18).

Históricamente el concepto de "aptitud para el uso" ha reflejado la principal medida de calidad de datos, aunque muchas de las características son subjetivas y son difíciles de medir cuantitativamente. Sin embargo, es razonable incorporar estas características y criterios en términos de posibles indicadores clave de rendimiento de calidad de datos (18).

CAPÍTULO III

ANÁLISIS COMPARATIVO Y DETERMINACION DE LAS MEJORES HERRAMIENTAS SELECIONADAS

En este capítulo se efectua un análisis de cada herramienta partiendo de un prototipo preparado con el fin de conocer las ventajas y desventajas que presenta cada una de las herramientas y seleccionar las más adecuadas para trabajar con un almacén de datos.

3.1. ANÁLISIS DE LOS DATOS DEL ESCENARIO DE PRUEBA.

Los resultados del análisis de los datos del escenario de prueba se detallan en el ANEXO 1.

3.2. ANÁLISIS DE LA CALIDAD DE LOS DATOS.

Todos los problemas presentados en el ANEXO 1 afectan determinadas dimensiones de calidad, las cuales se muestras a continuación como se solventaron para mejorar los datos de manera efectiva.

Tabla III. I Dimensiones de Calidad

Criterio	Parámetro	Metrica	Umbral de conformidad
Consitencia.	Precisión.	Numero de registros nulos o blancos.	=100%
	Valores aceptables.	Cantida de registos invalidos.	>=95%
Integridad	Duplicidad.	Cantidad de registros duplicados.	>=95%
Confiabilidad	Confianza.	(0.6) INTEGRIDAD + (0.4) CONSISTENCIA	>=95%

Elaborado por: Investigador

El umbral de conformidad que muestra la Tabla III.I ayuda ha establecer cuando un parámetro está dentro del nivel aceptable en la medición de la calidad de los datos, para ello se debe igualar este umbral de conformidad o superarlo como se muestra en la Tabla III. II anterior.

Para medir los criterios en los datos de cada tabla de la base de datos se tomarán en cuenta las siguientes fórmulas:

Parámetro de Precisión.

Porcentaje de registros no nulos ni blancos dentro de un campo en una tabla.

Parámetro de Valores Aceptables.

Porcentaje de registros con valores adecuados dentro de un campo en una tabla.

Parámetro de Duplicidad

Porcentaje de registros únicos dentro de una tabla.

Parámetro de Confianza.

Para el análisis de este parámetro se utilizó una formula planteada por el tesista debido a que no se encontró ninguna fórmula y se la planteo en base a la dimensión cualitativa de la confiabilidad partiendo de la teoría encontrada que se detalla en la página 74, la fórmula se identificó de acuerdo con el desempeño de una funcion satisfactoria de la integridad de los datos y la consistencia que son dimensiones las cuales si pueden ser medidas y son tomadas en cuenta en este estudio.

Donde se tiene la siguiente fórmula:

$$\mathbf{CFD} = (0.6)\ \mathbf{INT} + (0.4)\ \mathbf{CON}$$

Dónde:

CFD = CONFIABILIDAD

INT = INTEGRIDAD

CON = CONSISTENCIA

La consistencia se la puede obtener del promedio de valores obtenidos entre la precisión y los valores válidos partiendo que la consistencia es la unión y relación adecuada de todas las partes que forman un todo, como la precisión y valores válidos influyen en la consitencia por lo cual se porpone la siguiente fórmula.

$$\mathbf{CON} = (\mathbf{P}+\mathbf{VA})/2$$

Dónde:

P = PRECISIÓN

VA= VALORES ACEPTABLES

3.2.1. Descripción del proceso de análisis del indicador de confiabilidad de datos

El indicador de confiabilidad constituye una parte del parámetro de limpieza de datos que sera utilizado para la comparción de las herramientas.

Para la evaluación de este indicador en cada una de las herramientas se realizó el siguiente procedimiento:

- Para el análisis de los datos se utilizó la herramienta DataCleaner para obtener la linea base de resultados de la calidad de datos.
- Se realizó la limpieza de la datos con cada una de las herramientas a comparar Oracle Data Quality, SQL Power e Infromatica en escenarios similares.
- Una vez realizada la limpieza se procedio a envaluar los resultados nuevamente con la herramienta DataCleaner para cada uno de los casos.
- Se realizó la consolidación de los datos de la confiabilidad que servirá para el posterior análisis comparativo de las herramientas.

3.2.2. Análisis de la tabla TPM_CUS:

En los datos encontrados en la tabla TPM_CUS se encontraron:

- **Parámetro de Precisión:**

Tabla III. III. Parámetro s de Precisión

TPM_CUS			
COLUMNA	**VALORES NULOS**	**PRECISIÓN**	**CONFORMIDAD**
CODIGO	0%	100%	ACEPTABLE
CODIGO_CLIENTE	0%	100%	ACEPTABLE
NOMBRES	0%	100%	ACEPTABLE

IDENTIFICACION	0%	100%	ACEPTABLE
TIPO_IDENTIFICACION	0%	100%	ACEPTABLE
DIRECCION	2,568%	97,432%	ACEPTABLE
EMAIL	95,118%	4,882%	NO ACEPTABLE
TELEFONO	0,382%	99,618%	ACEPTABLE
CELULAR	39,794%	60,206%	NO ACEPTABLE

Elaborado por: Investigador

En la Tabla III.II muestra que el parámetro de la precisión se esta cumpliendo para muchas de las columnas de la tabla TPM_CUS excepto para las columnas EMAIL y CELULAR las cuales no cumplen con el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

- **Parámetro de Valores Aceptables:**

Tabla III. IV Parámetro de Valores Aceptables

TPM_CUS			
COLUMNA	**CARACTERES INVÁLIDOS**	**VALORES ACEPTABLES**	**CONFORMIDAD**
CODIGO	0%	100%	ACEPTABLE
CODIGO_CLIENTE	0%	100%	ACEPTABLE
NOMBRES	0%	100%	ACEPTABLE
IDENTIFICACION	0,073%	99,927%	ACEPTABLE
TIPO_IDENTIFICACION	0%	100%	ACEPTABLE
DIRECCION	0%	100%	ACEPTABLE
EMAIL	0%	100%	ACEPTABLE
TELEFONO	2,039%	97,961%	ACEPTABLE
CELULAR	0%	100%	ACEPTABLE

Elaborado por: Investigador

En la Tabla III.II muestra que el parámetro de valores aceptables se esta cumpliendo para todas las columnas de la tabla TPM_CUS que es el 95% o superior como se establece y se detalla en la Tabla III.I.

- **Parámetro de Duplicidad:**

Tabla III. V Parámetro de Duplicidad

TPM_CUS			
COLUMNA	**DUPLICADOS**	**DUPLICIDAD**	**CONFORMIDAD**
CODIGO	0%	100%	ACEPTABLE
CODIGO_CLIENTE	18,015%	81,985%	NO ACEPTABLE
NOMBRES	18,015%	81,985%	NO ACEPTABLE
IDENTIFICACION	18,015%	81,985%	NO ACEPTABLE
TIPO_IDENTIFICACION	18,015%	81,985%	NO ACEPTABLE
DIRECCION	9,250%	90,75%	NO ACEPTABLE
EMAIL	5,032%	94,968%	NO ACEPTABLE
TELEFONO	1,355%	98,645%	ACEPTABLE
CELULAR	0,576%	99,424%	ACEPTABLE

Elaborado por: Investigador

En la Tabla III.IV muestra que el parámetro de la duplicidad se esta cumpliendo para algunas de las columnas de la tabla TPM_CUS excepto para las columnas CODIGO_CLIENTE, NOMBRES, IDENTIFICACION, TIPO_IDENTIFICACION, DIRECCION y EMAIL las cuales no cumple con el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

- **Parámetro de Confianza:**

Tabla III. VI. Parámetro de Confianza

TPM_CUS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (PRESICIÓN + VALORES ACEPTABLES)/2	**CONFIABILIDAD** (0,6)INTEGRIDAD + (0,4)CONSITENCIA
CODIGO	100%	100%	100%
CODIGO_CLIENTE	81,985%	100%	89,191%
NOMBRES	81,985%	100%	89,191%
IDENTIFICACION	81,985%	100%	89,191%
TIPO_IDENTIFICACION	81,985%	99,9635%	89,1764%
DIRECCION	90,75%	98,716%	93,9364%
EMAIL	94,968%	52,441%	77,9572%
TELEFONO	98,645%	98,7895%	98,7028%
CELULAR	99,424%	80,103%	91,6956%

Elaborado por: Investigador

En la Tabla III.V muestra que el parámetro de la confiabilidad solo se esta cumpliendo para la columna CODIGO de la tabla TPM_CUS excepto para las otras columnas las cuales no cumple con el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

Resultados finales.

Tabla III. VII Resultados Finales

TPM_CUS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	90,19%
CONSISTENCIA	92,22%
CONFIABILIDAD	91%

Elaborado por: Investigador

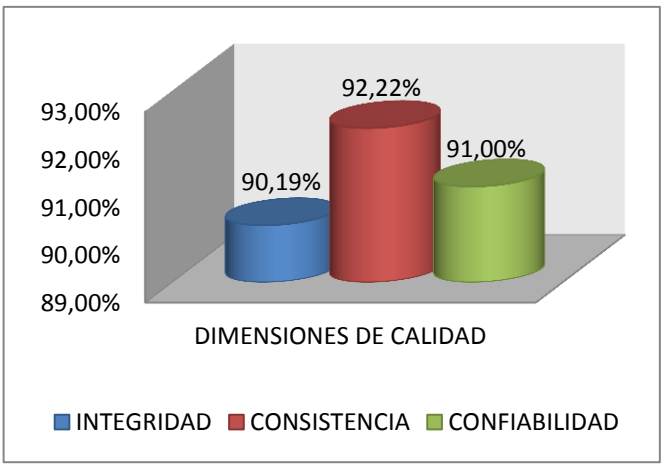


Figura III. 1 Resultados Finales

Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla TPM_CUS se puede evidenciar que ninguna de las dimensiones examinadas ha obtenido el nivel aceptable para indicar que los datos son de calidad como muestra en la Tabla III.VI debido a que las dimensiones de integridad, consistencia y confiabilidad se ubican por debajo del 95% que es el umbral de conformidad el cual se detalla en la Tabla III.I por esta razón se utilizará las herramientas antes determinadas con el fin de garantizar el aumento en la calidad de los datos con las funciones que estas puedan ofrecer.

3.2.3. Análisis de la tabla TPM_TRAN:

En los datos encontrados en la tabla TPM_TRAN se encontraron:

- **Parámetro de Precisión:**

Tabla III. VIII. Parámetro de Precisión

TPM_TRAN			
COLUMNA	**VALORES NULOS**	**PRECISIÓN**	**CONFORMIDAD**
CODIGO	0%	100%	ACEPTABLE
CODIGO_CLIENTE	0%	100%	ACEPTABLE
INL_AMT	81.503%	18,497%	NO ACEPTABLE

Elaborado por: Investigador

En la Tabla III. IX**.** muestra que el parámetro de la precisión se esta cumpliendo para las dos primeras columnas de la tabla TPM_TRAN excepto para las columnas INL_AMT la cual no cumple con el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

- **Parámetro de Valores Aceptables:**

Tabla III. X. Parámetro de Valores Aceptables

TPM_TRAN			
COLUMNA	**CARACTERES INVÁLIDOS**	**VALORES ACEPTABLES.**	**CONFORMIDAD**
CODIGO	0%	100%	ACEPTABLE
CODIGO_CLIENTE	0%	100%	ACEPTABLE
INL_AMT	81.503%	18,497%	NO ACEPTABLE

Elaborado por: Investigador

En la Tabla III.8 muestra que el parámetro de la precisión se esta cumpliendo para las dos primeras columnas de la tabla TPM_TRAN excepto para las columnas INL_AMT las cuales no cumple con el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.1.

- **Parámetro de Duplicidad:**

Tabla III. XI Parámetro de Duplicidad

TPM_TRAN			
COLUMNA	**DUCPLICADOS**	**DUPLICIDAD**	**CONFORMIDAD**
CODIGO	0%	100%	ACEPTABLE
CODIGO_CLIENTE	8.899%	91,101%	NO ACEPTABLE
INL_AMT	5.024%	94,976%	NO ACEPTABLE

Elaborado por: Investigador

En la Tabla III.IX muestra que el parámetro de la duplicidad se esta cumpliendo solo para la columna CODIGO de la tabla TPM_TRAN excepto para las columnas CODIGO_CLIENTE e INL_AMT la cual no cumple con el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

- **Parámetro de Confianza:**

Tabla III. XII Parámetro de Confianza

TPM_TRAN			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (PRESICIÓN + VALORES ACEPTABLES)/2	**CONFIABILIDAD** (0,6)INTEGRIDAD + (0,4)CONSITENCIA
CODIGO	100%	100%	100%
CODIGO_CLIENTE	91,101%	100%	94,6606%
INL_AMT	94,976%	81,503%	89,5868 %

Elaborado por: Investigador

En la Tabla III.X muestra que el parámetro de la confiabilidad se esta cumpliendo solo para la columna CODIGO de la tabla TPM_TRAN excepto para las columnas CODIGO_CLIENTE e INL_AMT las cuales no cumple con el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

Resultados finales.

Tabla III. XIII Resultados Finales

TPM_ TRAN	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	95,36%
CONSISTENCIA	93,83%
CONFIABILIDAD	94,75%

Elaborado por: Investigador

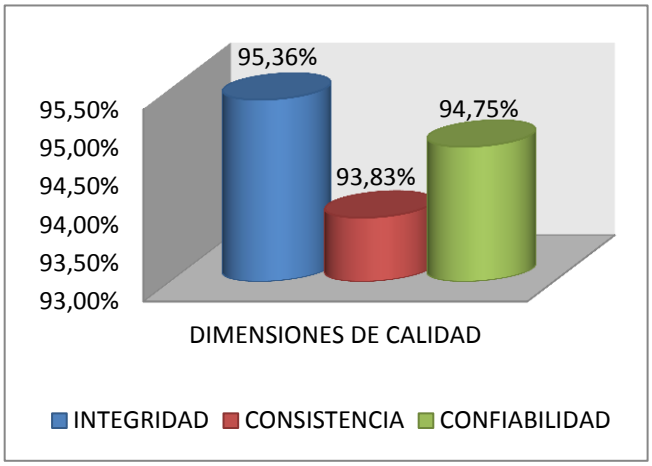


Figura III. 2 Resultados Finales

Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla TPM_TRAN se puede evidenciar que solo la dimensión de integridad ha obtenido el nivel de conformidad establecido con un 95,38% para indicar que los datos son de calidad como se muestra en la Tabla III.XI, además las dimensiones de consistencia y confiabilidad se ubican por debajo del 95% que es el umbral de conformidad el cual se detalla en la Tabla III.I por esta razón se utilizará cada una de las herramientas antes mensionadas con el fin de garantizar el aumento de la clidad de los datos con el uso de las herraientas y las funciones que estas puedan ofrecer.

3.3. LIMPIEZA DE LOS DATOS UTILIZANDO LAS DIFERENTES HERRAMIENTAS.

Lasa herramientas utilizadas para la limpieza de datos son las que se detallarón en el capítulo II tales como: Oracle Data Integrator e Informatica en herramientas propietarias y en herramientas libres DataCleaner y SQL Power.

3.3.1. SQL POWER

Limpieza: Los pasos que seguimos son los siguientes:

1. Creamos un nueva carpeta de proyectos
2. Dentro de la carpeta creada, creamos un proyecto de limpieza “Cleasing Project”
3. Seleccionamos la carpeta de transformaciones y creamos una nueva transformación
4. Utilizamos los objetos que la herramienta nos provee para la limpieza

La Figura III. 3 muestra los componentes que se utilizarón para la limpieza de los datos.

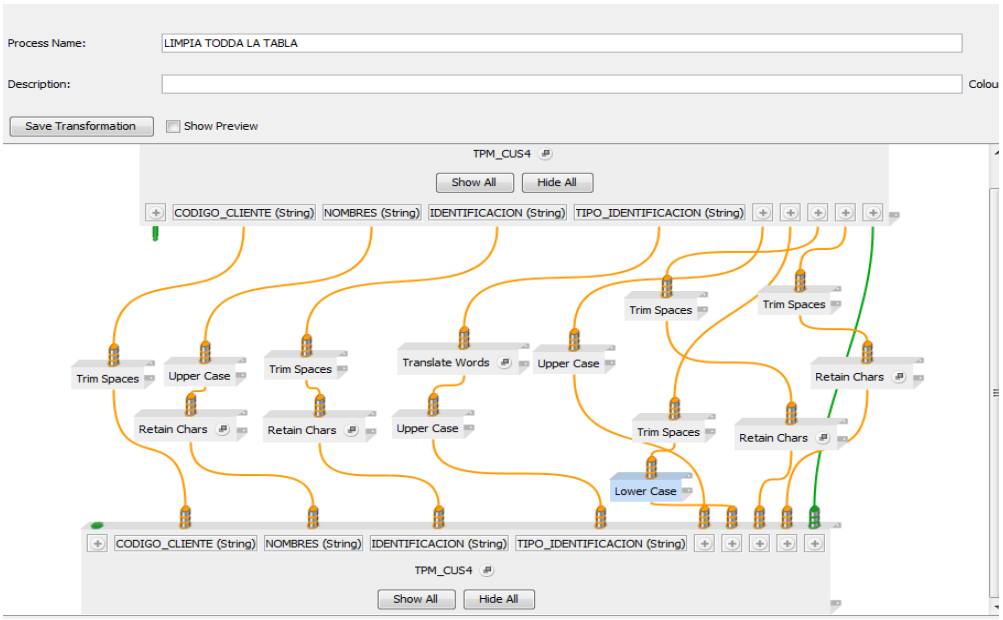


Figura III. 4 Limpieza

Fuente: Investigador

El resultado de la limpieza se puede visualizar de manera mas detallada en el ANEXO 1.

3.3.1.1. Análisis de los resultados

- **Análisis de la tabla TPM_CUS**

  - **Parámetro de Precisión:**

Tabla III. XIV Parámetros de Precisión-Tabla TPM_CUS

TPM_CUS		
COLUMNA	**VALORES NULOS**	**PRECISIÓN**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
NOMBRES	0%	100%
IDENTIFICACION	0%	100%
TIPO_IDENTIFICACION	0%	100%
DIRECCION	0%	100%
EMAIL	0%	100%
TELEFONO	0%	100%
CELULAR	0%	100%

Elaborado por: Investigador

En la Tabla III.XII muestra que el parámetro de la precisión se cumple para todas las columnas de la tabla TPM_CUS las cuales superan el umbral de conformidad que es el 95% o superior como se debe establece y se detalla en la Tabla III.I.

  - **Parámetro de Valores Aceptables:**

Tabla III. XV. Parámetro Valores Aceptables-Tabla TPM_CUS

TPM_CUS		
COLUMNA	**CARACTERES INVÁLIDOS**	**VALORES ACEPTABLES.**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
NOMBRES	0%	100%
IDENTIFICACION	0%	100%
TIPO_IDENTIFICACION	0%	100%
DIRECCION	1,568%	98,432%
EMAIL	5,118%	94,882%

TELEFONO	0,182%	99,9872%
CELULAR	9,794%	90,206%

Elaborado por: Investigador

En la Tabla III.XIII muestra que el parámetro de la valores aceptables se cumple para las cuatro priemeras columnas de la tabla TPM_CUS las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I excepto para las columnas de EMAIL y CELULAR.

- **Parámetro de Duplicidad:**

Tabla III. XVI . Parámetros de Duplicidad -Tabla TPM_CUS

TPM_CUS		
COLUMNA	**DUCPLICADOS**	**DUPLICIDAD**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
NOMBRES	0%	100%
IDENTIFICACION	0%	100%
TIPO_IDENTIFICACION	0%	100%
DIRECCION	0%	100%
EMAIL	0%	100%
TELEFONO	0%	100%
CELULAR	0%	100%

Elaborado por: Investigador

En la Tabla III.XVI muestra que el parámetro de la duplicidad se cumple para todas las columnas de la tabla TPM_CUS las cuales superan el umbral de conformidad que es el 95% o superior como se debe establece y se detalla en la Tabla III.I.

- **Parámetro de Confiabilidad:**

Tabla III. XVII Parámetros de Confiabilidad-Tabla TPM_CUS

TPM_CUS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (PRESICIÓN + VALORES ACEPTABLES)/2	**CONFIABILIDAD** (0,6)INTEGRIDAD + (0,4)CONSITENCIA
CODIGO	100%	100%	100%
CODIGO_CLIENTE	100%	100%	100%
NOMBRES	100%	100%	100%
IDENTIFICACION	100%	100%	100%
TIPO_IDENTIFICACION	100%	100%	100%
DIRECCION	100%	99,216%	99,6864%
EMAIL	100%	97,441%	98,9764 %
TELEFONO	100%	99,909%	99,9636%
CELULAR	100%	95,103%	98,0412%

Elaborado por: Investigador

En la Tabla III.XV muestra que el parámetro de la confiabilidad se cumple para todas las columnas de la tabla TPM_CUS las cuales superan el umbral de conformidad que es el 95% o superior como se debe establece y se detalla en la Tabla III.I.

Resultados finales.

Tabla III. XVIII Resultados Finales

TPM_CUS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	99,07%
CONFIABILIDAD	99,63%

Elaborado por: Investigador

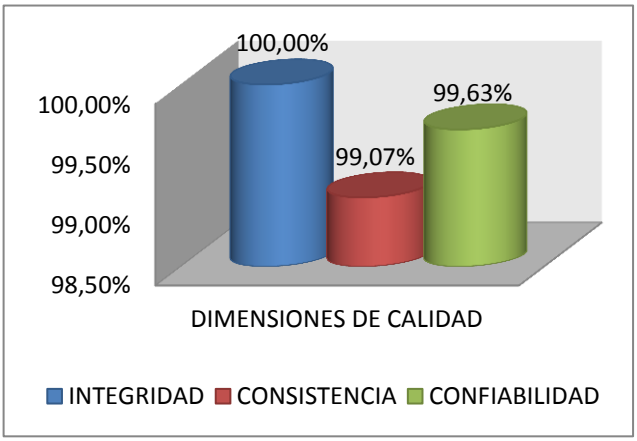


Figura III. 5 Resultados Finales

Fuente: Investigador

Interpretación de resultados finales.

Después de realizar el análisis posterior a la limpieza de los datos utilizando la herramienta SQL Power en la tabla TPM_CUS y observando las Tablas III.VI y III.XVI podemos evidenciar que las dimensiones examinadas han obtenido el nivel aceptable para indicar que los datos han mejorado en su calidad teniendo un porcentaje de 100% para la dimensión de integridad superando al análisis inicial en el cual se obtuvo 90,19% teniendo una optimización del 9.81% en lo que se refiere a la duplicidad de datos, para la dimensión de consistencia se han obtenido un 99.07% superando el análisis previo en el cual se obtuvo 92,22% con una ascenso del 6,85% en lo que se refiere a la precisión y validez de los datos, además el nivel de confiabilidad se ha superado obteniendo 99,63% con la limpieza de los datos en relación al primer análisis en el cual se obtuvo 91% superando esta dimensión en 8,63% con el uso de la herramienta optimizando así los datos de esta tabla.

- **Análisis de la tabla TPM_TRAN:**
  - **Parámetro de Precisión:**

Tabla III. XIX Parámetros de Precisión-Tabla TPM_TRAN

TPM_TRAN		
COLUMNA	**VALORES NULOS**	**PRECISIÓN**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
INL_AMT	0%	100%

Elaborado por: Investigador

En la Tabla III.XVII muestra que el parámetro de la precisión se cumple para todas las columnas de la tabla TPM_TRAN las cuales superan el umbral de conformidad que es el 95% o superior como se debe establece y se detalla en la Tabla III.I.

- **Parámetro de Valores Aceptables:**

Tabla III. XX Parámetros Valores Aceptables -Tabla TPM_TRAN

TPM_TRAN		
COLUMNA	**CARACTERES INVÁLIDOS**	**VALORES ACEPTABLES**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
INL_AMT	11.503%	88,497%

Elaborado por: Investigador

En la Tabla III. XXI muestra que el parámetro de valores aceptables cumple para las dos primeras columnas de la tabla TPM_TRAN excepto la columna INL_AMT las cuales superan el umbral de conformidad que es el 95% o superior como se debe establece y se detalla en la Tabla III.I.

- **Parámetro de Duplicidad:**

Tabla III. XXII Parámetros de Duplicidad -Tabla TPM_TRAN

TPM_TRAN		
COLUMNA	**DUCPLICADOS**	**DUPLICIDAD**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
INL_AMT	0%	100%

Elaborado por: Investigador

En la Tabla III. XXIII muestra que el parámetro de la duplicidad se cumple para todas las columnas de la tabla TPM_TRAN las cuales superan el umbral de conformidad que es el 95% o superior como se debe establece y se detalla en la Tabla III.I.

- **Parámetro de Confiabilidad:**

Tabla III. XXIV Parámetros de Confianza -Tabla TPM_TRAN

TPM_TRAN			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (PRESICIÓN + VALORES ACEPTABLES)/2	**CONFIABILIDAD** (0,6)INTEGRIDAD + (0,4)CONSITENCIA
CODIGO	100%	100%	100%
CODIGO_CLIENTE	100%	100%	100%
INL_AMT	100%	94,2485%	97,6994%

En la Tabla III. XXV muestra que el parámetro de la confiablilidad se cumple para todas las columnas de la tabla TPM_TRAN las cuales superan el umbral de conformidad que es el 95% o superior como se debe establece y se detalla en la Tabla III.I.

Resultados finales.

Tabla III. XXVI Resultados Finales

TPM_CUS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	98,08%
CONFIABILIDAD	99,23%

Elaborado por: Investigador

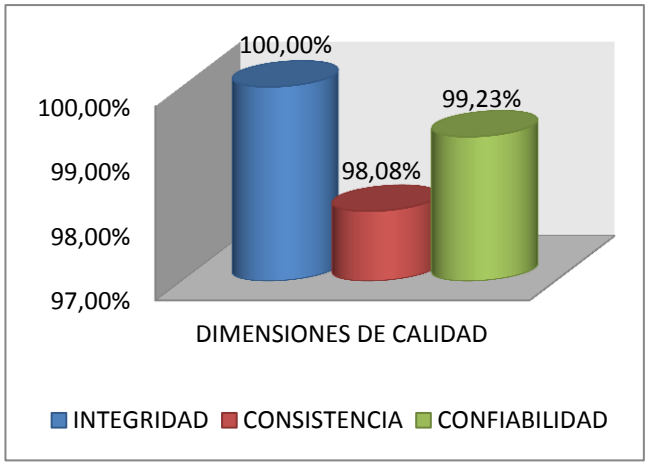


Figura III. 6 Resultados Finales

Fuente: Investigador

Interpretación de resultados finales.

Después de realizar el análisis posterior a la limpieza de los datos utilizando la herramienta SQL Power en la tabla TPM_TRAN y observando las Tablas III.XI y III.XXI podemos evidenciar que las dimensiones examinadas han mejorado en su nivel de aceptabilidad para indicar que los datos han mejorado en su calidad teniendo un porcentaje de 100% para la

dimensión de integridad superando al análisis inicial en el cual se obtuvo 95,36% teniendo una mejora del 4.64% en lo que se refiere a la duplicidad de datos, para la dimensión de consistencia se han obtenido un 98.08% superando el análisis previo en el cual se obtuvo 93,83% con una mejora del 4,25% en lo que se refiere a la precisión y validez de los datos, además el nivel de confiabilidad se ha superado obteniendo 99,23% con la limpieza de los datos en relación al primer análisis en el cual se obtuvo 94,75% superando esta dimensión en 4,48% con el uso de la herramienta mejorando así los datos de esta tabla .

3.3.2. ORACLE DATA INTEGRATOR

Limpieza

Pasos que se deben seguir para el uso de la herramienta:

- Identificar la tabla destino
- Identificar las tablas fuentes
- Identificar las tablas de Referencia (Lookup)
- Verificar los pareos de campos (mapping)
    - Emparejamiento Automáticos
    - Columnas no nulas
    - Añadir columnas adicionales
- Probar regularmente la extracción
- En las transformaciones
    - Identificar, verificar y validar las condiciones
    - Verificar y validar campos y funciones para convertir formatos de fecha

- Verificar tamaños de columnas para no truncar los datos extraídos que de algún tipo de error
- Verificar los tipos de datos (Datatype)
- Verificar las secuencias

En la Figura III. 7 se muestra los componentes utilizados para la limpieza.

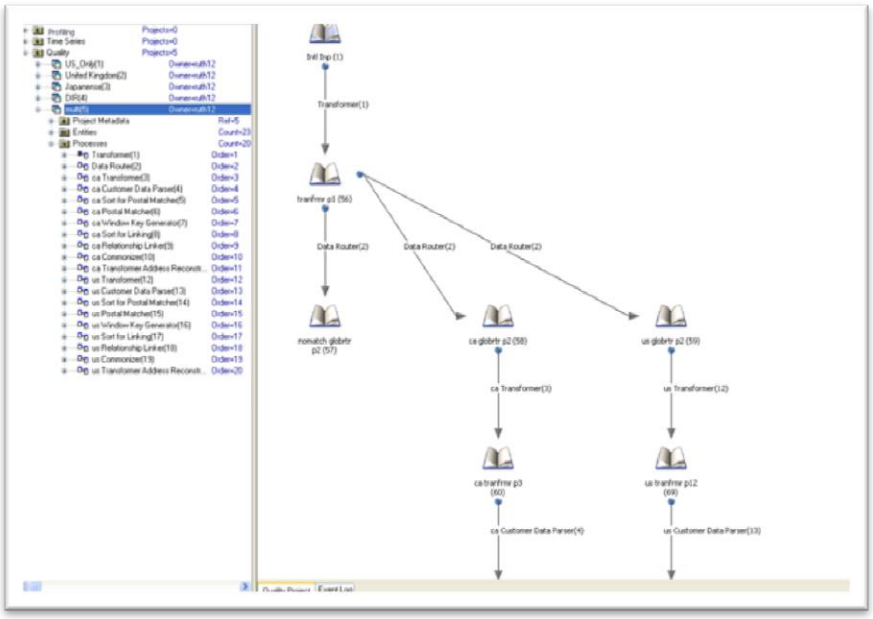


Figura III. 8 Limpieza con Oracle Data Integrator

Fuente: Investigador

El resultado de la limpieza se puede visualizar en el ANEXO 1.

3.3.2.1. ANALISIS DE LOS RESULTADOS

- **Análisis de la tabla TPM_CUS:**

- **Parámetro de Precisión:**

Tabla III. XXVII. Parámetros de Precisión-Tabla TPM_CUS

TPM_CUS		
COLUMNA	**VALORES NULOS**	**PRECISIÓN**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
NOMBRES	0%	100%
IDENTIFICACION	0%	100%
TIPO_IDENTIFICACION	0%	100%
DIRECCION	0%	100%
EMAIL	0%	100%
TELEFONO	0%	100%
CELULAR	0%	100%

Elaborado por: Investigador

En la Tabla III. XXVIII muestra que el parámetro de la precisión se cumple para todas las columnas de la tabla TPM_CUS las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

- **Parámetro de Valores Aceptables:**

Tabla III. XXIX. Parámetros Valores Aceptables -Tabla TPM_CUS

TPM_CUS		
COLUMNA	**CARACTERES INVÁLIDOS**	**VALORES ACEPTABLES**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
NOMBRES	0%	100%
IDENTIFICACION	0%	100%
TIPO_IDENTIFICACION	0%	100%
DIRECCION	2,134	97,866%
EMAIL	6,885	93,115 %
TELEFONO	0,271	99,729%
CELULAR	3,567	96,433%

Elaborado por: Investigador

En la Tabla III. XXX muestra que el parámetro de valores aceptables cumple casi en su totalidad de columnas de la tabla TPM_CUS las cuales superan el umbral de conformidad que es el 95% o superior como se debe establece y se detalla en la Tabla III.I exepto para la columna EMAIL la cual no supera el umbral de conformidad.

- **Parámetro de Duplicidad:**

Tabla III. XXXI Parámetros de Duplicidad -Tabla TPM_CUS

TPM_CUS		
COLUMNA	**DUCPLICADOS**	**DUPLICIDAD**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
NOMBRES	0%	100%
IDENTIFICACION	0%	100%
TIPO_IDENTIFICACION	0%	100%
DIRECCION	0%	100%
EMAIL	0%	100%
TELEFONO	0%	100%
CELULAR	0%	100%

Elaborado por: Investigador

En la Tabla III. XXXII muestra que el parámetro de la duplicidad se cumple para todas las columnas de la tabla TPM_CUS las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

- **Parámetro de Confiabilidad:**

Tabla III. XXXIII. Parámetros de Confiabilidad -Tabla TPM_CUS

TPM_CUS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (PRESICIÓN + VALORES ACEPTABLES)/2	**CONFIABILIDAD** (0,6)INTEGRIDAD + (0,4)CONSITENCIA
CODIGO	100%	100%	100%
CODIGO_CLIENTE	100%	100%	100%
NOMBRES	100%	100%	100%
IDENTIFICACION	100%	100%	100%

TIPO_IDENTIFICACION	100%	100%	100%
DIRECCION	100%	98,943%	99,5772%
EMAIL	100%	96,5575%	98,623%
TELEFONO	100%	99,8645%	99,9458%
CELULAR	100%	98,2165%	99,21648%

Elaborado por: Investigador

En la Tabla III. XXXIV muestra que el parámetro de la confiabilidad se cumple para todas las columnas de la tabla TPM_CUS las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

Resultados finales.

Tabla III. XXXV. Resultados Finales

TPM_CUS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	99,29%
CONFIABILIDAD	99,71%

Elaborado por: Investigador

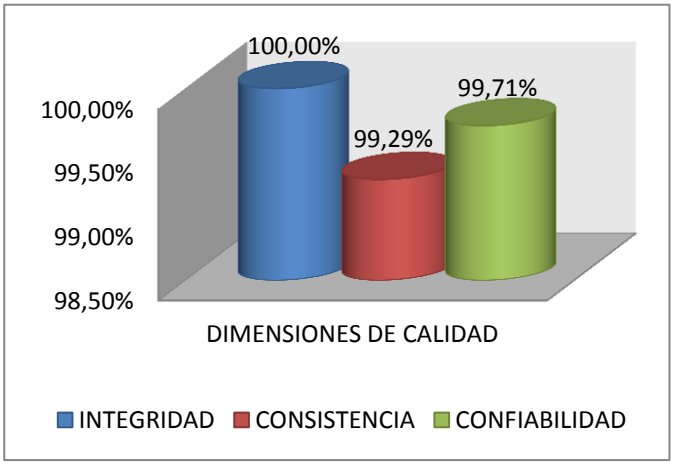


Figura III. 9 Resultados Finales

Fuente: Investigador

Interpretación de resultados finales.

Después de realizar el análisis posterior a la limpieza de los datos utilizando la herramienta Oracle Data Quality en la tabla TPM_CUS y observando las Tablas III.VI y III.XXVI podemos evidenciar que las dimensiones examinadas ha obtenido el nivel aceptable para indicar que los datos han mejorado en su calidad teniendo un porcentaje de 100% para la dimensión de integridad superando al análisis inicial en el cual se obtuvo 90,19% teniendo una mejora del 9.81% en lo que se refiere a la duplicidad de datos, para la dimensión de

consistencia se han obtenido un 99.29% superando el análisis previo en el cual se obtuvo 92,22% con una mejora del 7,07% en lo que se refiere a la precisión y validez de los datos, además el nivel de confiabilidad se ha superado obteniendo 99,71% con la limpieza de los datos en relación al primer análisis en el cual se obtuvo 91% superando esta dimensión en 8,71% con el uso de la herramienta mejorando así los datos de esta tabla.

- **Análisis de la tabla TPM_TRAN:**

  - **Parámetro de Precisión:**

Tabla III. XXXVI . Parámetros de Precisión-Tabla TPM_TRAN

TPM_TRAN		
COLUMNA	**VALORES NULOS**	**PRECISIÓN**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
INL_AMT	0%	100%

Elaborado por: Investigador

En la Tabla III. XXXVII muestra que el parámetro de la precisión se cumple para todas las columnas de la tabla TPM_TRAN las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

  - **Parámetro de Valores Aceptables:**

Tabla III. XXXVIII. Parámetros de Valores Aceptables -Tabla TPM_TRAN

TPM_TRAN		
COLUMNA	**CARACTERES INVÁLIDOS**	**VALORES ACEPTABLES.**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
INL_AMT	10.303%	89,697%

Elaborado por: Investigador

En la Tabla III. XXXIX muestra que el parámetro de valores aceptables se cumple para las dos primeras columnas de la tabla TPM_TRAN las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I excepto para las columnas INL_AMT.

- **Parámetro de Duplicidad:**

Tabla III. XL. Parámetros de Duplicidad -Tabla TPM_TRAN

TPM_TRAN		
COLUMNA	**DUCPLICADOS**	**DUPLICIDAD**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
INL_AMT	0%	100%

Elaborado por: Investigador

En la Tabla III. XLI muestra que el parámetro de la duplicidad se cumple para todas las columnas de la tabla TPM_TRAN las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

- **Parámetro de Confiabilidad:**

Tabla III. XLII. Parámetros de Confiabilidad -Tabla TPM_TRAN

TPM_TRAN			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (PRESICIÓN + VALORES ACEPTABLES)/2	**CONFIABILIDAD** (0,6)INTEGRIDAD + (0,4)CONSITENCIA
CODIGO	100%	100%	100%
CODIGO_CLIENTE	100%	100%	100%
INL_AMT	100%	94,8485%	97,9394%

Elaborado por: Investigador

En la Tabla III. XLIII muestra que el parámetro de la confiabilidad se cumple para todas las columnas de la tabla TPM_TRAN las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

Resultados finales.

Tabla III. XLIV. Resultados Finales

TPM_CUS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	98,28%
CONFIABILIDAD	99,31%

Elaborado por: Investigador

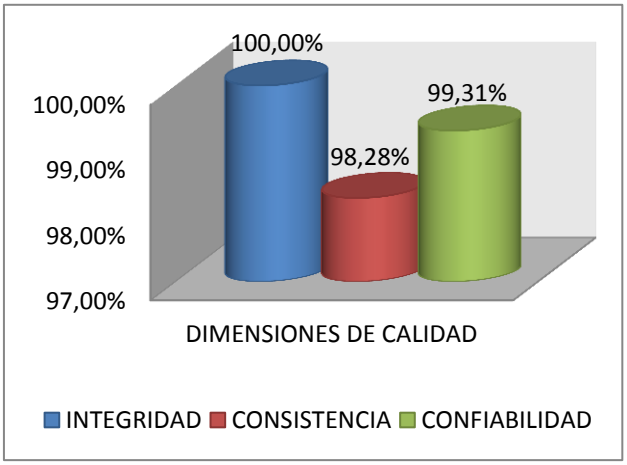


Figura III. 10. Resultados Finales

Fuente: Investigador

Interpretación de resultados finales.

Después de realizar el análisis posterior a la limpieza de los datos utilizando la herramienta Oracle Data Quality en la tabla TPM_TRAN y observando las Tablas III.XVI y III.XXXI podemos evidenciar que las dimensiones examinadas han mejorado en su nivel de aceptabilidad para indicar que los datos han mejorado en su calidad teniendo un porcentaje de 100% para la dimensión de integridad superando al análisis inicial en el cual se obtuvo 95,36% teniendo una mejora del 4.64% en lo que se refiere a la duplicidad de datos, para la dimensión de consistencia se han obtenido un 98.28% superando el análisis previo en el cual se obtuvo 93,83% con una mejora del 4,45% en lo que se refiere a la precisión y validez de los datos, además el nivel de confiabilidad se ha superado obteniendo 99,31% con la limpieza de los datos en relación al primer análisis en el cual se obtuvo 94,75% superando esta dimensión en 4,56% con el uso de la herramienta mejorando así los datos de esta tabla.

3.3.3. INFORMATICA

Limpieza

Los pasos que se siguen son los siguientes:

1. Se crea un nueva carpeta de proyectos
2. Dentro de la carpeta creada arrastramos todos los componentes necesarios para la limpieza de los datos.
3. Se selecciona todos los componentes de transformación y limpieza necesarios y se ejecuta el proyecto de limpieza.

La Figura III. 11 se muestra los objetos que la herramienta nos provee para la limpieza

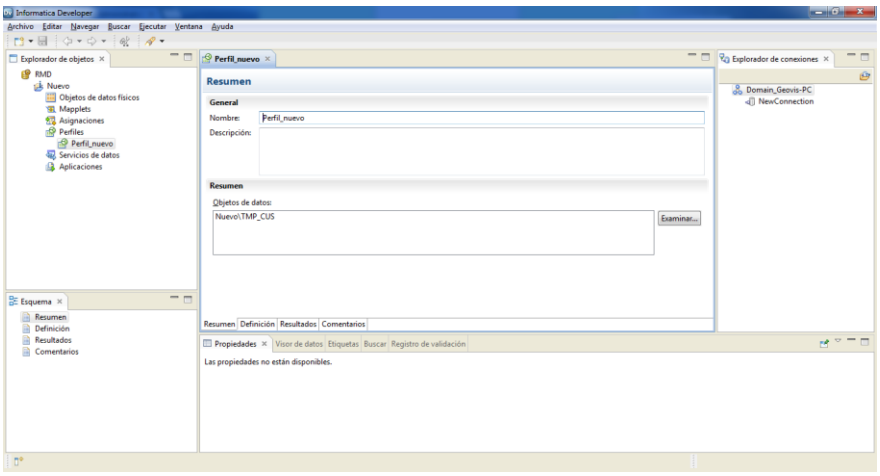


Figura III. 12. Limpieza herramienta Informatica

Fuente: Investigador

El resultado de la limpieza se puede visualizar en el ANEXO 1.

3.3.3.1. ANALISIS DE LOS RESULTADOS

- **Análisis de la tabla TPM_CUS:**

- **Parámetro de Precisión:**

Tabla III. XLV. Parámetros de Precisión-Tabla TPM_CUS

TPM_CUS		
COLUMNA	**VALORES NULOS**	**PRECISIÓN**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
NOMBRES	0%	100%
IDENTIFICACION	0%	100%
TIPO_IDENTIFICACION	0%	100%
DIRECCION	0%	100%
EMAIL	0%	100%
TELEFONO	0%	100%
CELULAR	0%	100%

Elaborado por: Investigador

En la Tabla III. XLVI muestra que el parámetro de la precisión se cumple para todas las columnas de la tabla TPM_CUS las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

- **Parámetro de Valores Aceptables:**

Tabla III. XLVII. Parámetros de Valores Aceptables -Tabla TPM_CUS

TPM_CUS		
COLUMNA	**CARACTERES INVÁLIDOS**	**VALORES ACEPTABLES**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
NOMBRES	0%	100%
IDENTIFICACION	0%	100%
TIPO_IDENTIFICACION	0%	100%
DIRECCION	2,134	97,866%
EMAIL	6,885	93,115 %
TELEFONO	0,271	99,729%
CELULAR	3,567	96,433%

Elaborado por: Investigador

En la Tabla III. XLVIII muestra que el parámetro de valores aceptables cumple para la mayoria de las columnas de la tabla TPM_CUS las cuales superan el umbral de conformidad que es el 95% o superior como se debe establece y se detalla en la Tabla III.I exepto para la columna EMAIL la cual no supera el umbral de conformidad.

- **Parámetro de Duplicidad:**

Tabla III. XLIX. Parámetros de Duplicidad -Tabla TPM_CUS

TPM_CUS		
COLUMNA	**DUCPLICADOS**	**DUPLICIDAD**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
NOMBRES	0%	100%
IDENTIFICACION	0%	100%
TIPO_IDENTIFICACION	0%	100%
DIRECCION	0%	100%
EMAIL	0%	100%
TELEFONO	0%	100%
CELULAR	0%	100%

Elaborado por: Investigador

En la Tabla III. L muestra que el parámetro de la duplicidad se cumple para todas las columnas de la tabla TPM_CUS las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

- **Parámetro de Confiabilidad:**

Tabla III. LI. Parámetros de Confiabilidad -Tabla TPM_CUS

TPM_CUS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (PRESICIÓN + VALORES ACEPTABLES)/2	**CONFIABILIDAD** (0,6)INTEGRIDAD + (0,4)CONSITENCIA
CODIGO	100%	100%	100%
CODIGO_CLIENTE	100%	100%	100%
NOMBRES	100%	100%	100%
IDENTIFICACION	100%	100%	100%
TIPO_IDENTIFICACION	100%	100%	100%
DIRECCION	100%	98,943%	99,5772%

EMAIL	100%	96,5575%	98,623%
TELEFONO	100%	99,8645%	99,9458%
CELULAR	100%	98,2165%	99,21648%

Elaborado por: Investigador

En la Tabla III. LII muestra que el parámetro de la confiabilidad se cumple para todas las columnas de la tabla TPM_CUS las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

Resultados finales.

Tabla III. LIII. Resultados Finales

TPM_CUS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	99,29%
CONFIABILIDAD	99,71%

Elaborado por: Investigador

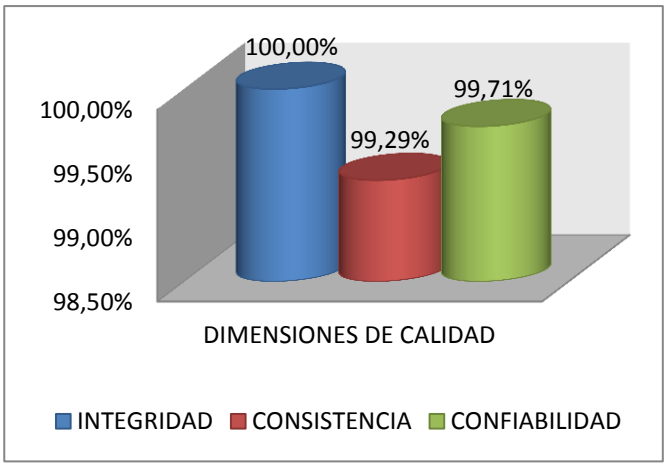


Figura III. 13. Resultados Finales

Fuente: Investigador

Interpretación de resultados finales.

Después de realizar el análisis posterior a la limpieza de los datos utilizando la herramienta Informatica en la tabla TPM_CUS y observando las Tablas III.VI y III. LIV podemos evidenciar que las dimensiones examinadas ha obtenido el nivel aceptable para indicar que los datos han mejorado en su calidad teniendo un porcentaje de 100% para la dimensión de integridad superando al análisis inicial en el cual se obtuvo 90,19% teniendo una mejora del 9.81% en lo que se refiere a la duplicidad de datos, para la dimensión de consistencia se han obtenido un 99.29% superando el análisis previo en el cual se obtuvo 92,22% con una

mejora del 7,07% en lo que se refiere a la precisión y validez de los datos, además el nivel de confiabilidad se ha superado obteniendo 99,71% con la limpieza de los datos en relación al primer análisis en el cual se obtuvo 91% superando esta dimensión en 8,71% con el uso de la herramienta mejorando así los datos de esta tabla.

- **Análisis de la tabla TPM_TRAN:**

  - **Parámetro de Precisión:**

Tabla III. LV. Parámetros de Precisión-Tabla TPM_TRAN

TPM_TRAN		
COLUMNA	**VALORES NULOS**	**PRECISIÓN**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
INL_AMT	0%	100%

Elaborado por: Investigador

En la Tabla III. LVI muestra que el parámetro de la precisión se cumple para todas las columnas de la tabla TPM_TRAN las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

  - **Parámetro de Valores Aceptables:**

Tabla III. LVII. Parámetros de Precisión-Tabla TPM_CUS

TPM_TRAN		
COLUMNA	**CARACTERES INVÁLIDOS**	**VALORES ACEPTABLES.**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
INL_AMT	10.303%	89,697%

Elaborado por: Investigador

En la Tabla III. LVIII muestra que el parámetro de valores aceptables se cumple para la mayoria las columnas de la tabla TPM_TRAN las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I excepto para las columnas INL_AMT.

- **Parámetro de Duplicidad:**

Tabla III. LIX. Parámetros de Duplicidad -Tabla TPM_CUS

TPM_TRAN		
COLUMNA	**DUCPLICADOS**	**DUPLICIDAD**
CODIGO	0%	100%
CODIGO_CLIENTE	0%	100%
INL_AMT	0%	100%

Elaborado por: Investigador

En la Tabla III. LX muestra que el parámetro de la duplicidad se cumple para todas las columnas de la tabla TPM_TRAN las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

- **Parámetro de Confiabilidad:**

Tabla III. LXI. Parámetros de Confiabilidad -Tabla TPM_CUS

TPM_TRAN			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (PRESICIÓN + VALORES ACEPTABLES)/2	**CONFIABILIDAD** (0,6)INTEGRIDAD + (0,4)CONSITENCIA
CODIGO	100%	100%	100%
CODIGO_CLIENTE	100%	100%	100%
INL_AMT	100%	94,8485%	97,9394%

Elaborado por: Investigador

En la Tabla III. **LXII** muestra que el parámetro de la confiabilidad se cumple para todas las columnas de la tabla TPM_TRAN las cuales superan el umbral de conformidad que es el 95% o superior como se establece y se detalla en la Tabla III.I.

Resultados finales.

Tabla III. LXIII. Resultados Finales

TPM_CUS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	98,28%
CONFIABILIDAD	99,31%

Elaborado por: Investigador

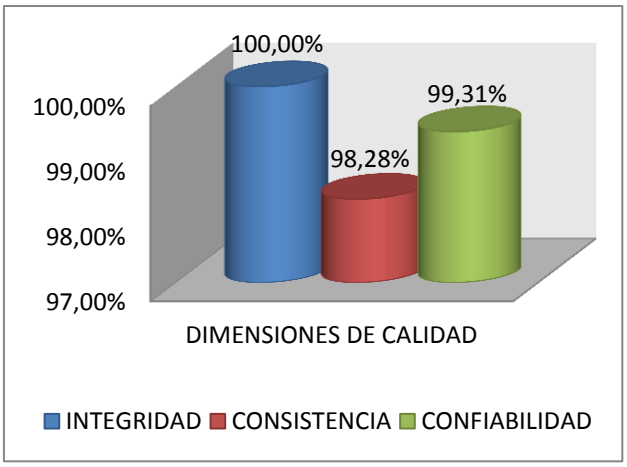


Figura III. 14. Resultados Finales

Fuente: Investigador

Interpretación de resultados finales.

Después de realizar el análisis posterior a la limpieza de los datos utilizando la herramienta Informatica en la tabla TPM_TRAN y observando las Tablas III.XI y III.XLI podemos evidenciar que las dimensiones examinadas han mejorado en su nivel de aceptabilidad para indicar que los datos han mejorado en su calidad teniendo un porcentaje de 100% para la dimensión de integridad superando al análisis inicial en el cual se obtuvo 95,36% teniendo una mejora del 4.64% en lo que se refiere a la duplicidad de datos, para la dimensión de consistencia se han obtenido un 98.28% superando el análisis previo en el cual se obtuvo 93,83% con una mejora del 4,45% en lo que se refiere a la precisión y validez de los datos, además el nivel de confiabilidad se ha superado obteniendo 99,31% con la limpieza de los datos en relación al primer análisis en el cual se obtuvo 94,75% superando esta dimensión en 4,56% con el uso de la herramienta mejorando así los datos de esta tabla.

3.4. RESULTADOS GENERALES POR PARÁMETROS

3.4.1. Resultado general de los datos sin utilizar las herramientas de limpieza.

Tabla III. LXIV. Resultados generales por parámetro sin limpieza

	INTEGRIDAD	CONSISTENCIA	CONFIABILIDAD
TPM_CUS	90,19%	92,22%	91%
TPM_TRAN	95,36%	93,83%	94,75%

Elaborado por: Investigador

En la Tabla III. LXV muestra el resultado de las dismensiones de calidad: integridad, consistencia y confiabilidad en el análisis preliminar de los datos.

3.4.2. Resultado general de los datos al utilizar las herramientas de limpieza.

Tabla III. LXVI. Resultados generales por parámetro con la herramienta.

HERRAMIENTA / TABLA		Integridad	Consistencia	Confiabilidad
SQL Power	**TPM_CUS**	100%	99,07%	99,63%
	TPM_TRAN	100%	98,08%	99,23%
Data Integrator	**TPM_CUS**	100%	99,29%	99,71%
	TPM_TRAN	100%	98,28%	99,31%
Informatica	**TPM_CUS**	100%	99,29%	99,71%
	TPM_TRAN	100%	98,28%	99,31%

Elaborado por: Investigador

En la Tabla III. **LXVII.** muestra el resultado de las dismensiones de calidad: integridad, consistencia y confiabilidad en el análisis preliminar de los datos.

Tabla TPM_CUS

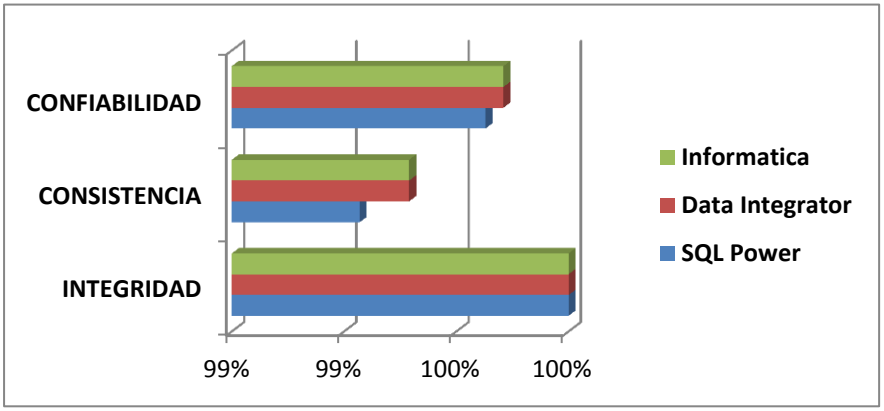


Figura III. 15. General Resultados por parámetro.

Fuente: Investigador

Interpretación de la gráfica de resultado final por parámetros para la tabla TPM_CUS

Como se puede observar en la Figura III. 12 las dimensiones de calidad analizadas: Integridad, Consistencia y Confiabilidad después de la utilización de las herramientas para calidad de datos han mejorado de manera aceptable para cada una de las dimensiones en lo que tinene que ver con superar valores nulos, corregir valores inválidos y los registros duplicados la calidad de los datos en la tabla TPM_CUS se ha afinado.

Tabla TPM_TRAN

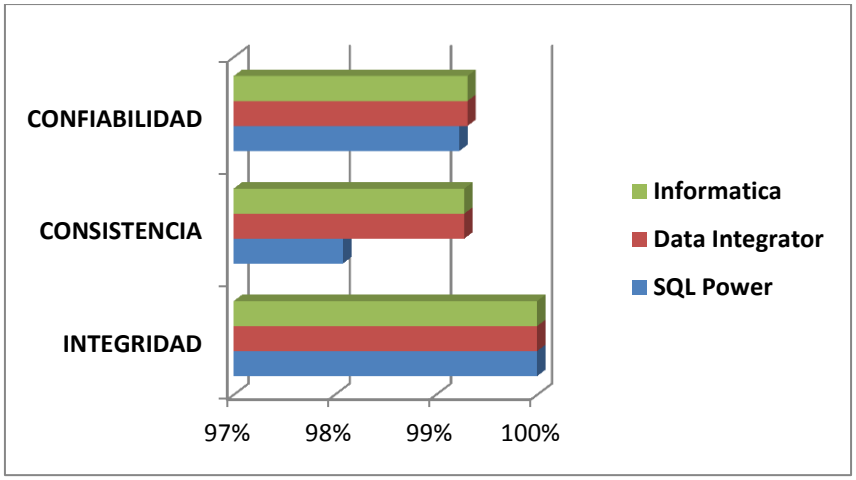


Figura III. 16. General Resultados por parámetro.

Fuente: Investigador

Interpretación de la gráfica de resultado final por parámetros para la tabla TPM_TRAN

Como se puede observar en la Figura III. 13 las dimensiones de calidad analizadas: Integridad, Consistencia y Confiabilidad después de la utilización de las herramientas para calidad de datos han mejorado de manera aceptable para cada una de las dimensiones en lo que tinene que ver con superar valores nulos, corregir valores inválidos y los registros duplicados, la calidad de los datos en la tabla TPM_CUS se ha afinado.

3.5. ANÁLISIS COMPARATIVO DE LAS HERRAMIENTAS DATA QUALITY FRENTE A LAS HERRAMIENTAS SOFTWARE LIBRE

Los resultados de los criterios de comparación con sus respectivos parámetros se realizarán en un cuadro comparativo de las herramientas DataCleaner, SQL Power, Oracle Oracle Data Quality e Informatica, cuyas pruebas de desarrollo fueron realizadas bajo los mismo escenario.

Los criterios y parámetros de evaluación para cada herramienta serán descritos a continuación:

Tabla III. LXVIII. Criterios y Parámetros.

N°	Criterio	Parámetro	Concepto
1	Compatibilidad.	Acceso a datos.	Compresión correcta de los componentes de la herramienta y funciones externas o sistemas de administración de datos.
2	Rendimiento.	Limpieza de datos.	Disminución de registros erróneos con funciones de limpieza de datos, eliminación de duplicados, etc.
		Configuración de la herramienta.	Aplicables a la elección de cualquier tipo de herramienta (interfaz de uso, la disponibilidad y el lenguaje de ayuda).

Elaborado por: Investigador

3.5.1. Definición de los indicadores

En las siguientes tablas se mencionan los indicadores para cada parámetro establecido en la Tabla III. LXIX, los mismos que serán analizados con el fin de entregar información específica.

Tabla III. LXX. Acceso a Datos.

ACCESO A DATOS.	
Indicador	**Descripción**
Soporte a múltiples bases de datos.	Números de motores de base de datos que soporta cada herramienta.
Manipulación con la base de datos.	Tiempo en configurar la conexión a la base de datos
Desempeño con la base de datos.	Manipulación de los datos: tipos de datos y exploración.

Elaborado por: Investigador

Tabla III. LXXI. Tabla. III. 1. Limpieza de Datos

LIMPIEZA DE DATOS.	
Indicador	**Descripción**
Eliminación de registros duplicados.	Tiempo de configuración y disminución de errores al eliminar registros duplicados para cada herramienta.
Funciones de limpieza.	Capacidad de emplear funciones de limpieza y complejidad de uso.
Confiabilidad de la calidad de datos	Resultados del uso de las funciones para la limpieza de los datos.

Elaborado por: Investigador

Tabla III. LXXII. Configuración de herramientas.

CONFIGURACIÓN DE LA HERRAMIENTA.	
Indicador	**Descripción**
Instalación de la herramienta.	Tiempo de instalación de cada herramienta.
Configuraciones.	Tiempo de configuraciones adicionales en las herramientas.
Documentación.	Existencia de suficiente información para el correcto uso de las herramientas.

Elaborado por: Investigador

3.5.2. Criterio de Evaluacion.

La calificación para cada parámetro se determinó de acuerdo a la escala que se mostrara a continuación, lo cual nos permitió determinar la tecnología que se adapta mejor para la calidad de datos.

3.5.2.1. Valoración cualitativa y cuantitativa.

Tabla III. LXXIII. Valoración.

Regular	Bueno	Muy Bueno	Excelente
<70%	>=70% y <80%	>=80% y <95%	>=95%

Elaborado por: Investigador

La evaluación para los parámetros es de acuerdo al tiempo y experiencia de desarrollo, para lo cual la valoración variará entre uno y cuatro.

3.5.2.2. Escala de valoración cualitativa y cuantitativa para los parámetros

Tabla III. LXXIV. Escala de Valoración

Valor Cualitativo		**Valor Representativo**
Insuficiente	No Satisfactorio	1
Parcial	Poco Satisfactorio	2
Suficiente	Satisfactorio	3
Excelente	Muy Satisfactorio	4

Elaborado por: Investigador

3.5.2.3. Equivalencias de los valores cuantitativos.

Tabla III. LXXV. Equivalencia de valores.

Valor Cuantitativo	1	2	3	4
	1 - 10	11 - 13	15 - 17	18 – 20
Equivalencias	0.25	0.50	0.75	1

Elaborado por: Investigador

Para la realización de la comparación se utilizará la siguiente nomenclatura:

W *= Representa el puntaje obtenido por la herramienta DataCleaner.*

X *= Representa el puntaje obtenido por la herramienta SQL Power.*

Y *= Representa el puntaje obtenido por la herramienta Oracle Oracle Data Quality.*

Z *= Representa el puntaje obtenido por la herramienta Informatica.*

M = *Representa el puntaje sobre el cual será evaluado el parámetro.*

Cdc = *Representa el puntaje alcanzado de DataCleaner en el parámetro.*

Csp = *Representa el puntaje alcanzado de SQL Power en el parámetro.*

Coi = *Representa el puntaje alcanzado de Oracle Oracle Data Quality en el parámetro.*

Cin **=** *Representa el puntaje alcanzado de Informatica en el parámetro.*

Ct **=** *Representa el puntaje por el cual es evaluado el parámetro.*

Pdc *= Calificación porcentual obtenida por DataCleaner.*

Psp *= Calificación porcentual obtenida por SQL Power.*

Poi *= Calificación porcentual obtenida por Oracle Oracle Data Quality.*

Pin *= Calificación porcentual obtenida por Informatica.*

Phl *= Calificación porcentual obtenida por las herramientas libres.*

Php *= Calificación porcentual obtenida por las herramientas propietarias.*

Las fórmulas que se utilizarán en el proceso del análisis comparativo son las siguientes:

$$Cdc = \sum W$$

$$Csp = \sum X$$

$$Coi = \sum Y$$

$$Cin = \sum Z$$

$$Ct = \sum M$$

$$Pdc = \left(\frac{Cdc}{Ct}\right) * 100\%$$

$$Psp = \left(\frac{Csp}{Ct}\right) * 100\%$$

$$Poi = \left(\frac{Coi}{Ct}\right) * 100\%$$

$$Pin = \left(\frac{Cin}{Ct}\right) * 100\%$$

$$Phl = \left(\frac{Pdc + Psp}{2}\right)$$

$$Phl = \left(\frac{Poi + Pin}{2}\right)$$

3.5.3. Análisis De Los Parámetros De Comparación

3.5.3.1. Acceso a datos

El acceso a la base de datos es un criterio muy importante, por lo cual se analizara los aspectos necesarios para la manipulación de los datos y acceso a los mismos.

- **Soporte a múltiples bases de datos:** En este parámetro se valorizará tomando en consideración los números de motores de base de datos que soporta cada herramienta, en base a la parte 1 de pruebas (Conexiones).

Tabla III. LXXVI. Soportesa mulples base de datos.

Valoración	
Numero de Bases de datos que soporta	**Valoración cualitativa**
Hasta 2	No Satisfactorio
Hasta 3	Poco Satisfactorio
Hasta 4	Satisfactorio
>= 5	Muy Satisfactorio

Elaborado por: Investigador

- **Manipulación con la base de datos:** En este parámetro se valorizará de acuerdo al tiempo en configurar la conexión a la base de datos, en base a la parte 1 de pruebas (Conexiones).

Tabla III. LXXVII. Manipulación con la base de datos

Valoración	
Tiempo en minutos	**Valoración cualitativa**
16 a 20	No Satisfactorio
11 a 15	Poco Satisfactorio
6 a 10	Satisfactorio
1 a 5	Muy Satisfactorio

Elaborado por: Investigador

- **Desempeño con la base de datos:** Este parámetro valorizará la manipulación de los datos tales como el soporte a varios tipos de datos y exploración.

Tabla III. LXXVIII. Desempeño con la base de datos.

Valoración	
Tiempo en minutos	**Valoración Cualitativa**
16 a 20	No Satisfactorio
11 a 15	Poco Satisfactorio
6 a 10	Satisfactorio
1 a 5	Muy Satisfactorio

Elaborado por: Investigador

Valoraciones de Acceso a datos

Tabla III. LXXIX. Resultados del Criterio para Herramientas Libres.

Parámetros	DataCleaner		SQL Power	
	Valor Cualitativo	Valor Obtenido/4	Valor Cualitativo	Valor Obtenido/4
Soporte a múltiples DBMS	Muy Satisfactorio	4	Muy Satisfactorio	4
Manipulación con la BD	Muy Satisfactorio	4	Muy Satisfactorio	4
Desempeño con la BD	No Satisfactorio	1	Poco Satisfactorio	2

Elaborado por: Investigador

Resultados del Criterio para Herramientas Propietarias

Tabla III. LXXX. Resultados del Criterio para Herramientas Propietarias.

Parámetros	Oracle Data Quality		Informática	
	Valor Cualitativo	Valor Obtenido/4	Valor Cualitativo	Valor Obtenido/4
Soporte a múltiples DBMS	Muy Satisfactorio	4	Muy Satisfactorio	4
Manipulación con la BD	Poco Satisfactorio	3	Poco Satisfactorio	2
Desempeño con la BD	Muy Satisfactorio	4	Muy Satisfactorio	4

Elaborado por: Investigador

Interpretación

Soporte para múltiples bases de datos: El soporte desde la herramienta hacia la administración de la información mediante los distintos motores de base de datos es muy importante, debido a que DataCleaner, SQL Power, Data Integrator e Informatica soportan más de 6 tipos de conexiones a motores de base de datos y otros tipos de

conexiones, por lo que todas las herramientas han obtenido una calificación de 4 puntos que equivale a Muy satisfactorio.

Manipulación con la base de datos: Este es un parámetro que nos permite identificar la eficiencia que presenta la tecnología para conectarse a uno u otro motor de base de datos por lo que el tiempo en realizar la conexión a la base de datos describe esta característica. Después de realizar las pruebas de conexión respecto al tiempo, las herramientas libres se lo puede implementar en un tiempo menor a cinco minutos, por esta razón ambas herramientas obtienen una calificación 4 puntos que equivale a Muy satisfactorio, mientras que las herramientas propietarias se debe tener una configuración más exhaustiva por esta razón ambas herramientas obtienen una calificación de 2 puntos que equivale a Poco Satisfactorio.

Desempeño con la base de datos: El desarrollo de los pruebas en las respectivas Herramientas, nos permitió determinar el desempeño que posee cada tecnología con la base de datos. Puesto que el trabajo de exploración de los datos de las herramientas libres obtuvo una calificación de 2 puntos que equivale a Poco Satisfactorio, no obstante las herramientas de software propietario presenta una mejor manipulación y facilidad en la exploración de los datos siendo estos desarrollados de mejor manera por esta razón obtiene una calificación de 4 puntos que equivale a Muy Satisfactorio.

Calificación

Cálculo de los porcentajes.

$$Cdc = \sum W$$

$$Csp = \sum X$$

$$Coi = \sum Y$$

$$Cin = \sum Z$$

$$Ct = \sum M$$

$$Pdc = \left(\frac{Cdc}{Ct}\right) * 100\%$$

$$Psp = \left(\frac{Csp}{Ct}\right) * 100\%$$

$$Poi = \left(\frac{Coi}{Ct}\right) * 100\%$$

$$Pin = \left(\frac{Cin}{Ct}\right) * 100\%$$

$Cdc: 4 + 4 + 1 = 9$

$Csp: 4 + 4 + 2 = 10$

$Coi: 4 + 3 + 4 = 11$

$Cin: 4 + 2 + 3 = 10$

$Ct: 4 + 4 + 4 = 12$

$Pdc: \left(\frac{9}{12}\right) * 100\% = 75\%$

$Psp: \left(\frac{10}{12}\right) * 100\% = 83.33\%$

$Poi: \left(\frac{11}{12}\right) * 100\% = 91.66\%$

$Pin: \left(\frac{10}{12}\right) * 100\% = 83.33\%$

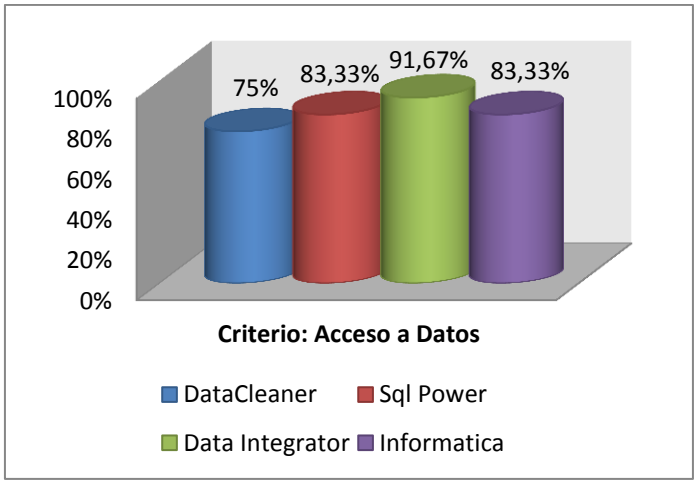


Figura III. 17. Representación Gráfica.

Fuente: Investigador

Interpretación de Resultados

Al realizar el análisis de los resultados obtenidos de las calificaciones de cada indicador para el criterio de acceso a datos se demostraron los diferencias existentes entre cada una de las herramientas, observando en la Figura III.14 se tiene como la mejor herramienta

para este criterio a Oracle Data Quality con un porcentaje del 91.67% de la calificación total por encima de SQL Power e Informatica con 8.33% del porcentaje y dejando a DataCleaner con apenas un 75% en lo que respecta al manejo de las conexiones, manipulación de los datos y desempeño al momento de trabajar con los estos, debido a su factibilidad de uso y su amigabilidad en la interfaz gráfica.

3.5.3.2. Limpieza de datos

La limpieza de los datos es muy importante para resolver los múltiples problemas que representan la mal calidad de los mismos, por lo cual analizaremos las distintas opciones para su solución.

- **Eliminación de registros duplicados:** Este parámetro se valorizará tomando en consideración el tiempo de configuración y disminución de errores al eliminar registros duplicados para cada herramienta, en base a la parte de limpieza de pruebas.

Tabla III. LXXXI. Valoración.

Valoración	
Tiempo en minutos	**Valoración cualitativa**
16 a 20	No Satisfactorio
11 a 15	Poco Satisfactorio
6 a 10	Satisfactorio
1 a 5	Muy Satisfactorio

Elaborado por: Investigador

- **Funciones de limpieza:** Este parámetro se valorizará tomando en consideración la capacidad de emplear funciones de limpieza y complejidad de uso tales como eliminar caracteres especiales o valores nulos, etc.

Tabla III. LXXXII. Valoración en fuentes de limpieza.

Valoración	
Numero de funciones que posee la herramienta	**Valoración cualitativa**
Hasta 2	No Satisfactorio
Hasta 3	Poco Satisfactorio
Hasta 4	Satisfactorio
>= 5	Muy Satisfactorio

Elaborado por: Investigador

- **Confiabilidad de los datos:** Este parámetro se valorizará tomando en consideración los resultados obtenidos de la limpieza de los datos en el escenario planteado.

Tabla III. 58. Valoración en fuentes de limpieza.

Valoración	
Resultados obtenidos	**Valoración cualitativa**
<70%	No Satisfactorio
>=70% y <80%	Poco Satisfactorio
>=80% y <95%	Satisfactorio
>=95%	Muy Satisfactorio

Elaborado por: Investigador

Valoraciones de limpieza de datos

Resultados del criterio para herramientas libres

Tabla III. LXXXIII. Resultados del Criterio para Herramientas Libres.

Parámetros	DataCleaner		SQL Power	
	Valor Cualitativo	Valor Obtenido/4	Valor Cualitativo	Valor Obtenido/4
Eliminación de registros	No Satisfactorio	1	Muy Satisfactorio	4
Funciones de limpieza	Poco Satisfactorio	2	Muy Satisfactorio	4
Confiabilidad de los datos	No Satisfactorio	1	Muy Satisfactorio	4

Elaborado por: Investigador

Resultados del Criterio para Herramientas Propietarias

Tabla III. LXXXIV. Resultados del Criterio para Herramientas Propietarios.

Parámetros	Oracle Data Integrator		Informatica	
	Valor Cualitativo	Valor Obtenido/4	Valor Cualitativo	Valor Obtenido/4
Eliminación de registros	Muy Satisfactorio	4	Satisfactorio	3
Funciones de limpieza	Poco Satisfactorio	3	Satisfactorio	3
Confiabilidad de los datos	Muy Satisfactorio	4	Muy Satisfactorio	4

Elaborado por: Investigador

Interpretación

Eliminación de registros duplicados: Para las herramientas libres el tiempo que toma configurar la eliminación de los registros duplicado es bueno pero las condiciones para

DataCleaner son malas ya que son bajo demanda es decir que en la herramienta se exige el pago por este servicio por esa razón se calificó con un 1 punto lo que equivale a No Satisfactorio, mientras que para la herramienta SQL Power el servicio viene incorporado en la herramienta por esa razón se calificó con 4 puntos lo que equivale a Muy Satisfactorio. Para las herramientas propietario en el caso de Oracle Data Quality el tiempo de configuración es bueno por esa razón se calificó con 4 punto lo que equivale a Muy Satisfactorio pero para Informatica el uso requiere de una configuración exhaustiva por esa razón se calificó con 3 punto lo que equivale a Satisfactorio.

Funciones de limpieza: Para las herramientas libres como DataCleaner las funciones de limpieza de datos no son muy buenas ya que no cuenta con muchas opciones para ello por esa razón se calificó con 2 puntos lo que equivale a Poco Satisfactorio, mientras que para SQL Power posee muchas funciones por esa razón se calificó con 4 puntos lo que equivale a Muy Satisfactorio. Para las herramientas propietarias las funciones de limpieza de datos son excelentes pero las configuración de uso son algo complicadas por esa razón se calificó a las dos herramientas con 3 puntos lo que equivale a Satisfactorio.

Confiabilidad de los datos: Los resultados de esta valoración se los puede visualizar en la Tabla III.LIII. Para las herramientas libres como DataCleaner la confiabilidad de los datos no se pudo apreciar debido a que las funciones son bajo demanda y no se tomó en cuenta a esta herramienta por lo que obtuvo la calificación más baja 1 punto lo que equivale a Poco Satisfactorio, mientras que, para SQL Power la confiabilidad de las tablas en el escenario planteado superó el 95% por lo que obtuvo una calificación de 4 lo que equivale a Muy Satisfactorio. Para las herramientas propietarias también cumplieron

con un porcentaje superior al 95% por lo que obtuvieron una calificación de 4 lo que equivale a Muy Satisfactorio

Calificación

Cálculo de los porcentajes.

$$Cdc = \sum W$$

$$Csp = \sum X$$

$$Coi = \sum Y$$

$$Cin = \sum Z$$

$$Ct = \sum M$$

$$Pdc = \left(\frac{Cdc}{Ct}\right) * 100\%$$

$$Psp = \left(\frac{Csp}{Ct}\right) * 100\%$$

$$Poi = \left(\frac{Coi}{Ct}\right) * 100\%$$

$$Pin = \left(\frac{Cin}{Ct}\right) * 100\%$$

$$Cdc: 1 + 2 + 1 = 4$$

$$Csp: 4 + 4 + 4 = 12$$

$Coi: 4 + 3 + 4 = 11$

$Cin: 3 + 3 + 4 = 10$

$Ct: 4 + 4 + 4 = 12$

$Pdc: \left(\frac{4}{12}\right) * 100\% = 33.33\%$

$Psp: \left(\frac{12}{12}\right) * 100\% = 100\%$

$Poi: \left(\frac{11}{12}\right) * 100\% = 91{,}67\%$

$Pin: \left(\frac{10}{12}\right) * 100\% = 83{,}33\%$

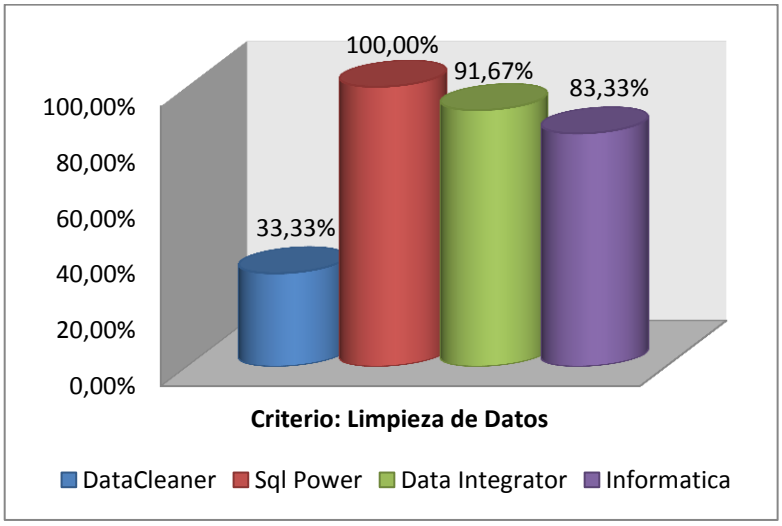


Figura III. 18. Respresentación Gráfica.

Fuente: Investigador

Interpretación de Resultados

Al realizar el análisis de los resultados obtenidos de las calificaciones de cada indicador para el criterio de limpieza de datos se demostraron los diferencias existentes entre cada una de las herramientas, y observando la Figura III.XV se tiene como la mejor

herramienta para este criterio a SQL Power con un porcentaje del 100% de la calificación total por encima de Data Integrator con 8,33% del porcentaje, además Informatica con el 83,33% y DataCleaner con el 33,33% de la calificación en lo que respecta a la confiabilidad de los datos después de una limpieza.

3.5.3.3.Configuración de la herramienta

El proceso de instalación y configuración de las herramientas es un criterio fundamental para la utilización de las mismas en la obtención de los mejores resultados de nuestros análisis.

- **Instalación de la herramienta:** Este parámetro se valorizará tomando en consideración el tiempo de instalación, en base a la parte de instalación de las pruebas.

Tabla III. LXXXV. Valoración para instalación de la herramienta .

Valoración	
Tiempo en minutos	**Valoración cualitativa**
16 a 20	No Satisfactorio
11 a 15	Poco Satisfactorio
6 a 10	Satisfactorio
1 a 5	Muy Satisfactorio

Elaborado por: Investigador

- **Configuraciones:** Este parámetro se valorizará tomando en consideración el tiempo de configuraciones adicionales en las herramientas, en base a la parte de instalación de las pruebas

Tabla III. LXXXVI. Valoración para la Configuracióon .

Valoración	
Tiempo en minutos	**Valoración cualitativa**
16 a 20	No Satisfactorio
11 a 15	Poco Satisfactorio
6 a 10	Satisfactorio
1 a 5	Muy Satisfactorio

Elaborado por: Investigador

- **Documentación:** Este parámetro se valorizará tomando en consideración si existe suficiente información para el correcto uso de las herramientas.

Tabla III. LXXXVII. Valoración en la Documentación.

Valoración	
Publicaciones	**Valoración cualitativa**
La herramienta no provee manuales	No Satisfactorio
La herramienta provee manuales	Poco Satisfactorio
La herramienta provee manuales y ayuda en línea	Satisfactorio
La herramienta provee todo tipo de información	Muy Satisfactorio

Elaborado por: Investigador

Valoraciones de Acceso a datos

Resultados del Criterio para Herramientas Libres

Tabla III. LXXXVIII. Resultados de los criterios para herramientas libres .

Parámetros	DataCleaner		SQL Power	
	Valor Cualitativo	Valor Obtenido/4	Valor Cualitativo	Valor Obtenido/4
Instalación de la herramienta	No Satisfactorio	4	Muy Satisfactorio	4
Configuraciones	Poco Satisfactorio	4	Muy Satisfactorio	4
Documentación	Satisfactorio	3	No Satisfactorio	1

Elaborado por: Investigador

Resultados del Criterio para Herramientas Propietarias

Tabla III. LXXXIX. Resultados de los criterios para herramientas propietarias .

Parámetros	**Oracle Data Integrator**		**Informatica**	
	Valor Cualitativo	**Valor Obtenido/4**	**Valor Cualitativo**	**Valor Obtenido/4**
Instalación de la herramienta	Satisfactorio	3	Satisfactorio	3
Configuraciones	Satisfactorio	4	Satisfactorio	3
Documentación	Muy Satisfactorio	4	Muy Satisfactorio	4

Elaborado por: Investigador

Interpretación

Instalación de la herramienta: En las herramientas libres no se requieren de una instalación exhaustiva por esa razón se calificó con 4 puntos lo que equivale a Muy

Satisfactorio para las dos herramientas. Para las herramientas propietarias la instalación toma varios minutos por esa razón se calificó con 3 puntos lo que equivale a Satisfactorio para las dos herramientas.

Configuraciones: En las herramientas libres no se requieren de una configuración previa por esa razón se calificó con 4 puntos lo que equivale a Muy Satisfactorio para las dos herramientas. Para las herramientas propietarias si se requiere de configuraciones adicionales para Data Integrator no lleva mucho tiempo su configuración por esa razón se calificó con 4 puntos lo que equivale a Muy Satisfactorio, mientras que la configuración en informática si se necesita de un poco más de tiempo por esa razón se calificó con 3 puntos lo que equivale a Satisfactorio.

Documentación: Para las herramientas libres en el caso de DataCleaner la herramienta si se cuenta con la documentación necesaria para su uso por esa razón se calificó con 3 puntos lo que equivale a Satisfactorio, mientras que para SQL Power la herramienta no cuenta con la documentación necesaria y es muy difícil de conseguir por esa razón se calificó con 1 punto lo que equivale a No Satisfactorio Para las herramientas propietarias la documentación es muy buena por esa razón se calificó con 4 puntos lo que equivale a Muy Satisfactorio para las dos herramientas.

Calificación

Cálculo de los porcentajes.

$$Cdc = \sum W$$

$$Csp = \sum X$$

$$Coi = \sum Y$$

$$Cin = \sum Z$$

$$Ct = \sum M$$

$$Pdc = \left(\frac{Cdc}{Ct}\right) * 100\%$$

$$Psp = \left(\frac{Csp}{Ct}\right) * 100\%$$

$$Poi = \left(\frac{Coi}{Ct}\right) * 100\%$$

$$Pin = \left(\frac{Cin}{Ct}\right) * 100\%$$

$Cdc: 4 + 4 + 3 = 11$

$Csp: 4 + 4 + 1 = 9$

$Coi: 3 + 4 + 4 = 11$

$Cin: 3 + 3 + 4 = 10$

$Ct: 4 + 4 + 4 = 12$

$$Pdc: \left(\frac{11}{12}\right) * 100\% = 91.67\%$$

$$Psp: \left(\frac{9}{12}\right) * 100\% = 75\%$$

$$Poi: \left(\frac{11}{12}\right) * 100\% = 91.67\%$$

$$Pin: \left(\frac{10}{12}\right) * 100\% = 83.33\%$$

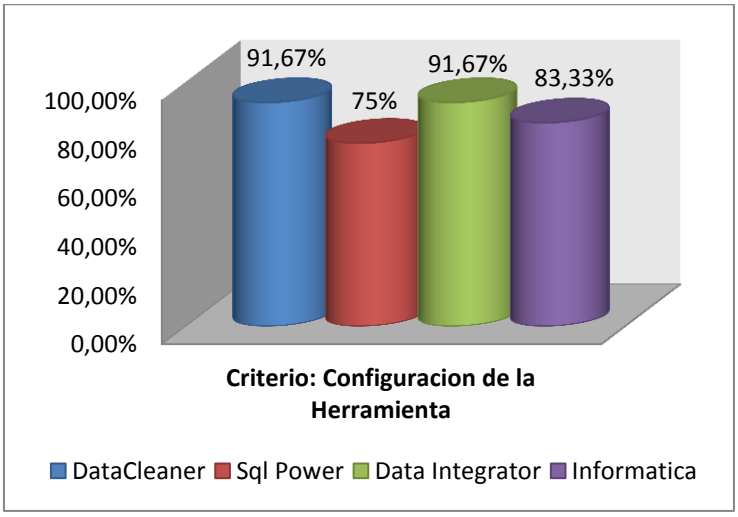


Figura III. 19. Resultados

Fuente: Investigador.

Interpretación de Resultados

Al realizar el análisis de los resultados obtenidos de las calificaciones de cada indicador para el criterio de configuración de la herramienta se demostraron los diferencias existentes entre cada una de las herramientas, y observando la Figura III.16 se tiene como las mejores herramientas para este criterio a Oracle Data Quality y DataCleaner con un porcentaje del 91.67% de la calificación total por encima de Informatica con

8.33% del porcentaje, además SQL Power con el 75% de la calificación en lo que respecta la instalación de cada herramienta y la configuración después de cada instalación además de proveer una buena documentación para la realización de este trabajo.

3.5.3.4. Puntajes Alcanzados

Después del análisis de los indicadores propuestos, mediante la aplicación de experimentación y observación en la construcción del prototipo, se obtuvieron como resultado valores cuantitativos que reflejan el desenvolvimiento de cada una de las herramientas de acuerdo a sus características en función a la desempeño del desarrollo.

Tabla de resultados obtenido en el análisis:

La Tabla III.LXVI es una consolidación de los resultados obtenidos en el análisis realizado los cuales se detallan en las tablas: Tabla III. LIV, Tabla III.LV, Tabla III.LVIII, Tabla III.LIX, Tabla III.LXIV y Tabla III.LXV.

Tabla III. XC. Tabla de resultados.

Criterios	Parámetros	Indicadores	Herramientas				Pesos Máximos
			DataCleaner	SQL Power	Oracle Data Quality	Informatica	
Compatibilidad	Acceso a datos.	Soporte a múltiples bases de datos	4	4	4	4	4
		Manipulación con la base de datos	4	4	3	2	4
		Desempeño con la base de datos	1	2	4	4	4
Rendimiento.	Limpieza de datos.	Eliminación de registros duplicados	1	4	4	3	4
		Funciones de limpieza	2	4	3	3	4
		Confiabilidad de los datos	1	4	4	4	4
	Configuración de la	Instalación de la herramienta.	4	4	3	3	4

Criterios	Parámetros	Indicadores	Herramientas				Pesos Máximos
			DataCleaner	SQL Power	Oracle Data Quality	Informatica	
	herramienta.	Configuraciones.	4	4	4	3	4
		Documentación.	3	1	4	4	4
Suma			27	31	33	30	36
Promedios Totales			3	3,44	3,67	3,33	4
Totales en porcentaje			75%	86%	91,75%	83,25%	100%

Elaborado por: Investigador

Resultados generales por parámetros

La Tabla III.LXVII se obtubo de consolidar los resultados por parametos de cada herramienta en el analisis antes realizado, los datos se obtubieron de los resultados mostrados en las figuras: Figura III.XIV, Figura III.XV, Figura III.XVI.

Tabla III. XCI. Resultados generales por Parámetro.

	Acceso a datos.	**Limpieza de datos.**	**Configuración de la herramienta.**
DataCleaner	75%	33,33%	91.67%
SQL Power	83.33%	100%	75%
Oracle	91.67%	91.67%	91.67%
Informatica	83.33%	83,33%	83.33%

Elaborado por: Investigador

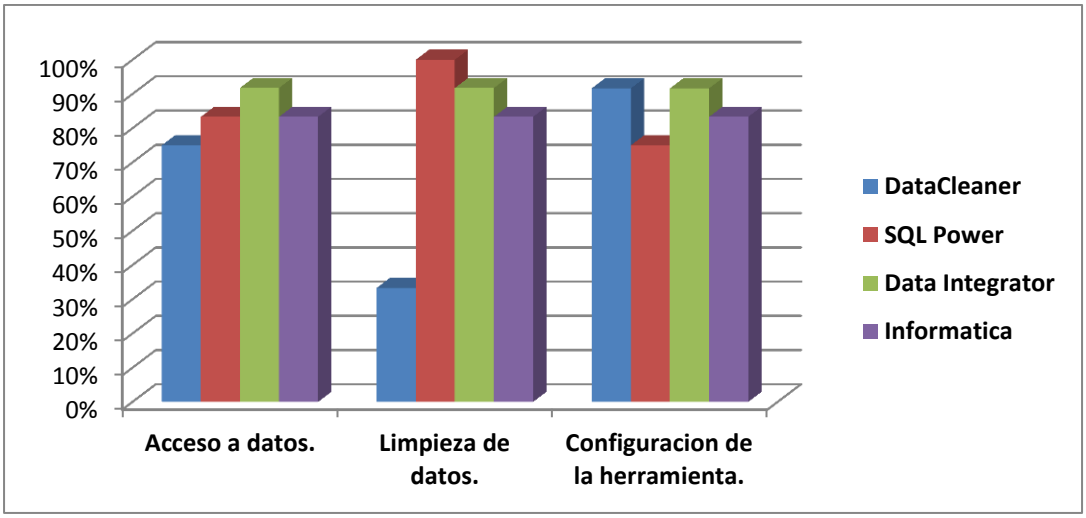


Figura III. 20. Resultados generales por Parámetros.

Fuente: Investigador

Interpretacion de los resultados generales por parámetros.

En la Figura III.17 se muestra la gráfica estadística del resultado final de la comparación de las de las herramientas para calidad de datos DataCleaner, SQL Power, Oracle Data Quality e Informatica donde se muestra los resultados obtenido para cada uno de los criterios los cuales se detallan en la Tabla III.LXVII.

Resultado final de cada herramienta

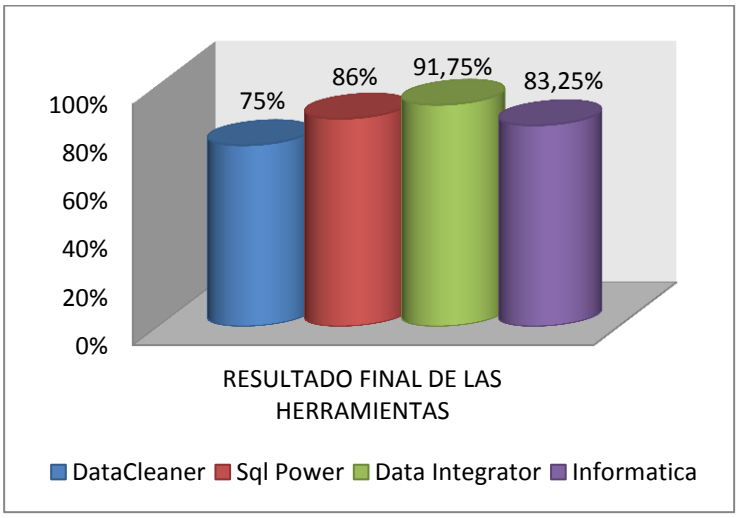


Figura III. 21. Resultados generales de Resultados Finales.

Fuente: Investigador

Interpretación de resultados finales.

En el Figura III.18 se puede observar el resultado final del análisis de las herramientas donde muestra la herramienta con mejores ventajas en el momento de realizar un proyecto de calidad de datos los resultados obtenidos fueron, para Oracle Data Quality se obtuvo el mejor resultado de entre todas las herramientas analizadas con un 91,75% por encima de SQL Power la cual obtuvo 86% la cual se ubicó en el segundo lugar en este análisis dejado

a Informatica Data Quality con un 83,25% en el tercer lugar de este análisis y al final DataCleaner con 75%. Con los resultados obtenidos se puede deducir que las herramientas con las cuales se puede continuar la siguiente fase de este estudio son las tres primeras debido a su facilidad de uso y por contar con todas las ventajas que se necesitan en este estudio.

Resultado entre herramientas libres y propietarias

La Tabla III.LXVIIIes el resultado de consolidar la Tabla III.LXVII y agrupará las herramientas en su respectivo grupo de herramienta: libres DataCleaner y SQL Power y propietarias Oracle Data Quality e Informatica.

Tabla III. XCII. Resultados entre herramientas.

	Herramientas Libre		**Herramientas Propietaria**	
	Valor Representativo	**Valor Calificativo**	**Valor Representativo**	**Valor Calificativo**
Porcentaje final	80,50%	Muy Bueno	87,50%	Muy Bueno

Elaborado por: Investigador

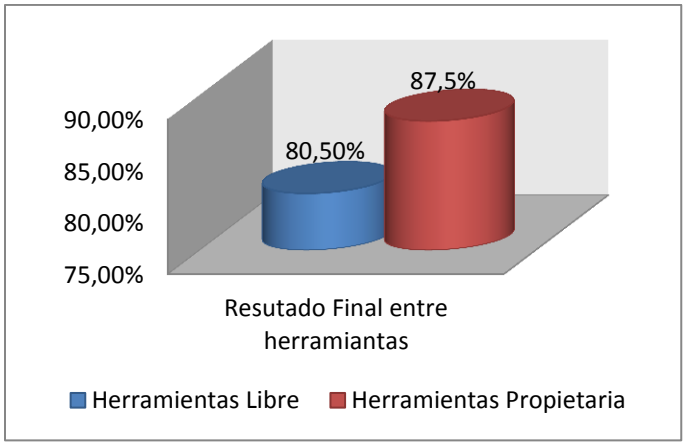


Figura III. 22.. Resultados entre herramientas.

Fuente: Investigador

Interpretación de resultados entre herramientas

En la Figura III.19 se puede visualizar que las herramientas propietarias Oracle Data Quality e Informatica han superado a las herramientas libres DataCleaner y SQL Power en este análisis de comparación con los criterios establecidos y analizados en el mismo escenario de prueba obteniendo 87,5% para las herramientas propietarias en comparación con el 80,5% que obtuvo las herramientas libres dejando como resultado que en la generación de un proyecto de calidad de datos sería recomendable la utilización de las herramientas propietarias según el estudio realizado en esta investigación.

3.6. Análisis de Resultados

Los resultados que presenta cada herramienta al establecer los mismos criterios y parámetros en el escenario de prueba son claros debido a que las herramientas Oracle Data Quality, Informatica, DataCleaner y SQL Power presentan características similares y se puede conocer la herramienta que se adapta de mejor manera a un proceso de calidad de datos.

Luego de la interpretación de los resultados realizados anteriormente se destacan las herramientas Data Integrator, Informatica y SQL Power para realizar un proceso de limpieza adecuado, ideal para desarrollar el proceso de limpieza para la base de datos OASIS con los objetivos planteados al principio de la tesis, debido a que las herramientas antes mencionadas presenta grandes características en la conexión de base de datos con soporte para varias plataformas, exploración de datos, manejo de datos más óptimo, proceso de instalación y configuración con los entornos de trabajo.

CAPÍTULO IV

APLICACIÓN DE LA METODOLOGÍA SII- ESPOCH EN LA BASE DE DATOS OASIS.

El presente capítulo tiene como objetivo aplicar la limpieza de datos y garantizar la calidad de la información en la base de datos OASIS, para lo cual se utilizará la metodología SII – ESPOCH propuesto por Margarita Isabel Solís Velasco en su trabajo "PROPUESTA METODOLÓGICA PARA LA GESTIÓN DE LA CALIDAD DE DATOS EN PROYECTOS DE INTEGRACIÓN".

La Metodología SII-ESPOCH propone las siguientes fases: Fase 1: Estudio y preparación., fase 2: Análisis de la información., fase 3: Evaluación y análisis inicial de calidad de datos., fase 4: Limpieza de datos., fase 5: Evaluación y análisis final de calidad de datos., fase 6: Mejoramiento y prevención., fase 7: Seguimiento y control.

Para el estudio, se inició, con la preparación del proyecto para lo cual se presentan todas las fuentes de datos preparados para luego pasar al proceso de evaluación de los mismos. Posterior se procede con la limpieza de datos y posteriormente la evaluación y análisis de los resultados para de esta manera establecer planes de mejora y control.

4.1. FASE I. ESTUDIO Y PREPARACIÓN

4.1.1. **Etapa 1.1** Planificar

Estructura del proyecto para la gestión de calidad de datos.

Nombre del Proyecto: Calidad de Datos SII-ESPOCH

Fecha: 01 de Febrero 2014

Realizado por: Carlos Javier Medina Benalcázar.

Recurso Humano:

Jefe del Proyecto: Ing. Paúl Paguay

Analista – Desarrollador: Carlos Javier Medina B.

Recursos Primarios:

Estación de Trabajo para desarrollo

PC Intel Core I5 8GB RAM

Herramientas de desarrollo para calidad de datos

- Perfil de datos
    - DataCleaner
- Limpieza
    - Informatica Data Quality

Objetivo:

- Gestionar Calidad de datos para la base de datos OASIS.

Propósito:

- Obtener datos de calidad en la fuente.

Justificación del proyecto:

Los datos históricos con los que cuenta la institución se encuentran almacenados con una serie de errores lo cual causa datos de mala calidad por lo que es necesario gestionar esta información de tal manera que no cause problemas en el momento de su uso.

Ámbito del Proyecto:

La información es uno de los activos más importantes de un negocio y, cada vez más, acceder a información de calidad de una manera eficaz resulta una necesidad primordial.

El presente proyecto se centra en la gestión de la calidad de datos de la base de datos OASIS que permita a las personas interesadas que laboran en la institución acceder a información precisa y confiable.[3]

Plan y cronograma de trabajo

A continuación se indica el cronograma de trabajo que se aplicará para el desarrollo del proyecto para cada una de las fases:

[3] Tomado de: http://www.partenon.net/proyectos/Descripcion_completa_Proyecto_3.pdf

	Modo de	Nombre de tarea	Duración	Comienzo	Fin	Pre
1		ESTUDIO Y PREPARACION	5 días	lun 06/01/14	vie 10/01/14	
2		ANALISIS DE LA INFORMACION	10 días	sáb 11/01/14	jue 23/01/14	
3		EVALUACION Y ANALISIS INICIAL	10 días	vie 24/01/14	jue 06/02/14	
4		LIMPIEZA DE DATOS	15 días	vie 07/02/14	jue 27/02/14	
5		EVALUACION Y ANALISIS FINAL	10 días	vie 28/02/14	jue 13/03/14	
6		MEJORAMIENTO Y PREVENCION	5 días	vie 14/03/14	jue 20/03/14	
7		SEGUIMIENTO Y CONTROL	5 días	vie 21/03/14	jue 27/03/14	

Figura IV. 1. Cronograma de Trabajo.

Fuente: Investigador

4.1.2. Etapa 1.2 Identificar el Negocio

Entre las principales tareas que se debe realizar es el estudio del negocio para saber de esta manera en que rumbo se encuentra y de qué manera se puede aportar a cumplir con los objetivos deseados.

- **Información General acerca del Negocio**

Nombre del Negocio: Escuela Superior Politécnica de Chimborazo

- **Misión:** “Ser una institución universitaria líder en la Educación Superior y en el soporte científico y tecnológico para el desarrollo socioeconómico y cultural de la provincia de Chimborazo y del país, con calidad, pertinencia y reconocimiento social".
- **Visión:** "Formar profesionales competitivos, emprendedores, conscientes de su identidad nacional, justicia social, democracia y preservación del ambiente sano, a través de la generación, transmisión, adaptación y aplicación del conocimiento científico y tecnológico para contribuir al desarrollo sustentable de nuestro país".

- **OBJETIVOS**

  - Lograr una administración moderna y eficiente en el ámbito académico, administrativo y de desarrollo institucional.
  - Establecer en la ESPOCH una organización sistémica, flexible, adaptativa y dinámica para responder con oportunidad y eficiencia a las expectativas de nuestra sociedad.
  - Desarrollar una cultura organizacional integradora y solidaria para facilitar el desarrollo individual y colectivo de los politécnicos.
  - Fortalecer el modelo educativo mediante la consolidación de las unidades académicas, procurando una mejor articulación entre las funciones universitarias.
  - Dinamizar la administración institucional mediante la desconcentración de funciones y responsabilidades, procurando la optimización de los recursos en el marco de la Ley y del Estatuto Politécnico.
  - Impulsar la investigación básica y aplicada, vinculándola con las otras funciones universitarias y con los sectores productivos y sociales.
  - Promover la generación de bienes y prestación de servicios basados en el potencial científico-tecnológico de la ESPOCH.

- **Tecnología Involucrada:** La tecnología que utiliza la institución para la administración y almacenamiento de datos es: SQL Server 2008 para el almacenamiento de la Información de la Institución.

4.2. FASE II. ANÁLISIS DE LA INFORMACIÓN

4.2.1. Etapa 2.1 Plan de captura de datos

Captura de datos

Tabla IV. I. Captura de Datos

Datos	Método de acceso	Herramientas
BD OASIS	Restauración del back up de la BD	SQL SERVER 2008

Elaborado por: Investigador

4.2.2. Etapa 2.2 Datos Disponibles

En la institución el ciclo de vida se puede evidenciar de la siguiente manera:

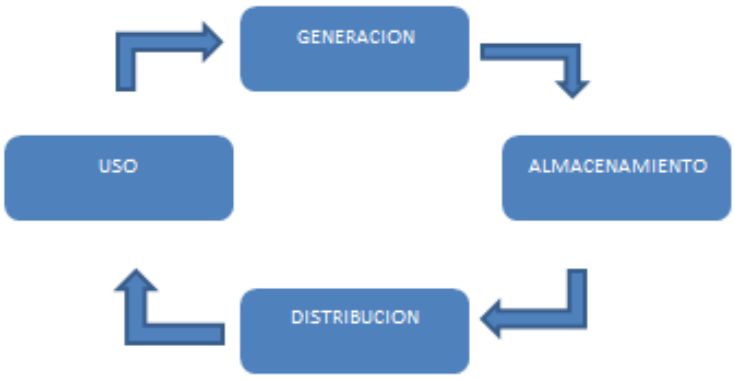


Figura IV. 2. Ciclo de Vida
Fuente: Investigador

- **Diagrama de flujo de datos**

En la figura V.6 se muestra como entidades de la institución mediante un Diagrama de Flujo de Datos - DFD de nivel 1:

Figura IV.3. Diagrama de Flujo
Fuente: Investigador

En las tablas: Tabla IV. II y Tabla IV. III que se muestran a continuación se indica el nivel de detalle del diagrama de flujo de datos expuesto anteriormente:

Tabla IV. IV. Detalle del diagrama de flujo

ALTO NIVEL	DETALLE	DETALLE COMPLETO
Datos	Ingreso, Modificación, Eliminación.	Utilización de los datos del Estudiante.

Fuente: Metodología SII-ESPOCH

Tabla IV. V. Detalle del diagrama de flujo

ALTO NIVEL	DETALLE	DETALLE COMPLETO
Secretarias	Manipulación de los datos.	Utilización de los datos del Estudiante para procesos que realiza la institución que pueden ser inscripciones y matriculación.

Fuente: Metodología SII-ESPOCH

- **Información para el flujo de datos**

Tabla IV. VI. Información para el Flujo de Datos**.**

Nombre del Departamento/Negocio	**Quien colecciona los datos**	**Que datos son coleccionados**	**Quien usa los datos**	**Donde están almacenados los datos**	**Quien es el propietario de los datos**	**Frecuencia de actualización de los datos**
ESPOCH	Administradores de base de datos	Estudiantes Docentes Periodos Notas Semestres Cursos Inscritos Estados Evaluaciones Exámenes Horarios Pensum Permisos Requisitos Materias	Secretarias de la institución	Servidores	ESPOCH	Semestral

Fuente: Metodología SII-ESPOCH

4.2.3. Etapa 2.3 Especificación de datos

- **Ámbito de Especificaciones de datos**

En la siguiente tabla se indica las especificaciones de datos con los que cuenta la institución actualmente:

Tabla IV. VII. Especificación de Datos

Especificación	Existe Especificación	Evaluar la calidad de la especificación
Estándares de datos	No	Si
Modelos de datos	Si	Si
Reglas de negocio	No	Si
Metadatos	Si	Si
Referencia de datos	No	Si

Fuente: Metodología SII-ESPOCH

- **Nivel de madurez de la calidad de datos**

La institución en cuanto a calidad de datos se encuentra en el nivel 2, ya que no existe un proyecto anterior de mejorar de calidad de datos lo cual causa problemas en los procesos administrativos. Pero para el personal que labora en la institución es de vital importancia contar de inmediato con una gestión de calidad de datos eficiente.

4.3. FASE III. EVALUACIÓN Y ANÁLISIS INICIAL DE LOS DATOS

4.3.1. Etapa 3.1 Obtención de requisitos

Para obtener los requerimientos de la calidad de datos del negocio se procedió a analizar los datos de trabajo con los involucrados en el proyecto.

- **Requerimientos de la calidad de datos**

Los requerimientos que se solicitaron en el análisis de los datos con el Ing. Paul Paguay jefe del proyecto se resume en la siguiente tabla los mismos que fueron facilitados por el administrador del Sistema OASIS:

Tabla IV. VIII. Requisitos.

N°	Problemas	Tabla(s)	Columna(s)	Requerimientos	Acción	Herramientas
1	Datos NULL y blancos	CESTUD Estudiantes Docentes Periodos Notas_Examenes Matriculas Materias Evaluaciones	CEDIDE DIRPER DIRCUR strCedula	Medir la cantidad de valores nulos o blancos.	Sustituir por valores referenciales.	-DataCleaner -SQL Power -Data Integrator -Informatica
2	Datos incompletos	CESTUD Estudiantes Docentes Periodos Notas_Examenes Matriculas Materias Evaluaciones	CEDIDE strCedula	Medir la cantidad de registros incompletos.	Valores referenciales en caso de encontrar cedulas sin la longitud exacta. Ejemplo 999999999-9	
3	Datos duplicados	CESTUD Estudiantes Docentes Periodos Notas_Examenes Matriculas Materias Evaluaciones	CEDIDE strCedula	Medir la cantidad de registros duplicados.	Eliminar los registros duplicados conservando el mayor número de información.	
4	Inconsistencia en los datos	CESTUD Estudiantes Docentes Periodos Notas_Examenes Matriculas Materias Evaluaciones	FECING	Medir el nivel de inconsistencias en los datos	Reducir las inconsistencias en el nivel que sea necesario. Valor referencial 1900-01-01	
5	Datos sin estándar.	CESTUD Estudiantes Docentes Periodos Notas_Examenes Matriculas Materias Evaluaciones	strEmail	Medir la cantidad de registros se encuentra con minúsculas y mayúsculas.	Estándar definido: Nombres y dirección con mayúsculas. Email con minúsculas.	

Elaborado por: Investigador

En los requerimientos antes revisados se tomaron en cuenta las tablas las que se detallan en la IV. IX debido a su nivel de importancia dentro de este análisis de calidad e importancia en el negocio ya que existen tablas que son solo utilizadas para la parametrización de otras tal como la tabla Sexos, Títulos, etc. Las cuales no serán tomadas en cuenta en este estudio.

Datos del sistema académico:

Tabla IV. X. Datos del Sistema

TABLA	REGISTROS
CESTUD	3597
Estudiantes	2286
Docentes	100
Periodos	56
Notas_Examenes	99099
Matriculas	16592
Materias	222
Evaluaciones	17801

Elaborado por: Investigador

4.3.2. Etapa 3.2 Medición de Datos

Perfilado de datos

Para realizar el perfilado de los datos se utilizó las siguientes herramientas software:

- Data Cleaner2.1.1 (herramienta Open Source)

Configuración:

Para empezar se necesita hacer una conexión a la base de datos con la que se va a trabajar como muestra la Figura IV.3, para las bases de datos del Sistema Académico se utiliza la Base de datos SQL Server 2008:

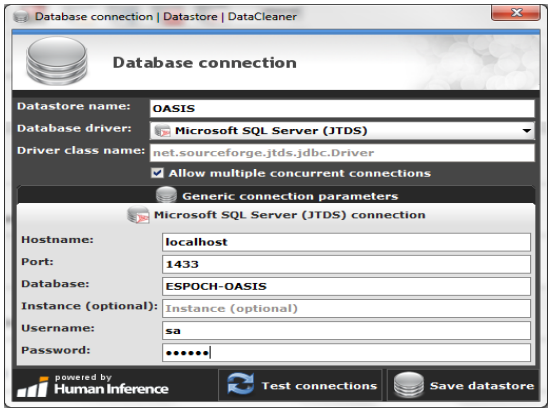


Figura IV. 3. Conexión a la Base de Datos
Fuente: Investigador

Ejecución: Para realizar el análisis de los datos se presiona el botón Analizar y se procede a escoger las columnas que se analizaran como muestra la Figura IV.4.

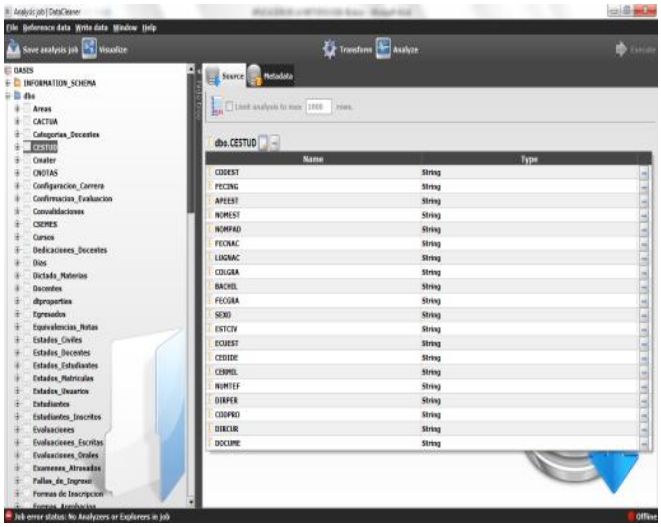


Figura IV. 4. Perfilado
Fuente: Investigador

Se puede visualizar la metadata de la tabla que se analizará y se selecciona el analizador de datos para comenzar con el perfilado de los datos como se muestra en la Figura IV.5.

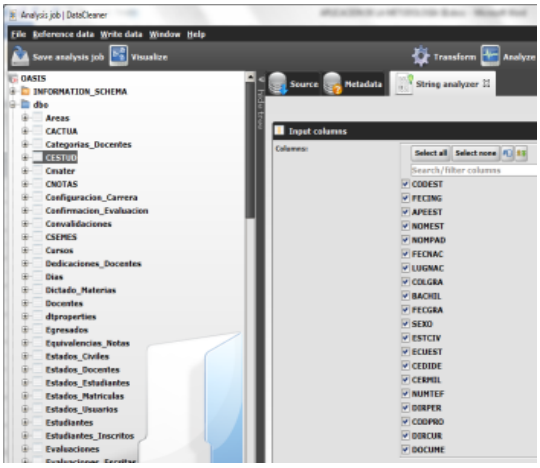


Figura IV. 5. Metadata
Fuente: Investigador

Se visualiza el resultado del análisis de los datos que se muestra en la Figura IV.6.

Figura IV. 6. Resultado del Análisis
Fuente: Investigador

Etapa 3.3 Análisis de la calidad de datos iniciales

Evaluación inicial de las fuentes de datos

Para el análisis de la calidad de los datos se tomarán en cuenta las dimensiones de calidad, las mismas que fueron utilizadas para el análisis del escenario de pruebas de las herramientas.

Los problemas que se presenten en las tablas de la base de datos OASIS afectan determinadas dimensiones de calidad, las cuales se muestras a continuación:

Tabla IV. XI. Criterios- Parámetros y Métricas
Fuente: Investigador

Criterio	Parámetro	Métrica
Consistencia.	Precisión.	Número de registros nulos o blancos.
	Valores Aceptables.	Cantidad de registros inválidos.
Integridad	Duplicidad.	Cantidad de registros duplicados.
Confiabilidad	Confianza.	(0.6) Integridad + (0.4) Consistencia.

Para medir el criterio de confiabilidad de los datos de cada tabla de la base de datos se tomarán en cuenta las siguientes fórmulas:

Parámetro de Precisión.

Porcentaje de registros no nulos ni blancos dentro de un campo en una tabla.

Parámetro de Valores Aceptables.

Porcentaje de registros con valores adecuados dentro de un campo en una tabla.

Parámetro de Duplicidad

Porcentaje de registros únicos dentro de una tabla.

Parámetro de Confianza.

Para el análisis de la confianza se utilizará las fórmulas planteadas y utilizadas en el Capítulo III.

Donde se tiene la siguiente fórmula:

CFD = (0.6) **INT** + (0.4) **CON**

Dónde:

CFD = CONFIABILIDAD

INT = INTEGRIDAD

CON = CONSISTENCIA

La consistencia se la puede obtener del promedio de valores obtenidos entre la precisión y los valores válidos, de lo cual tenemos.

CON = (**P**+**VA**)/2

Dónde:

P = PRECISION

VA= VALORES ACEPTABLES

4.3.3. Análisis preliminar de la calidad de los datos de la base OASIS

Para obtener los resultados totales del análisis de cada tabla de manera más explícita al revisar el ANEXO 2.

- **CESTUD**.

En la siguiente tabla se muestran los resultados del análisis de la calidad de datos para la tabla CESTUD.

Tabla IV. XII. Calidad de Datos en tabla CESTUD

TABLA CESTUD			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6) INTEGRIDAD + (0.4) CONSISTENCIA
APEEST	99%	99,81%	99,32%
FECING	99%	99,45%	99,18%
NOMPAD	99,05%	98,70%	98,91%
LUGNAC	99%	72,24%	88,30%
COLGRA	99%	99,99%	99,40%
BACHIL	99%	99,36%	99,14%
FECGRA	99%	32,47%	72,39%
CEDIDE	99%	50,08%	79,43%
CERMIL	99%	99,89%	99,36%
NUMTEF	99,19%	99,38%	99,27%
DIRPER	99%	50,00%	79,40%
DIRCUR	99%	87,06%	94,22%
DOCUME	99%	99,46%	99,18%
CODEST	99,11%	99,81%	99,39%

Elaborado por: Investigador

Resultados finales.

Tabla IV. XIII. Resultado Final

TABLA CESTUD	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	99,03%
CONSISTENCIA	84,84%
CONFIABILIDAD	93,35%

Elaborado por: Investigador

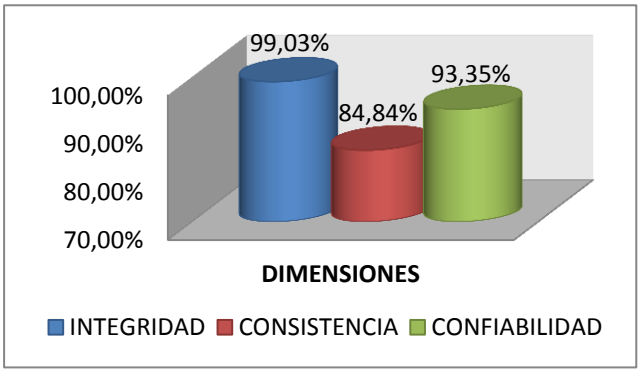


Figura IV. 7. Resultado Final de la Tabla CESTUD
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla CESTUD y observando la Tabla IV.X se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos como muestra la Figura IV.8, pero los cuales se podrían mejorar con el uso de las herramientas dedicadas a esta función.

- **DOCENTES**

En la siguiente tabla se muestran los resultados del análisis de la calidad de datos para la tabla Docentes.

Tabla IV. XIV. Resultados del Análisis - Tabla Docentes

TABLA DOCENTES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6) INTEGRIDAD + (0.4) CONSISTENCIA
strCedulaMil	100%	53,50%	81,40%
strCarnetSeg	100%	52,50%	81,00%
strDireccion	100%	82,00%	92,80%
strTel	100%	79,00%	91,60%
strMail	100%	73,26%	89,30%
strWww	100%	70,50%	88,20%
strCodTipoSan	100%	50,00%	80,00%
strCodEstCiv	100%	99,00%	99,60%
strTitulos	100%	51,00%	80,40%
strCargos	100%	51,00%	80,40%
strCodTipTit	100%	95,00%	98,00%
strNacionalidad	100%	99,98%	99,99%
strNombres	100%	99,52%	99,81%
strApellidos	100%	99,50%	99,80%

Elaborado por: Investigador

Resultados finales.

Tabla IV. XV. Resulta Final

TABLA DOCENTES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	75,41%
CONFIABILIDAD	90,16%

Elaborado por: Investigador

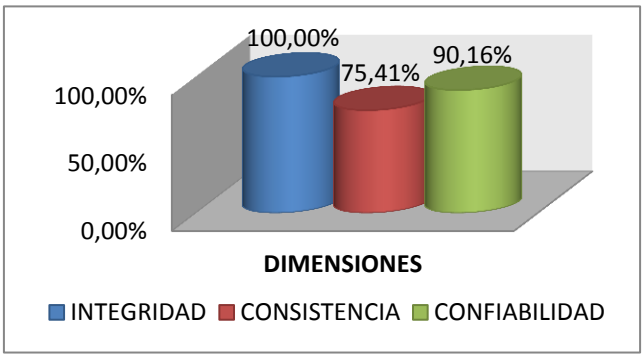


Figura IV. 8. Resultado Final de la Tabla docente
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla Docentes y observando la Tabla IV.XII se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos como muestra la figura IV.9, excepto para la dimensión de consistencia, pero los cuales se podrían mejorar con el uso de las herramientas dedicadas a esta función.

- **ESTUDIANTES**

En la siguiente tabla se muestran los resultados del análisis de la calidad de datos para la tabla Estudiantes.

Tabla IV. XVI. Resultados del Análisis - Tabla Estudiante

TABLA ESTUDIANTES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6)INTEGRIDAD + (0.4)CONSISTENCIA
FECING	100%	99,37%	99,75%
strMail	100%	73,63%	89,45%
strDocumentacion	100%	50,00%	80,00%
strCodTit	100%	75,35%	90,14%
strCedulaMil	100%	99,15%	99,66%
strCodInt	100%	75,35%	90,14%
strNacionalidad	100%	99,98%	99,99%
strNombres	100%	99,52%	99,81%
strApellidos	100%	99,50%	99,80%

Elaborado por: Investigador

Resultados finales.

Tabla IV. XVII. Resultados Finales

TABLA ESTUDIANTES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	85,76%
CONFIABILIDAD	94,30%

Elaborado por: Investigador

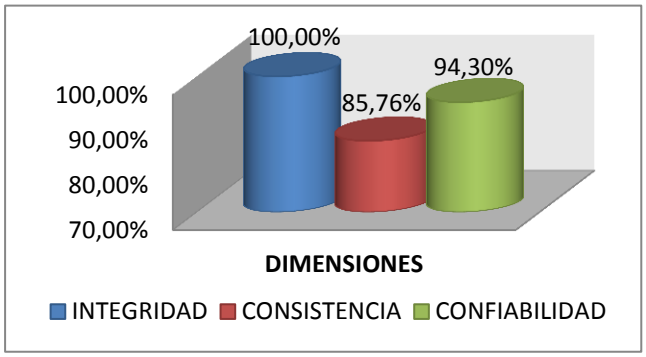


Figura IV. 9. Resultado final Tabla Estudiantes
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla Estudiante y observando la Tabla IV.XIV se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos como muestra la figura IV.10, excepto para la dimensión de consistencia, pero los cuales se podrían mejorar con el uso de las herramientas dedicadas a esta función.

- **MATERIAS**

En la siguiente tabla se muestran los resultados del análisis de la calidad de datos para la tabla Materias.

Tabla IV. XVIII. Resultados del Análisis - Tabla Materias

TABLA MATERIAS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6)INTEGRIDAD + (0.4)CONSISTENCIA
srtCodigo	100%	100%	100%
strNombre	100%	100%	100%
dtFechaCreada	100%	100%	100%
dtFechaElim	100%	70,72%	88,29%
blnActiva	100%	100%	100%

Elaborado por: Investigador

Resultados finales.

Tabla IV. XIX. Resultados Finales

TABLA MATERIAS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	94,14%
CONFIABILIDAD	97,66%

Elaborado por: Investigador

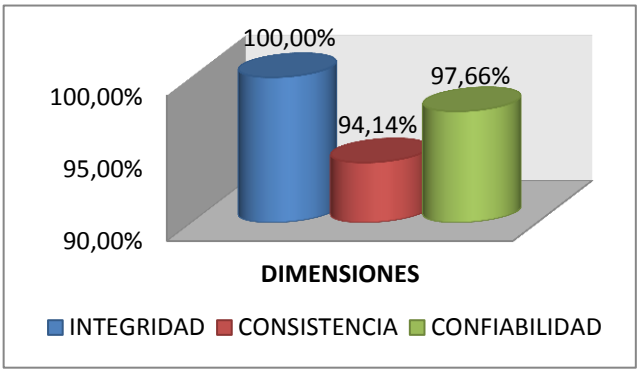


Figura IV. 10. Resultados Final- Tabla Materias
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla Materias y observando la Tabla IV.XVI se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos como muestra la figura IV.11, excepto para la dimensión de consistencia, pero los cuales se podrían mejorar con el uso de las herramientas dedicadas a esta función.

- **MATRICULAS**

En la siguiente tabla se muestran los resultados del análisis de la calidad de datos para la tabla Matriculas.

Tabla IV. XX. Resultados del Análisis - Tabla Matriculas
Fuente: Investigador

TABLA MATRICULAS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6)INTEGRIDAD + (0.4)CONSISTENCIA
sintCodigo	100%	100%	100%
strCodPeriodo	100%	100%	100%
strCodEstud	100%	100%	100%
strCodNivel	100%	100%	100%
strAutorizadaPor	100%	96,80%	98,72%
dtFechaAutorizada	100%	100%	100%
strCreadaPor	100%	96,80%	98,72%
dtFechaCreada	100%	100%	100%
strCodEstado	100%	100%	100%
strObservaciones	100%	50,03%	80,01%

Elaborado por: Investigador

Resultados finales.

Tabla IV. XXI.Resultado Final

TABLA MATRICULAS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	94,36%
CONFIABILIDAD	97,75%

Elaborado por: Investigador

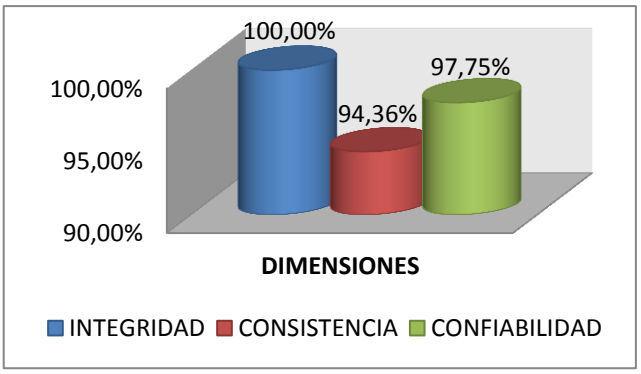


Figura IV. 11. Resultado final - Tabla Matricula
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla Matriculas y observando la Tabla IV.XVIII se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos como muestra la figura IV.12, excepto para la dimensión de consistencia, pero los cuales se podrían mejorar con el uso de las herramientas dedicadas a esta función.

- **EVALUACIONES**

En la siguiente tabla se muestran los resultados del análisis de la calidad de datos para la tabla Evaluaciones.

Tabla IV. XXII. Resultados del Análisis - Tabla Evaluaciones

TABLA EVALUACIONES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6)INTEGRIDAD + (0.4)CONSISTENCIA
sintCodMatricula	100%	100%	100%
strCodPeriodo	100%	100%	100%
strCodMateria	100%	100%	100%
bytNota1	100%	100%	100%
bytNota2	100%	100%	100%
bytNota3	100%	100%	100%
strObservaciones	100%	50%%	75%

Elaborado por: Investigador

Resultados finales.

Tabla IV. XXIII. Resultado Final

TABLA EVALUACIONES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	92,86%
CONFIABILIDAD	96,43%

Elaborado por: Investigador

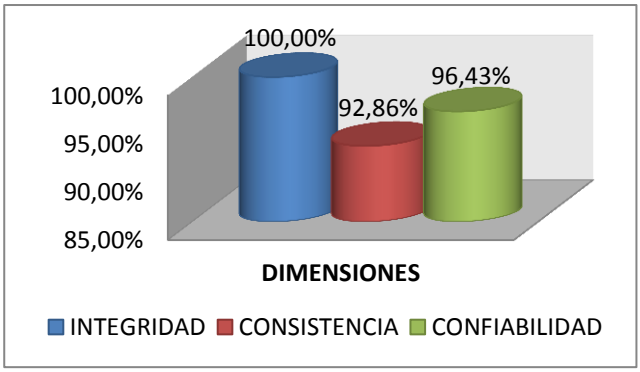


Figura IV. 12. Resultado final
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla Evaluaciones y observando la Tabla IV.XX se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos como muestra la figura IV.13, excepto para la dimensión de consistencia, pero los cuales se podrían mejorar con el uso de las herramientas dedicadas a esta función.

- **NOTAS EXAMENES**

En la siguiente tabla se muestran los resultados del análisis de la calidad de datos para la tabla Notas_Examenes.

Tabla IV. XXIV. Resultados del Análisis - Tabla Notas Exámenes

TABLA NOTAS_EXAMENES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6)INTEGRIDAD + (0.4)CONSISTENCIA
sintCodMatricula	100%	100%	100%
strCodPeriodo	100%	100%	100%
strCodMateria	100%	100%	100%
strCodTipoExamen	100%	100%	100%
bytAcumulado	100%	100%	100%
bytNota	100%	99,14%	99,67%
strCodEquiv	100%	100%	100%
strObservaciones	100%	52,69%	81,08%

Elaborado por: Investigador

Resultados finales.

Tabla IV. XXV. Resultado Final

TABLA NOTAS_EXAMENES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	93,98%
CONFIABILIDAD	97,59%

Elaborado por: Investigador

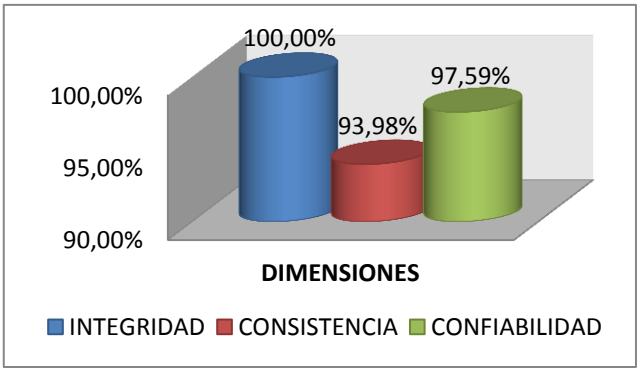


Figura IV. 13. Resultado final - Tabla Notas Exámenes
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla Notas_Examenes y observando la Tabla IV.XXII se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos como muestra la figura IV.14, excepto para la dimensión de consistencia, pero los cuales se podrían mejorar con el uso de las herramientas dedicadas a esta función.

- **PERIODOS**

En la Tabla IV. XXVI se muestran los resultados del análisis de la calidad de datos para la tabla Periodos.

Tabla IV. XXVII. Resultados del Análisis - Tabla Periodos

TABLA PERIODOS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6)INTEGRIDAD + (0.4)CONSISTENCIA
strCodigo	100%	100%	100%
strDescripcion	100%	100%	100%
dtFechaInic	100%	100%	100%
dtFechaFin	100%	100%	100%
sintUltNumMat	100%	100%	100%
strCodPensum	100%	100%	100%
blnTransicion	100%	100%	100%
blnVigente	100%	100%	100%
dtFechaTopeMatOrd	100%	58,93%	83,57%
dtFechaTopeMatExt	100%	58,93%	83,57%
dtFechaTopeMatPro	100%	58,93%	83,57%
dtFechaTopeRetMat	100%	58,93%	83,57%
strCodReglamento	100%	100%	100%

Elaborado por: Investigador

Resultados finales.

Tabla IV. XXVIII. Resultado Final

TABLA PERIODOS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	87,36%
CONFIABILIDAD	94,94%

Elaborado Por: Investigador

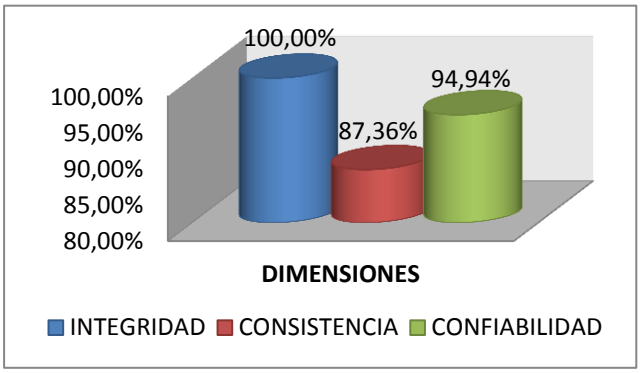


Figura IV. 14. Resultado final - Tabla Períodos
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla Periodos y observando la Tabla IV. XXIX se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos como muestra la figura IV.15, excepto para la dimensión de consistencia, pero los cuales se podrían mejorar con el uso de las herramientas dedicadas a esta función.

4.3.4. Etapa 3.4 Políticas Internas De Calidad

Tabla IV. XXX. Políticas internas de calidad

N°	Problemas	Tabla(s)	Columna(s)	Políticas
1	Datos NULL y blancos	CESTUD Estudiantes Docentes Periodos Notas_Examenes Matriculas Materias Evaluaciones	CEDIDE DIRPER DIRCUR strCedula	Sustituir por valores referenciales. CEDIDE=999999999-9 StrCedula=999999999-9 DIRPER, DIRCUR y strObservaciones se remplazara con la palabra DESCONOCIDO
2	Datos incompletos	CESTUD Estudiantes Docentes Periodos Notas_Examenes Matriculas Materias Evaluaciones	CEDIDE strCedula	Valores referenciales en caso de encontrar cedulas sin la longitud exacta. Ejemplo 999999999-9 Para valores numéricos la referencia será 0
3	Datos duplicados	CESTUD Estudiantes Docentes Periodos Notas_Examenes Matriculas Materias Evaluaciones	CEDIDE strCedula	Eliminar los registros duplicados conservando el mayor número de información.
4	Inconsistencia en los datos	CESTUD Estudiantes Docentes Periodos Notas_Examenes Matriculas Materias Evaluaciones	FECING strTel	Reducir las inconsistencias en el nivel que sea necesario. Valor referencial 1900-01-01 Telefono:9999999
5	Datos sin estándar.	CESTUD Estudiantes Docentes Periodos Notas_Examenes Matriculas Materias Evaluaciones	strEmail strNomres strApellidos	Estándar definido: Nombres y dirección con mayúsculas. Email con minúsculas.

Fuente: Metodología SII-ESPOCH

4.4. FASE IV. LIMPIEZA DE DATOS

4.4.1. Etapa Limpieza

Ejecución de Limpieza

Para la limpieza de datos se utilizó la siguiente herramienta la cual fue elegida de acuerdo al resultado del estudio previo realizado:

- Informatica Data Quality

Limpieza de los datos utilizando las herramientas para calidad de datos.

Para la realización de la limpieza se tomó en cuenta todas las tablas antes mencionadas, los resultados de la limpieza de cada una de las tablas se las puede visualizar más detalladamente en el ANEXO 2.

4.5. FASE V. EVALUACIÓN Y ANÁLISIS FINAL DE LA CALIDAD DE DATOS

4.5.1. Etapa 5.1 Evaluación Final de los datos

El resumen de los resultados obtenidos después de la limpieza con la herramienta se muestra en la Tabla IV. XXXI.

- **CESTUD**

Tabla IV. XXXII. Resultados obtenidos después de la limpieza

TABLA CESTUD			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6)INTEGRIDAD + (0.4)CONSISTENCIA
APEEST	100%	100%	100%
FECING	100%	100%	100%
NOMPAD	100%	100%	100%
LUGNAC	100%	100%	100%
COLGRA	100%	100%	100%
BACHIL	100%	100%	100%
FECGRA	100%	100%	100%
CEDIDE	100%	100%	100%
CERMIL	100%	100%	100%
NUMTEF	100%	100%	100%
DIRPER	100%	100%	100%
DIRCUR	100%	100%	100%
DOCUME	100%	100%	100%
CODEST	100%	100%	100%

Elaborado por: Investigador

Resultados finales.

Tabla IV. XXXIII. Resultado Final

TABLA CESTUD	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

Elaborado por: Investigador

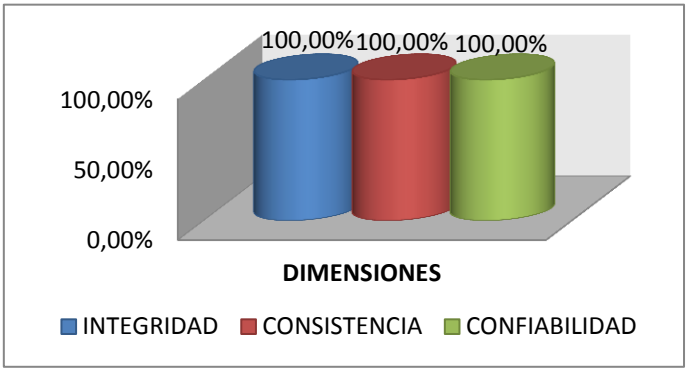


Figura IV. 15. Resultado limpieza - Tabla CESTUD
Fuente: Investigador

Interpretación de los resultados finales.

Como se puede observar en la Tabla XXXIV y en la Figura IV.16 las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla CESTUD, con un nivel de integridad del 100% en

comparación al análisis preliminar de los datos que se puede visualizar en la Tabla IV.X en el cual se obtuvo 99,03% mejorándolo en un 1,97%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación al 84,84% obtenido en el análisis preliminar mejorando a esta dimensión en un 15,16% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 93,35% obtenido en el análisis preliminar optimizando esta dimensión en 6,65% dejando a la tabla con un mejor nivel de conformidad de datos.

- **DOCENTES**

Tabla IV. XXXV. Resultados de Limpieza - Tabla Docentes

TABLA DOCENTES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6)INTEGRIDAD + (0.4)CONSISTENCIA
strCedulaMil	100%	100%	100%
strCarnetSeg	100%	100%	100%
strDireccion	100%	100%	100%
strTel	100%	100%	100%
strMail	100%	100%	100%
strWww	100%	100%	100%
strCodTipoSan	100%	100%	100%
strCodEstCiv	100%	100%	100%
strTitulos	100%	100%	100%
strCargos	100%	100%	100%
strCodTipTit	100%	100%	100%
strNacionalidad	100%	100%	100%
strNombres	100%	100%	100%
strApellidos	100%	100%	100%

Elaborado por: Investigador

Resultados finales.

Tabla IV. XXXVI. Resultado Final

TABLA DOCENTES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

Elaborado por: Investigador

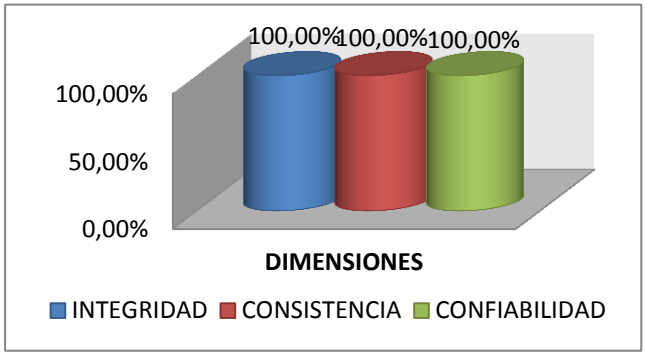


Figura IV. 16. Resultado limpieza - Tabla Docentes
Fuente: Investigador

Interpretación de los resultados finales.

Como se puede observar en la Tabla IV. XXXVIIy en la Figura IV.17 las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla Docentes, con un nivel de integridad del 100% al igual que en el análisis preliminar de los datos que se puede visualizar en la Tabla IV.XII en el cual se obtuvo 100%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación al 75,41% obtenido en el análisis preliminar mejorando a esta dimensión en un 24,59% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 90,16% obtenido en el análisis preliminar mejorando esta dimensión en 9,84% dejando a la tabla con un mejor nivel de conformidad de datos.

- **ESTUDIANTES**

Tabla IV. XXXVIII. Resultados de Limpieza - Tabla Estudiantes

TABLA ESTUDIANTES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6)INTEGRIDAD + (0.4)CONSISTENCIA
FECING	100%	100%	100%
strMail	100%	100%	100%
strDocumentacion	100%	100%	100%

strCodTit	100%	100%	100%
strCedulaMil	100%	100%	100%
strCodInt	100%	100%	100%
strNacionalidad	100%	100%	100%
strNombres	100%	100%	100%
strApellidos	100%	100%	100%

Elaborado por: Investigador

Resultados finales.

Tabla IV. XXXIX. Resultado Final

TABLA ESTUDIANTES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

Elaborado por: Investigador

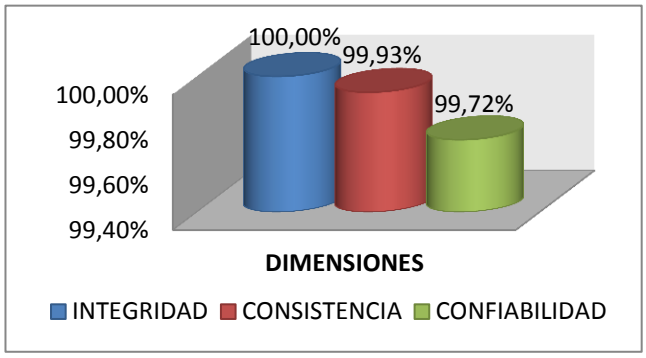


Figura IV. 17. Resultado limpieza – Tabla Estudiante
Fuente: Investigador

Interpretación de los resultados finales.

Como se puede observar en la Tabla IV. XL y en la Figura IV.18 las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla Estudiantes, con un nivel de integridad del 100% al igual que en el análisis preliminar de los datos el cual se puede visualizar en la Tabla IV.XIV en el cual se obtuvo 100%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación al 85,76% obtenido en el análisis preliminar mejorando a esta dimensión en un 14,24% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 94,30% obtenido en el análisis preliminar mejorando esta dimensión en 5,7% dejando a la tabla con un mejor nivel de conformidad de datos.

- **MATERIAS**

Tabla IV. XLI. Resultados de Limpieza - Tabla Materias

TABLA MATERIAS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6)INTEGRIDAD + (0.4)CONSISTENCIA
srtCodigo	100%	100%	100%
strNombre	100%	100%	100%
dtFechaCreada	100%	100%	100%
dtFechaElim	100%	100%	100%
blnActiva	100%	100%	100%

Elaborado por: Investigador

Resultados finales.

Tabla IV. XLII. Resultado Final

TABLA MATERIAS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

Elaborado por: Investigador

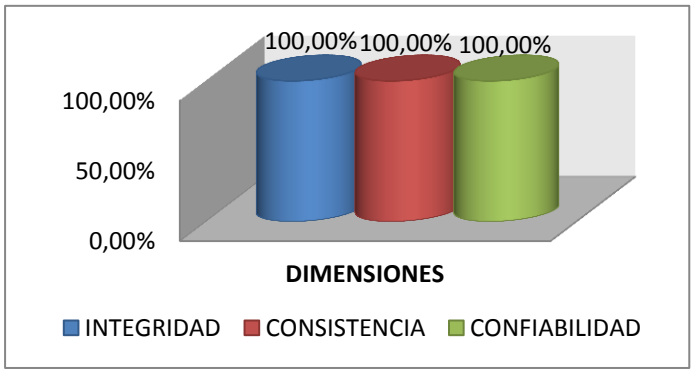


Figura IV. 18. Resultado limpieza - Tabla Materias
Fuente: Investigador

Interpretación de los resultados finales.

Como se puede observar en la Tabla IV. XLIII y en la Figura IV.19 las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla Materias, con un nivel de integridad del 100% al igual que en el análisis preliminar de los datos el cual se puede visualizar en la Tabla IV.XVI en el cual se obtuvo 100%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación al 94,14% obtenido en el análisis preliminar mejorando a esta dimensión en un 5,86% y en la dimensión de confiabilidad

de esta tabla obteniendo 100% después de la limpieza en comparación con el 97,66% obtenido en el análisis preliminar mejorando esta dimensión en 2,34% dejando a la tabla con un mejor nivel de conformidad de datos.

- **MATRICULAS**

Tabla IV. XLIV. Resultados de Limpieza - Tabla Matriculas

TABLA MATRICULAS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6)INTEGRIDAD + (0.4)CONSISTENCIA
sintCodigo	100%	100%	100%
strCodPeriodo	100%	100%	100%
strCodEstud	100%	100%	100%
strCodNivel	100%	100%	100%
strAutorizadaPor	100%	100%	100%
dtFechaAutorizada	100%	100%	100%
strCreadaPor	100%	100%	100%
dtFechaCreada	100%	100%	100%
strCodEstado	100%	100%	100%
strObservaciones	100%	100%	100%

Elaborado por: Investigador

Resultados finales.

Tabla IV. XLV. Resultado Final

TABLA MATRICULAS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

Elaborado por: Investigador

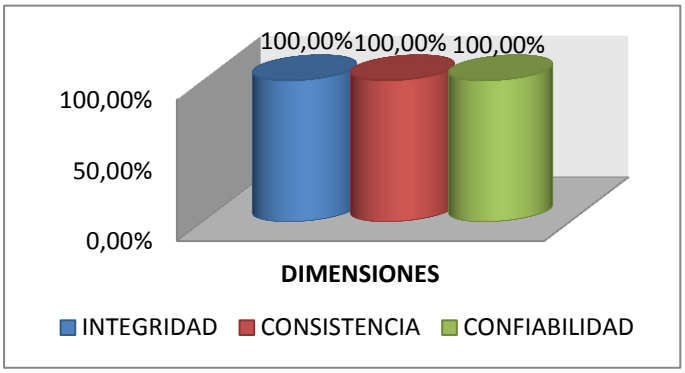


Figura IV. 19. Resultado limpieza - Tabla Matriculas
Fuente: Investigador

Interpretación de los resultados finales.

Como se puede observar en la Tabla IV. XLVI y en la Figura IV.20 las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad

mucho más alta a la tabla Matriculas, con un nivel de integridad del 100% al igual que en el análisis preliminar de los datos el cual se puede visualizar en la Tabla IV.XVIII en el cual se obtuvo 100%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación al 94,36% obtenido en el análisis preliminar mejorando a esta dimensión en un 5,64% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 97,75% obtenido en el análisis preliminar mejorando esta dimensión en 2,25% dejando a la tabla con un mejor nivel de conformidad de datos.

- **EVALUACIONES**

Tabla IV. XLVII. Resultados de Limpieza - Tabla Evaluaciones
Fuente: Investigador

TABLA EVALUACIONES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6)INTEGRIDAD + (0.4)CONSISTENCIA
sintCodMatricula	100%	100%	100%
strCodPeriodo	100%	100%	100%
strCodMateria	100%	100%	100%
bytNota1	100%	100%	100%
bytNota2	100%	100%	100%
bytNota3	100%	100%	100%
strObservaciones	100%	100%	100%

Elaborado por: Investigador

Resultados finales.

Tabla IV. XLVIII. Resultado Final

TABLA EVALUACIONES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

Elaborado por: Investigador

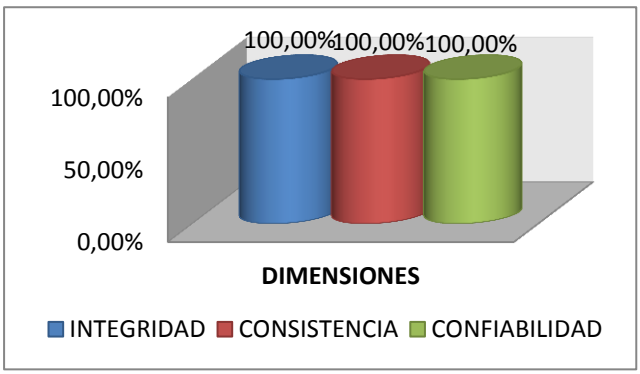


Figura IV. 20. Resultado limpieza – Tabla Evaluaciones
Fuente: Investigador

Interpretación de los resultados finales.

Como se puede observar en la Tabla IV. XLIX y en Figura IV.21 las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla Evaluaciones, con un nivel de integridad del 100% al igual que en el análisis preliminar de los datos el cual se puede visualizar en la Tabla IV.XX en el cual se obtuvo 100%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación al 92,86% obtenido en el análisis preliminar mejorando a esta dimensión en un 7,14% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 96,43% obtenido en el análisis preliminar mejorando esta dimensión en 3,57% dejando a la tabla con un mejor nivel de conformidad de datos.

- **NOTAS_EXAMENES**

Tabla IV. L. Resultados de Limpieza - Tabla Notas Exámenes

TABLA NOTAS_EXAMENES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6)INTEGRIDAD + (0.4)CONSISTENCIA
sintCodMatricula	100%	100%	100%
strCodPeriodo	100%	100%	100%
strCodMateria	100%	100%	100%
strCodTipoExamen	100%	100%	100%
bytAcumulado	100%	100%	100%
bytNota	100%	100%	100%
strCodEquiv	100%	100%	100%
strObservaciones	100%	100%	100%

Elaborado por: Investigador

Resultados finales.

Tabla IV. LI. Resultado Final

TABLA NOTAS_EXAMENES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

Elaborado por: Investigador

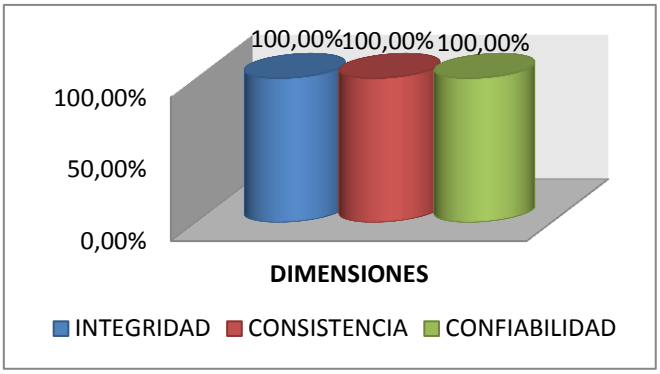


Figura IV. 21. Resultado limpieza - Tabla Notas Exámenes

Fuente: Investigador

Interpretación de los resultados finales.

Como se puede observar en la Tabla IV. LIIy en la Figura IV.22 las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla Notas_Examenes, con un nivel de integridad del 100% al igual que en el análisis preliminar de los datos el cual se puede visualizar en la Tabla IV.XXII en el cual se obtuvo 100%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación al 93,98% obtenido en el análisis preliminar mejorando a esta dimensión en un 6,02% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 97,59% obtenido en el análisis preliminar mejorando esta dimensión en 2,41% dejando a la tabla con un mejor nivel de conformidad de datos.

- **PERIODOS**

Tabla IV. LIII. Resultados de Limpieza - Tabla Periodos

TABLA PERIODOS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** (0.6)INTEGRIDAD + (0.4)CONSISTENCIA
strCodigo	100%	100%	100%
strDescripcion	100%	100%	100%
dtFechaInic	100%	100%	100%
dtFechaFin	100%	100%	100%
sintUltNumMat	100%	100%	100%
strCodPensum	100%	100%	100%
blnTransicion	100%	100%	100%
blnVigente	100%	100%	100%
dtFechaTopeMatOrd	100%	100%	100%
dtFechaTopeMatExt	100%	100%	100%
dtFechaTopeMatPro	100%	100%	100%
dtFechaTopeRetMat	100%	100%	100%
strCodReglamento	100%	100%	100%

Elaborado por: Investigador

Resultados finales.

Tabla IV. LIV. Resultado Final

TABLA PERIODOS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

Elaborado por: Investigador

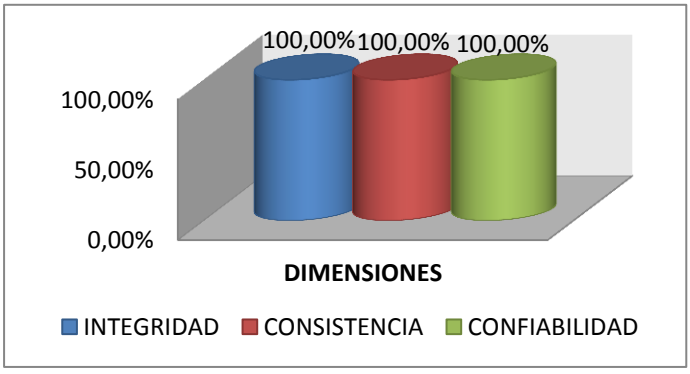


Figura IV. 22. Resultado limpieza - Tabla Períodos
Fuente: Investigador

Interpretación de los resultados finales.

Como se puede observar en la Tabla IV. LV y en la Figura IV.23 las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla Periodos, con un nivel de integridad del 100% al igual que en

el análisis preliminar de los datos el cual se puede visualizar en la Tabla IV.XXIV en el cual también se obtuvo 100%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación con el 87,36% obtenido en el análisis preliminar mejorando a esta dimensión en un 12,64% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 94,94% obtenido en el análisis preliminar mejorando esta dimensión en 5,06% dejando a la tabla con un mejor nivel de conformidad de datos.

4.6. FASE VI. MEJORAMIENTO Y PREVENCIÓN

4.6.1. Etapa 6.1 Analizar Causas de origen

Tabla IV. LVI. Causas de origen

Problema	**Causa**
Datos NULL y blancos	Falta de información por parte del estudiante
Datos incompletos	Falta de información por parte del estudiante
Datos duplicados	Falta de controles en la aplicación software para el ingreso de datos
Datos sin formatos necesarios	Falta de políticas establecidas antes de la construcción del software para el ingreso de datos y del diseño de la base de datos
Datos sin un estándar especifico	Falta de políticas establecidas antes de la construcción del software para el ingreso de datos y del diseño de la base de datos
Inconsistencias en los datos	Falta de políticas establecidas antes de la construcción del software para el ingreso de datos y del diseño de la base de datos Digitación errónea por parte del personal involucrado

Fuente: Metodología SII-ESPOCH

4.6.2. Etapa 6.2 Diseñar plan de mejoramiento

Tabla IV. LVII. Plan de mejoramiento

Problema	**Acción**	**Viabilidad**	**Responsable de verificación**
Datos NULL y blancos	Mejorar el proceso de matriculación.	100%	Autoridades. Equipo de desarrollo. Equipo de gestión de calidad de datos
Datos incompletos	Mejorar el proceso de matriculación. Exigir datos completos a los estudiantes.	100%	Equipo de desarrollo. Equipo de gestión de calidad de datos
Datos duplicados	Desarrollar mejoras para el software de matriculación.	100%	Equipo de desarrollo.
Datos sin formatos necesarios	Desarrollar mejoras para el software de matriculación.	100%	Equipo de desarrollo.
Datos sin un estándar especifico	Desarrollar mejoras para el software de matriculación, en la manipulación de los datos.	100%	Equipo de desarrollo.
Inconsistencias en los datos	Desarrollar mejoras para el software de matriculación, en la manipulación de los datos.	100%	Equipo de desarrollo.

Fuente: Metodología SII-ESPOCH

4.7. FASE VII. SEGUIMIENTO Y CONTROL

4.7.1. Etapa 7.1 Diseñar Plan de seguimiento y control

Se detalla el plan de seguimiento y control en la Tabla IV. LVIII.

Plan de seguimiento y control

Tabla IV. LIX. Plan de seguimiento y control

Problema	**Causa**	**Proceso de control**	**Responsable**	**Frecuencia**
Datos NULL y blancos	Falta de información por parte del estudiante	Perfilamiento de datos. Análisis de la calidad de datos.	Equipo de gestión de calidad de datos	Semestral
Datos incompletos	Falta de información por parte del estudiante	Perfilamiento de datos. Análisis de la calidad de datos.	Equipo de gestión de calidad de datos	Semestral
Datos duplicados	Falta de controles en la aplicación software para el ingreso de datos	Perfilamiento de datos. Análisis de la calidad de datos.	Equipo de gestión de calidad de datos	Semestral
Datos sin formatos necesarios	Falta de políticas establecidas antes de la construcción del software para el ingreso de datos y del diseño de la base de datos	Perfilamiento de datos. Análisis de la calidad de datos.	Equipo de gestión de calidad de datos	Semestral
Datos sin un estándar especifico	Falta de políticas establecidas antes de la construcción del software para el ingreso de datos y del diseño de la base de datos	Perfilamiento de datos. Análisis de la calidad de datos.	Equipo de gestión de calidad de datos	Semestral
Inconsistencias en los datos	Falta de políticas establecidas antes de la construcción del software para el ingreso de datos y del diseño de la base de datos Digitación errónea por parte del personal involucrado	Perfilamiento de datos. Análisis de la calidad de datos.	Equipo de gestión de calidad de datos	Semestral

Fuente: Metodología SII-ESPOCH

4.8. COMPROBACIÓN DE HIPÓTESIS

La hipótesis planteada es:

H1: El uso de las herramientas para Data Quality permite mejorar la confiabilidad y consistencia de los datos en la base de datos “OASIS”.

Tipo de hipótesis

La hipótesis de esta investigación es de tipo Causa – Efecto

Determinación de variables

- **Variable Independiente:** Uso de las herramientas para Data Quality.
- **Variable Dependiente:** Mejorar la confiabilidad y consistencia de los datos.

Operacionalización Conceptual

Tabla IV. LX. Operacionalización
Fuente: Investigador

VARIABLE	TIPO	CONCEPTO
Uso de las herramientas para Data Quality	Independiente	Las herramientas para la calidad de datos son soluciones dedicadas a detectar y corregir problemas en los datos dentro de un almacén de datos, aquellos datos que afectan a la precisión y la eficiencia de las aplicaciones de análisis de datos.
Mejorar la confiabilidad y consistencia de los datos en la base de datos “OASIS”.	Dependiente	Son indicadores de precisión, fiabilidad y veracidad de la información dentro de los almacenes de datos.

4.8.1. Operacionalización Metodológica.

Tabla IV. LXI. Operacionalización Metodológica

Hipótesis	Variables	Criterio	Indicadores	Técnicas	Fuentes de verificación
El uso de las herramientas para Data Quality permite mejorar la confiabilidad y consistencia de los datos en la base de datos "OASIS"	Uso de las herramientas para Data Quality.	Investigación	Herramientas: SQL Power, Data Integrator e Informatica.	Revisión de documentos	Internet. Manuales. Guías de Usuario.
	Mejorar la confiabilidad y consistencia de los datos	Consistencia.	Precisión.	Observación	Limpieza de los datos. Aplicación metodológica.
			Valores Aceptables.	Observación	Limpieza de los datos. Aplicación metodológica.
		Integridad.	Duplicidad.	Observación	Limpieza de los datos. Aplicación metodológica.
		Confiabilidad.	(0,6)INTEGRIDA + (0,4)CONSISTENCIA	Observación	Limpieza de los datos. Aplicación metodológica.

Fuente: Metodología SII-ESPOCH

Para la demostración de la hipótesis se utilizó los datos de las tablas de resultados obtenidas en el análisis previo al uso de las herramientas y las que se obtuvieron en el análisis posterior.

4.8.2. Resultados generales por parámetros

Resultado general de los datos sin utilizar las herramientas de limpieza.

Tabla IV. LXII. Resultados generales por parámetro.

DIMENSIONES DE CALIDAD	ANALISIS PRELIMINAR DE LOS DATOS							
	CESTUD	**Estudiantes**	**Docentes**	**Periodos**	**Notas_Examenes**	**Matriculas**	**Materias**	**Evaluaciones**
INTEGRIDAD	99,03%	100%	100%	100%	100%	100%	100%	100%
CONSISTENCIA	84,84%	85,76%	75,41%	87,36%	93,98%	94,36%	94,14%	92,86%
CONFIABILIDAD	93,35%	94,30%	90,16%	94,94%	97,59%	97,75%	97,66%	96,43%

Elaborado por: Investigador

La Tabla IV. LXIIImuestra un resumen de los datos consolidados del análisis previo de los datos.

Resultado general de los datos al utilizar las herramientas de limpieza.

Tabla IV. LXIV. Resultados generales por parámetro.

DIMENSIONES DE CALIDAD	ANALISIS FINAL DE LOS DATOS							
	CESTUD	**Estudiantes**	**Docentes**	**Periodos**	**Notas_Examenes**	**Matriculas**	**Materias**	**Evaluaciones**
INTEGRIDAD	100%	100%	100%	100%	100%	100%	100%	100%
CONSISTENCIA	100%	100%	100%	100%	100%	100%	100%	100%
CONFIABILIDAD	100%	100%	100%	100%	100%	100%	100%	100%

Elaborado por: Investigador

La Tabla IV. LXVmuestra un resumen de los datos consolidados del análisis posterior de los datos después de haber utilizado la herramienta para calidad de datos.

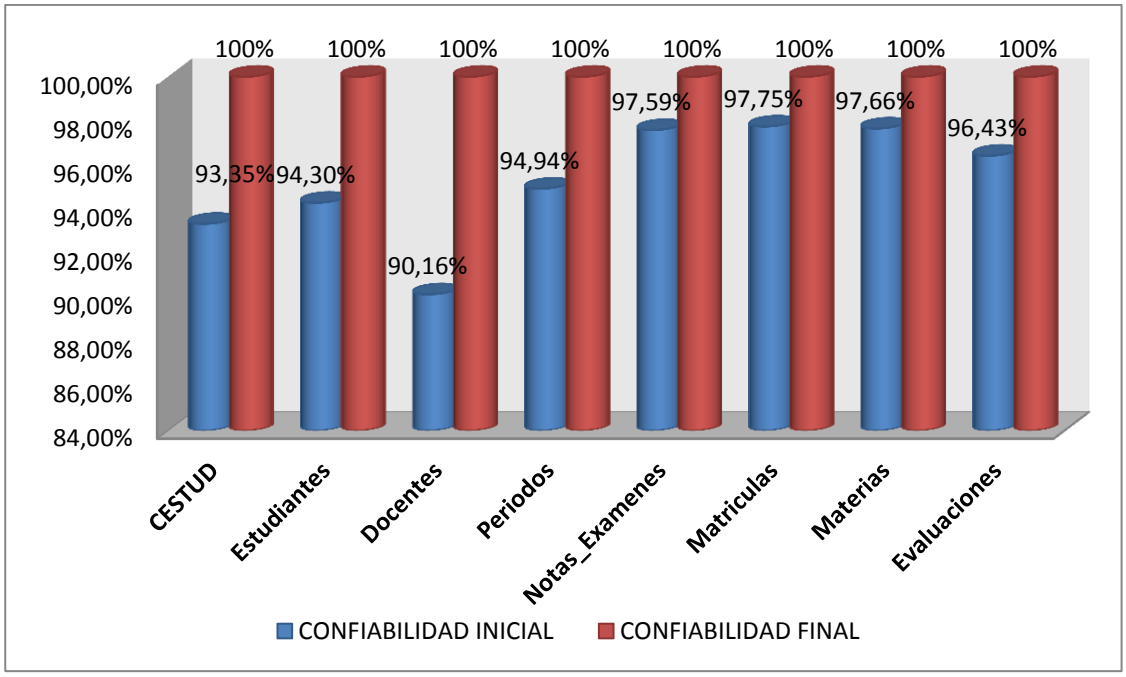


Figura IV. 23.Resultado Final de Confiabilidad.

Fuente: Investigador

4.8.3. RESULTADO FINAL DE LA CONFIABILIDAD DE LAS TABLAS

Como se puede apreciar en la gráfica anterior la dimensión de confiabilidad de los datos de cada tabla ha aumentado después del uso de las herramientas para calidad de datos obteniendo un 100% para cada una de las tablas dando como resultado un nivel perfecto de conformidad de esta dimensión dejando así los datos con un alto índice de calidad, corrigiendo los problemas de valores nulos, valores incorrectos y duplicados, encontrados en el análisis previo, de esta manera se garantiza la calidad de la información.

Consolidación de los resultados antes de utilizar las herramientas.

Tabla IV. LXVI. Resultados antes de utilizar las herramientas

CONSISTENCIA	CONFIABILIDAD
88,59%	95,27%

Elaborado por: Investigador

Consolidación de los resultados después de utilizar las herramientas.

Tabla IV. LXVII. Resultados después de utilizar las herramientas

CONSISTENCIA	CONFIABILIDAD
100%	100%

Elaborado por: Investigador

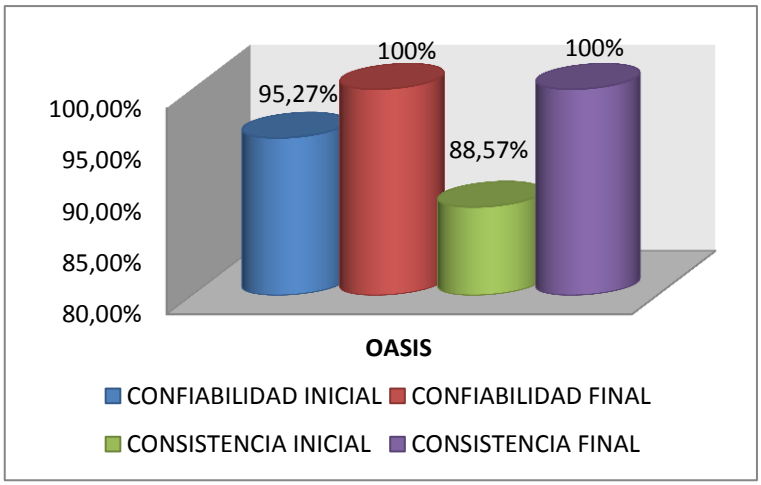


Figura IV. 24. Consolidación Final
Fuente: Investigador

Análisis de la gráfica de consolidación.

La confiabilidad y consistencia de los datos se mide en base a los indicadores descritos en la Tabla IV.46, en la gráfica anterior podemos observar que la confiabilidad inicial obtenida sin el uso de las herramientas ha sido de 95,27% a diferencia de la confiabilidad final donde se ha obtenido un 100% mejorando esta dimensión en un

4,73% dejando así los problemas de valores nulos, valores incorrectos y los duplicados solucionados, además de la consistencia inicial la cual se obtuvo un 88,57% en el análisis preliminar al contrario del análisis final donde se obtuvo 100% mejorando esta dimensión en 11,43% aumentando la aceptabilidad denlas dimensiones de precisión y valores válidos, consolidando los datos de mejor calidad.

La hipótesis planteada es:

H1: El uso de las herramientas para Data Quality permite mejorar la confiabilidad y consistencia de los datos en la base de datos “OASIS”.

H1: H1.1 ∧ H1.2

H1.1: Confiabilidad con el uso de las herramientas mayor que confiabilidad sin el uso de las herramientas.

μ1: Confiabilidad sin el uso de las herramientas.

μ2: Confiabilidad con el uso de las herramientas.

Datos

H1.1: (μ2 > μ1)

H1.1: (100%>95,27%

Por estadística descriptiva se acepta H1.1

H1.2: Consistencia con el uso de las herramientas mayor que consistencia sin el uso de las herramientas.

μ3: Consistencia sin el uso de las herramientas.

µ4: Consistencia con el uso de las herramientas.

Datos

H1.2: (µ4 > µ3)

H1.2: (100% > 88,57%)

Por estadística descriptiva se acepta H1.2

H1: (100% > 95,27%) ∧ (100% > 88,57%)

Dela observación de los resultados obtenidos en el estudio de este trabajo para las dimensiones de calidad: confiabilidad y consistencia y aplicando estadística descriptiva se acepta la hipótesis planteada: "El uso de las herramientas para Data Quality permite mejorar la confiabilidad y consistencia de los datos en la base de datos OASIS". Por lo tanto se concluye que la hipótesis H1 es verdadera.

CONCLUSIONES

1. El análisis comparativo realizado entre las herramientas para calidad de datos permite determinar las herramientas más aceptables que se adaptan de mejor manera para un proyecto de calidad de datos obteniendo los siguientes resultados para cada una de las herramientas, la herramienta Oracle Data Quality ha alcanzado el puntaje más alto con un porcentaje de 90,63% mientras que la herramienta SQL Power ha logrado un valor de 84,74%, además la herramienta Informatica ha conseguido un valor de 81,25% y al final con el porcentaje más bajo DataCleaner ha obtenido un valor de 71,88%.

2. El escenario de pruebas ayuda a determinar de las herramientas seleccionadas Oracle Data Quality, Informatica, SQL Power y DataCleaner cual asiste de mejor manera en la selección de una herramienta al momento de realizar un proyecto de calidad de datos.

3. El uso de las herramientas para calidad de datos permite garantizar la calidad de la información de la base de datos OASIS mejorando su confiabilidad en un 4,73% y su consistencia en un 11,43% consolidando la información con la mejora de sus datos y fortaleciendo las dimensiones de calidad analizadas.

4. Se obtuvo un porcentaje del 87,5% para las herramientas propietarias y para las herramientas libres el 80,5% dejando como resultado que las herramientas propietarias son más recomendables según el estudio realizado en esta investigación.

RECOMENDACIONES

1. Al momento de definir las dimensiones de calidad que se desean analizar se recomienda tener un alto conocimiento del negocio con el fin de obtener los mejores resultados para la empresa donde se realizará el proyecto de calidad de datos.

2. Al realizar un proyecto de calidad de datos se recomienda se tome en cuenta las herramientas que se piensan utilizar debido a que estas pueden ser un limitante al momento de la realización del proyecto por falta de conocimiento en la herramienta por parte del desarrollador.

3. Incluir en el pensum académico de la Escuela de Ingeniería en Sistemas de la ESPOCH el estudio de las herramientas para calidad de datos y las dimensiones de calidad, debido a la importancia de información de calidad de datos dentro de los almacenes de datos en cualquier aplicación que se esté diseñando o ya diseñada, debido a que se comprobó que el conocimiento de las mismas, ayudan a garantizar la calidad de la información en el ámbito empresarial.

RESUMEN

La investigación de herramientas para calidad de datos propietarias frente a herramientas de software libre disponibles en el mercado, aplicado en el sistema académico OASIS de la Escuela Superior Politécnica de Chimborazo.

En la investigación se utilizó método científico y técnicas estadísticas para determinar las principales herramientas para limpieza de datos mejoren la confiabilidad y consistencia de los datos, se utilizó las herramientas: DataCleaner, SQL Power, Oracle Data Integrator e Informatica Data Quality, todas las herramientas fueron adaptadas a los mismos ambientes de desarrollo y pruebas.

Se alcanzó los siguientes resultados mediante la comparación de parámetros como: Acceso a datos, limpieza de datos y configuración, permitiendo evaluar el desempeño de cada herramienta, los cuales fueron: Oracle Data Integrator 91,75% (Muy Bueno), SQL Power 86% (Muy Bueno), Informatica Data Quality 83,25% (Muy Bueno) y DataCleaner 75% (Bueno).

Del análisis realizado se tiene que: el uso de herramientas para calidad de datos permitió garantizar la calidad de información del Sistema Académico OASIS de la Escuela Superior Politécnica de Chimborazo mejorando su confiabilidad en un 4,73% y su consistencia en un 11,43% además de fortalecer las dimensiones de calidad analizadas.

Se recomienda al personal de desarrollo el uso de herramientas para calidad de datos en sus proyectos debido a que su utilidad ayuda y garantiza la optimización de calidad de los datos.

Palabras claves: análisis, calidad, consistencia, confiabilidad.

SUMMARY

This investigation is about owners' data quality tools in comparison with free-software tools available in the market applied in the OASIS academic system of Escuela Superior Politecnica de Chimborazo.

Scientific method and statistical techniques were used in this investigation so that the main data cleaning tools improving data reliability and consistency, can be determined. The following tools were used: Data Cleaner, SQL Power, Oracle Data Integrator and Informatics Data Quality. All of these tools were adapted to the same development and test environments.

Having compared parameters, the following results were gotten: Data Access, data cleaning and configuration which allowed evaluating the performance of each tool, that is, Oracle Data Integrator 91,75% (Very good), SQL Power 86% (Very good), Informatics Data Quality 83,25%(Very good) and DataCleaner 75%(Good).

By using tools for quality, it was possible to guarantee the OASIS academic system information quality of ESPOCH improving its reliability in 4,73% and its consistency in 11,43%. Moreover, the analyzed quality dimensions were improved.

It is recommended that the development staff use the data quality tool in their projects because it helps and guarantees the data quality optimization.

BIBLIOGRAFÍA

[1]. NORMA ISO / IEC 25012:2008.

E-book:

http://www.iso.org/iso/catalogue_detail.htm?csnumber=35736

2013-11-21.

[2]. CALIDAD DE DATOS.

E-book:

http://es.wikipedia.org/wiki/Calidad_de_datos

2013-11-01.

[3]. PÉREZ, C., En Los Datos, La calidad Importa. 2013

E-book:

http://liberix.es/blog/en-los-datos-la-calidad-importa/

[4]. DATACLEANER., Concepts., Belfast, U.K. 2011

E-book:

http://datacleaner.eobjects.org/resources/docs/2.5.2/pdf/

[5]. INFORMATICA., Gestión de datos maestro., Estados Unidos., 2008

E-book:

http://www.informatica.com/INFA_Resources/br_mdm_es.pdf

[6]. NORMA ISO 8000-110:2009.

E-book:

http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=51653

2013-11-10.

[7]. LIMPIEZA DE DATOS.

E-book:

http://es.wikipedia.org/wiki/Limpieza_de_datos

2013-11-01.

[8]. AUDITORIA DE DATOS.

E-book:

http://en.wikipedia.org/wiki/Data_auditing

2013-11-01.

[9]. DATA QUALITY.

E-book:

http://www.buenastareas.com/ensayos/Data-Quality/7365766.html

2013-11-01.

[10]. MAGIC QUADRANT FOR DATA QUALITY TOOLS. Belfast, U.K.

E-book:

http://www.citia.co.uk/content/files/50_161-377.pdf

2013-11-15.

[11]. PERVASIVE DATA QUALITY FOR TRUSTED DATA.

Belfast, U.K.

E-book:

http://www.informatica.com/INFA_Resources/ds_idq_6710.pdf

2013-11-15.

[12]. INFORAMTICA GETTING STARTED. Belfast, U.K. 2013

E-book:

http://www.informatica.com/INFA_Resources/ds_idq_6710.pdf

[13]. ORACLE ENTERPRISE DATA QUALITY PRODUCT FAMILY

Belfast, U.K.

E-book:

http://www.oracle.com/us/products/middleware/data-integration/enterprise-data-quality/oracle-enterprise-data-quality-ds-430148.pdf

2013-11-15.

[14]. SQL POWER. Belfast, U.K.

E-book:

http://www.sqlpower.ca/page/company

2013-11-15.

[15]. SQL POWER DQGURU. Belfast, U.K.

E-book:

http://www.sqlpower.ca/page/company

2013-11-15.

[16]. SQL POWER ARQUITECTURA. Belfast, U.K.

E-book:

http://www.sqlpower.ca/page/company

2013-11-15.

[17]. SOLIS, I., Propuesta metodológica para la gestión de la calidad de datos en proyectos de integración**.,** Facultad Informática y Electrónica – Escuela Ingeniería en Sistemas., Escuela Superior Politécnica de Chimborazo., Riobamba – Ecuador., Tesis., 2011., Pp. 64-91, 107-135.

[18]. LOSHIN, D., The practitioner's guide to data quality improvement**.**

2014-01-15.

http://common.books24x7.com/toc.aspx?bookid=40139.

[15]. CONFIALBILIDAD Belfast, U.K.

E-book:

http://clubensayos.com/Psicolog%C3%ADa/Confiabilidad/1454932.html

2013-11-15.

ANEXO I

ESCENARIO DE PRUEBAS

ESCENARIO DE PRUEBA.

Se tomó una base de datos en SQL Server 2008 que consta de dos tablas con una cantidad aproximada de cien mil registros en cada tabla el nombre de la tablas son TPM_CUS y TPM_TRAN, la tabla TPM_CUS contiene nueve columnas: CODIGO, CODIGO_CLIENTE, NOMBRES, IDENTIFICACION, TIPO_IDENTIFICACION, DIRECCION, EMAIL, TELEFONO, CELULAR y en la tabla TPM_TRAN contiene tres columnas CODIGO, CODIGO_CLIENTE, INL_AMT, en las cuales podemos encontrar varios problemas como valores incoherentes, caracteres alfabéticos en columnas donde no deben existir letras, espacios en blanco entre palabras, etc. y además registros duplicados.

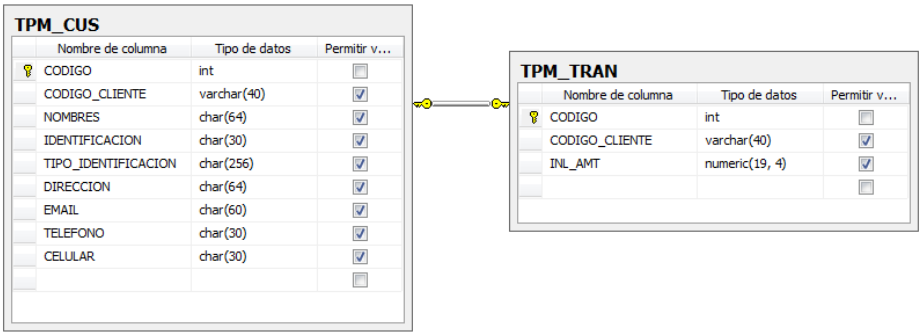


Figura 1. Estructuras de las tablas
Fuente: Investigador

Requerimientos de calidad de datos

Tabla 1. Requerimientos de calidad de datos
Fuente: Investigador

N°	Problemas	Tabla(s)	Columna(s)	Requerimientos	Acción	Herramientas
1	Datos NULL y blancos	TPM_CUS TPM_TRAN	DIRRECION EMAIL TELÉFONO CELULAR INL_AMT	Medir la cantidad de valores nulos o blancos.	Sustituir por valores referenciales.	-DataCleaner -SQL Power -Data Integrator -Informatica
2	Datos incompletos	TPM_CUS	CEDULA	Medir la cantidad de registros incompletos.	Valores referenciales en caso de encontrar cedulas sin la longitud exacta. Ejemplo 9999999999.	
3	Datos duplicados	TPM_CUS	CÓDIGO_CLIENTE	Medir la cantidad de registros duplicados.	Eliminar los registros duplicados conservando el mayor número de información.	
4	Inconsistencia en los datos	TPM_CUS	IDENTIFICACIÓN TELÉFONO	Medir el nivel de inconsistencias en los datos	reducir las inconsistencias en el nivel que sea necesario	
5	Datos sin estándar.	TPM_CUS	NOMBRES DIRECCIÓN EMAIL	Medir la cantidad de registros se encuentra con minúsculas y mayúsculas.	Estándar definido: Nombres y dirección con mayúsculas. Email con minúsculas.	

PRUEBAS DE LAS HERRAMIENTAS

Las herramientas fueron instaladas en un mismo computador de las siguientes características:

Procesador: Intel Core i5 de 2.4
Memoria: 8 Gb de RAM
Sistema: Windows Ultimate de 64 bits

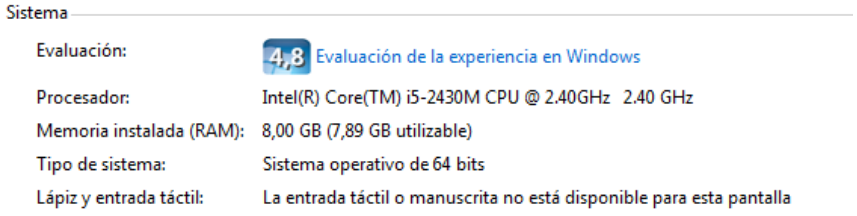


Figura 2. Detalle del Equipo
Fuente: Investigador

DataCleaner

Instalación

La herramienta DataCleaner no necesita una instalación previa debido a que es una aplicación ejecutable, solo se debe ejecutar la aplicación después de descomprimirla cuando la descargamos del sitio oficial de la herramienta.

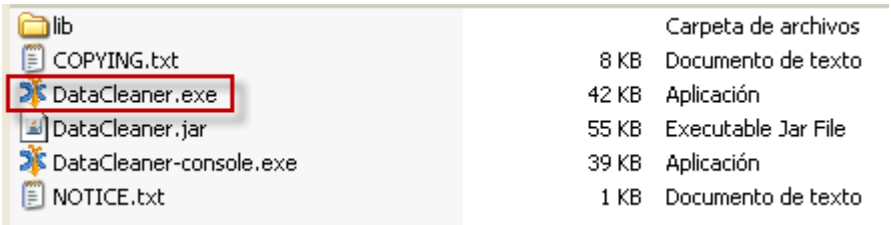


Figura 3. Estructuras de las tablas
Fuente: Investigador

Nota: El único requisito para que la herramienta pueda ejecutarse sin ningún problema es que en el equipo se encuentre pre instalado la versión JRE 6.0 de java o una superior.

Conexión a los orígenes de datos.

Para realizar una conexión a un almacén de datos se selecciona de la pantalla principal los accesos rápidos que ahí se muestra, si se desea realizar una conexión a otro origen se debe seleccionar el botón **MORE** de la pantalla principal.

Para conocer los **DRIVERS** que están instalados o que se pueden instalar cuando se los necesite se debe configurar las opciones desde el menú principal -> **Windows** la opción **Options.**

Los drivers que se encuentran listos para ser usados se muestran con un icono verde y los otros con un icono plomo para ser descargados de manera directa si es necesario su uso.

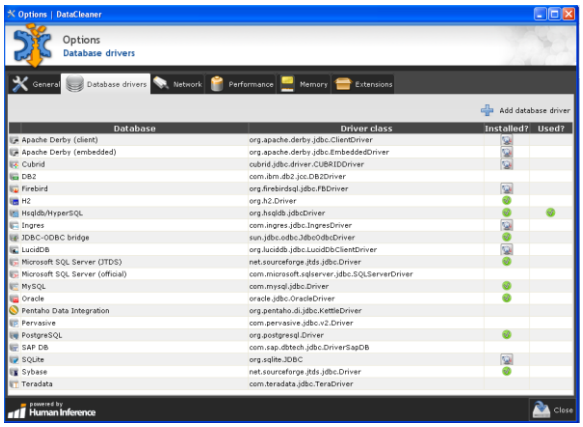


Figura 4. Database drivers
Fuente: Investigador

Conexión.
Conectarse a la base de datos.
Para conectarnos a la base de datos DB_PRUEBAS_DQ se utiliza el driver que se tiene por defecto en la herramienta la cual presenta la siguiente pantalla para la conexión con el DBMS.

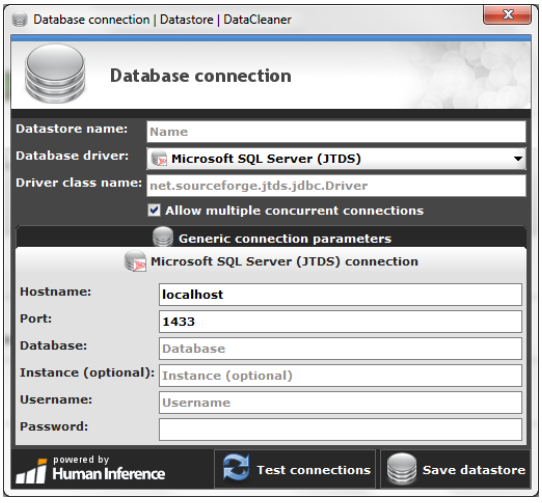


Figura 5. Conexión de base de datos
Fuente: Investigador

Donde se debe configurar:
1.- El nombre de la conexión.
2.- Revisar que el driver sea propio de MICROSOFT o el recomendado por la herramienta.
3.- El nombre del servidor al cual se va a conectar.
4.- El puerto por el cual se comunica.
5.- El nombre de la base a la cual se va a conectar.
6.- El nombre de la instancia.
7.- El usuario y la clave del servidor con permisos de acceso a la base de datos.

Se prueba la conexión de la herramienta con el DBMS de Microsoft.

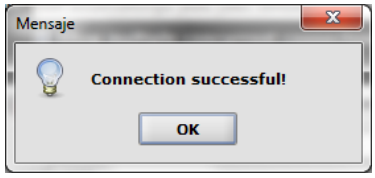


Figura 6. Mensaje
Fuente: Investigador

Se procede con el análisis de los datos, una vez conectados con el servidor de datos se puede visualizar las columnas que comprenden en cada tabla si la selecciona.

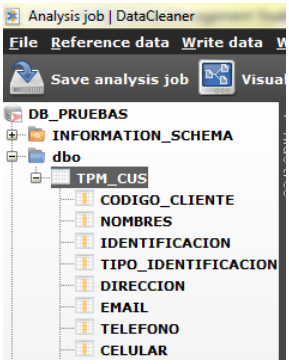


Figura 7. Tablas de análisis
Fuente: Investigador

Para un rápido trabajo con los datos se puede dar un clic secundario en la tabla y seleccionar una de las opciones que ahí aparecen tal como:

Remover la tabla de origen.
Análisis rápido
Exportar a tablas de Excel

Exportar como un archivo CVS
Ver los datos.

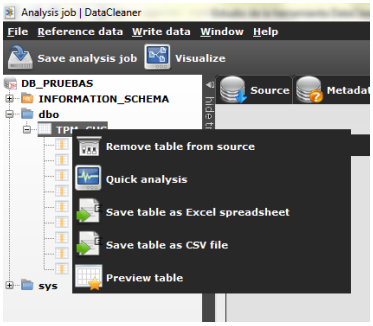


Figura 8. Opciones
Fuente: Investigador

Se realiza un análisis de los datos de la tabla TMP_CUS

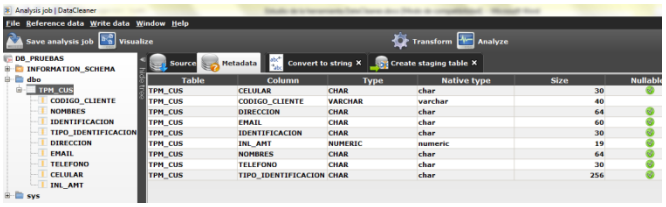


Figura 9. Tabla TMP_CUS
Fuente: Investigador

Se realiza una conversión de los datos de tipo CHAR a tipo STRING para realizar un análisis de estos. Selecciona las columnas que se van a convertir.

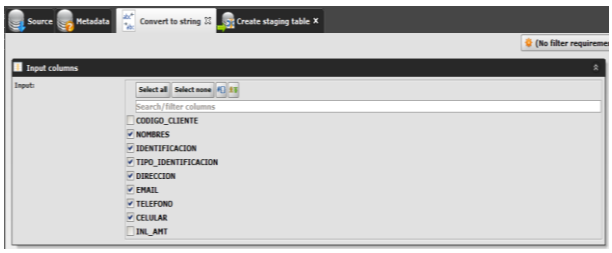


Figura. 10 Columnas de entrada
Fuente: Investigador

El nombre de las columnas de salida.

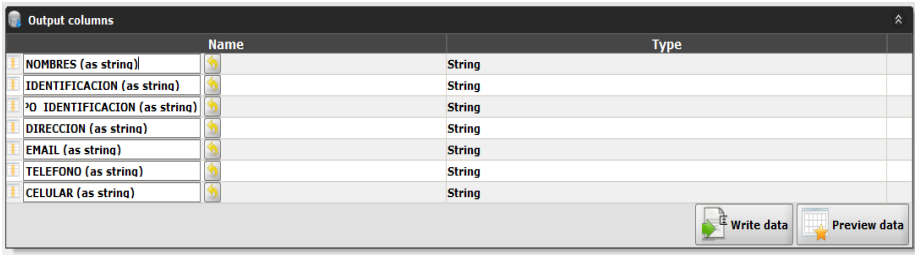


Figura. 11 Columnas de salida
Fuente: Investigador

La grafica del trabajo que se está realizando.

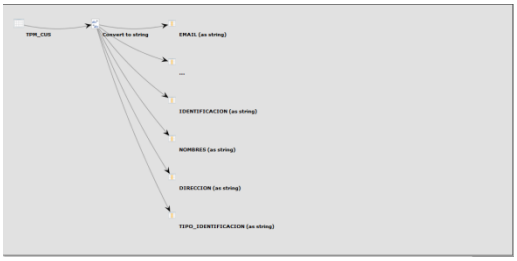


Figura 12 Diagrama de ejecución
Fuente: Investigador

El resultado lo podemos ver de las siguientes maneras:

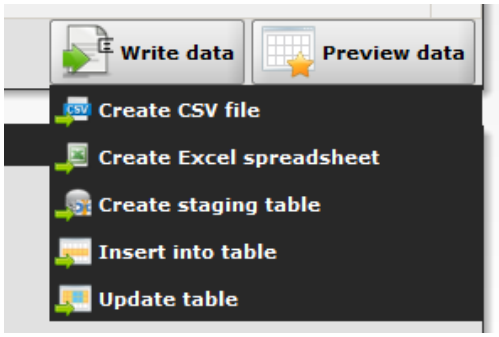


Figura 13. Formatos
Fuente: Investigador

Crear un archivo CSV
Crear un archivo de Excel
Crear una tabla temporal
Insertar en una tabla
Actualizar la tabla
Se puede visualizar los datos convertidos

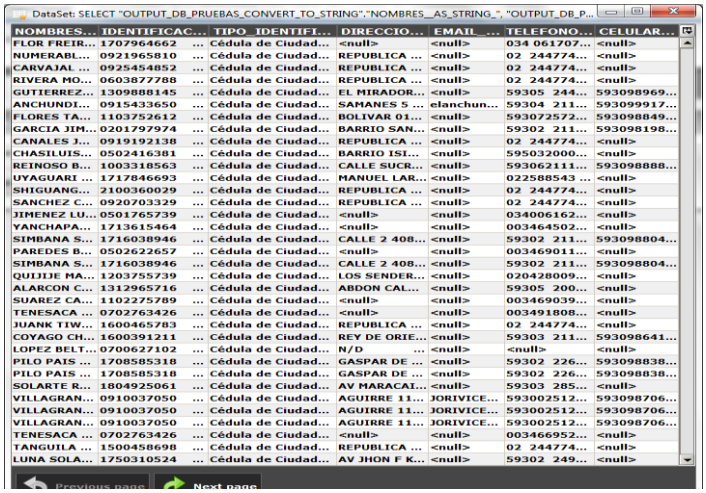


Figura 14. Resultados
Fuente: Investigador

Se pueden visualizar las tablas temporales creadas cuando se desee

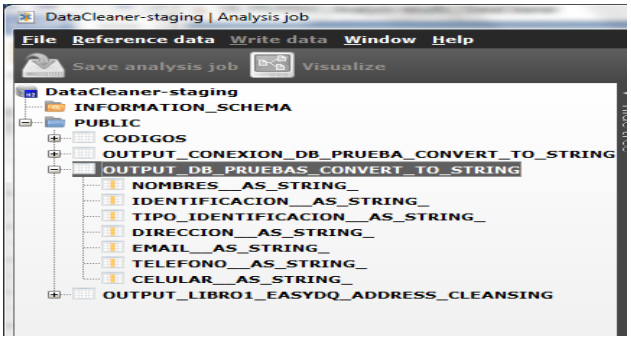


Figura 15. Tablas Temporales
Fuente: Investigador

Seleccionamos "analize this datastorge" para pasar a la ventana de análisis del repositorio temporal.

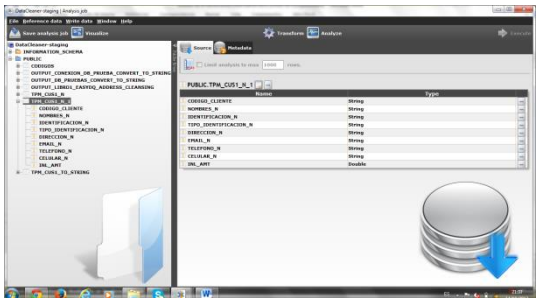


Figura 16. Análisis del repositorio temporal
Fuente: Investigador

En esta parte se selecciona la tabla creada y se procede a seleccionar todas sus columnas y con un clic contextual se despliega un menú donde se selecciona la opción “Quick Analysis” para conocer los datos

Para conocer los metadatos de los campos se selecciona la pestaña METADATA del origen de los datos.

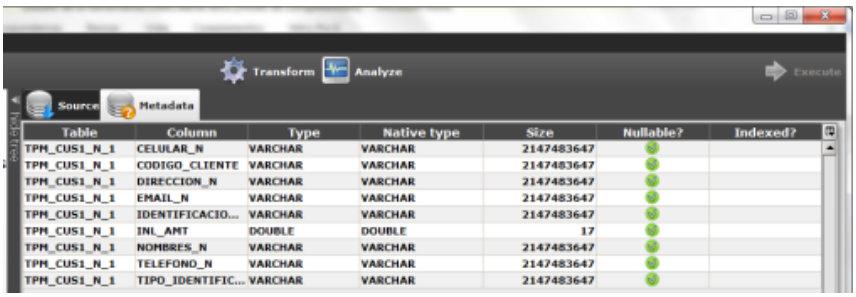


Table	Column	Type	Native type	Size	Nullable?	Indexed?
TPM_CUS1_N_1	CELULAR_N	VARCHAR	VARCHAR	2147483647		
TPM_CUS1_N_1	CODIGO_CLIENTE	VARCHAR	VARCHAR	2147483647		
TPM_CUS1_N_1	DIRECCION_N	VARCHAR	VARCHAR	2147483647		
TPM_CUS1_N_1	EMAIL_N	VARCHAR	VARCHAR	2147483647		
TPM_CUS1_N_1	IDENTIFICACIO...	VARCHAR	VARCHAR	2147483647		
TPM_CUS1_N_1	INL_AMT	DOUBLE	DOUBLE	17		
TPM_CUS1_N_1	NOMBRES_N	VARCHAR	VARCHAR	2147483647		
TPM_CUS1_N_1	TELEFONO_N	VARCHAR	VARCHAR	2147483647		
TPM_CUS1_N_1	TIPO_IDENTIFIC...	VARCHAR	VARCHAR	2147483647		

Figura 17. Metadata
Fuente: Investigador

Al convertir los datos de CHAR a STRING se modifican las variables lo cual no es muy recomendable y eso no se puede cambiar ya que no deja elegir la longitud de los datos y cambia los valores originales.

Perfiles

Después de ejecutar el análisis rápido se visualiza una tabla de resultados muy buena con un resumen de cómo se encuentra los datos en la tabla temporal.

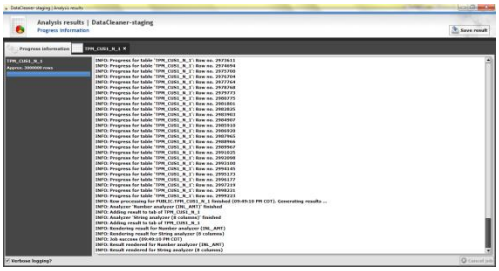


Figura 18. Perfiles
Fuente: Investigador

En tabla de resúmenes se observa que no existen muy buenos resultados los cuales se pueden guardar como resultados iniciales de análisis.

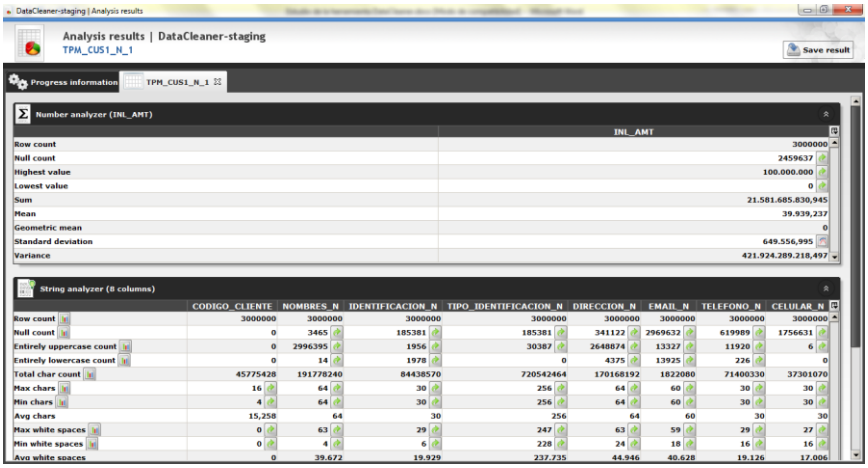


Figura 19. Tabla Resumen
Fuente: Investigador

Se puede visualizar los resultados de análisis de números y STRINGS tales como valores números de registros analizados, valor máximo, mínimo, promedios, suma, desviación estándar, varianza, etc. Para los valores numéricos y para los valores STRING número de registros analizados, valores nulos, máximo de palabras, mínimo de palabras, total de caracteres, espacios en blanco, numero de caracteres utilizados.

Resultados obtenidos del análisis.

Los resultados obtenidos de la herramienta se muestran a continuación:

Tabla: TPM_CUS

Los valores nulos.

Tabla 2. Resultados
Fuente: Investigador

TPM_CUS	
COLUMNA	**VALORES NULOS**
CODIGO	0
CODIGO_CLIENTE	0
NOMBRES	0
IDENTIFICACION	0
TIPO_IDENTIFICACION	0
DIRECCION	2652
EMAIL	95118
TELEFONO	382

CELULAR	39794

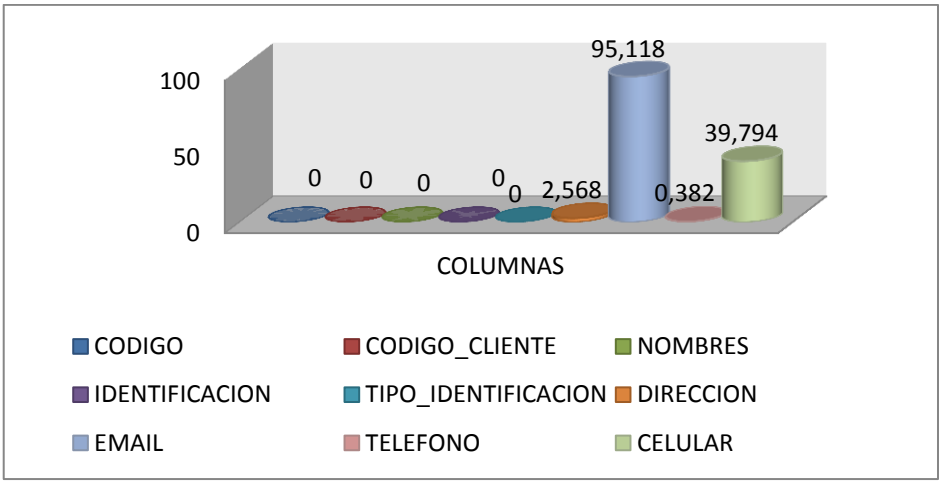


Figura 20. Resultados
Fuente: Investigador

Interpretación de resultados para valores nulos.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan valores nulos a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Caracteres Inválidos.

Tabla 3 . Caracteres Inválidos
Fuente: Investigador

TPM_CUS	
COLUMNA	**CARACTERES INVÁLIDOS**
CODIGO	0
CODIGO_CLIENTE	0
NOMBRES	0
IDENTIFICACION	73
TIPO_IDENTIFICACION	0
DIRECCION	0
EMAIL	0
TELEFONO	2039
CELULAR	0

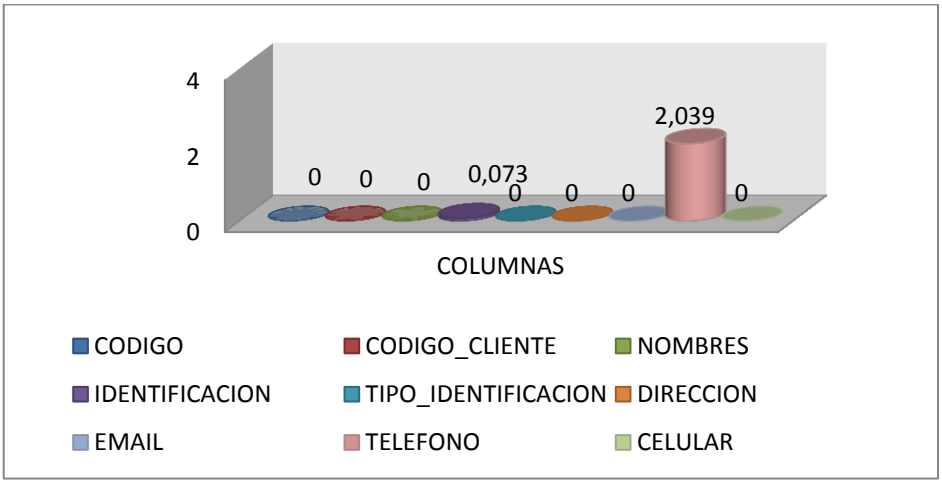


Figura 21 Tabla 4. Resultados
Fuente: Investigador

Interpretación de resultados para caracteres inválidos.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan caracteres inválidos a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Registros duplicados.

Tabla 5 . Registros Duplicados
Fuente: Investigador

TPM_CUS	
COLUMNA	**DUCPLICADOS**
CODIGO	0
CODIGO_CLIENTE	18015
NOMBRES	18015
IDENTIFICACION	18015
TIPO_IDENTIFICACION	18015
DIRECCION	9250
EMAIL	5032
TELEFONO	1355
CELULAR	576

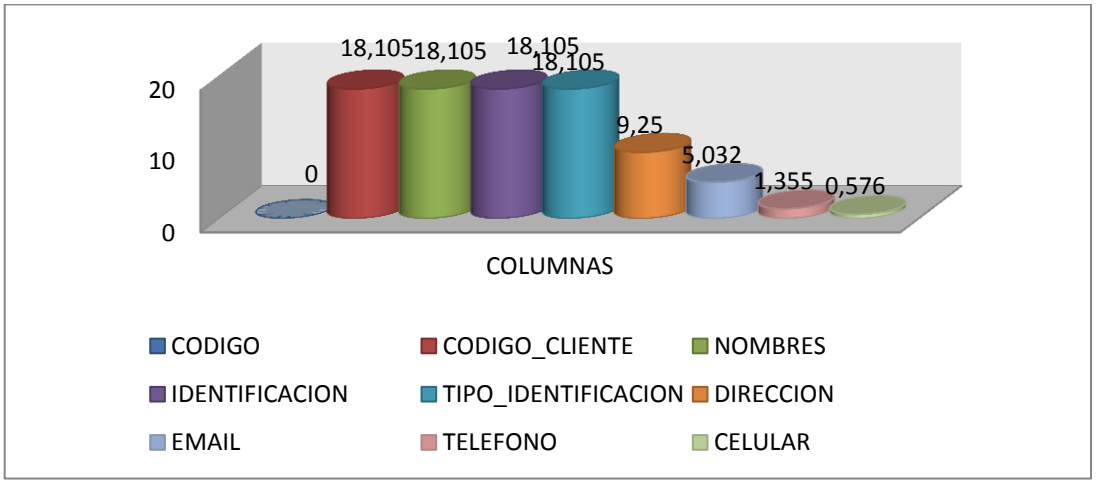


Figura 22. Resultados
Fuente: Investigador

Interpretación de resultados para duplicados.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan registros duplicados a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Total de registros duplicados.

Tabla 6 . Resultados
Fuente: Investigador

TOTAL REGISTRO DUPLICADOS	
REGISTROS	46492

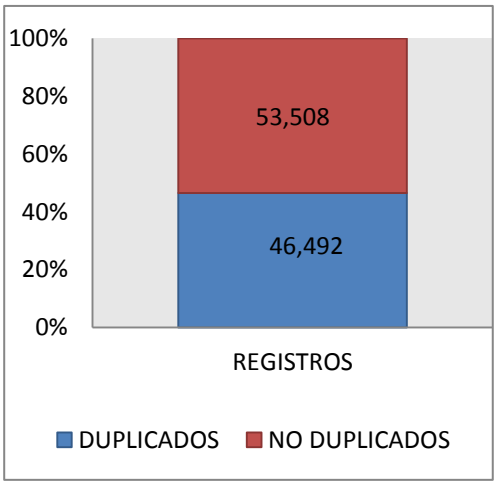


Figura 23. Resultados
Fuente: Investigador

Interpretación de resultados para total de duplicados.

Como se visualizar en la figura anterior existe un gran número de registros duplicados para la tabla TPM_CUS los cuales deben ser solventados para obtener una mejor calidad de datos en esta tabla.

Tabla: TPM_TRAN

Valores Nulos.

Tabla 7. **Resultados**
Fuente: Investigador

TPM_TRAN	
COLUMNA	**VALORES NULOS**
CODIGO	0
CODIGO_CLIENTE	0
INL_AMT	81503

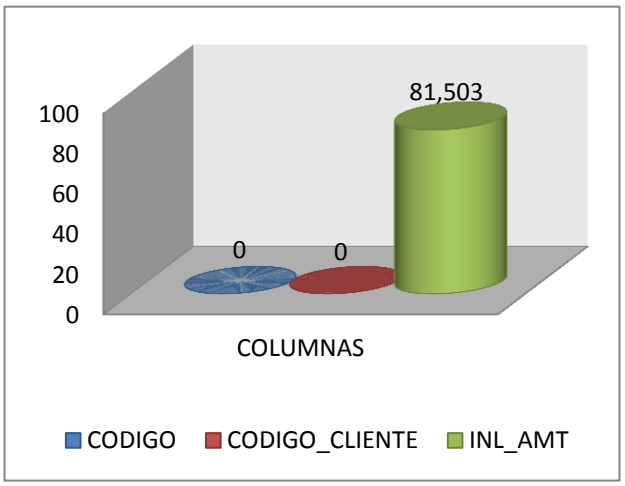


Figura 24 . Resultados
Fuente: Investigador

Interpretación de resultados para valores nulos.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan valores nulos a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Caracteres Inválidos.

No se presentaron valores invalidos en esta tabla.
Interpretación de resultados para caracteres inválidos.

Al realizar el análisis de los datos para esta tabla se evidencio que no existían valores inválidos dentro de los campos de la tabla TPM_TRAN.

Registros duplicados.

Tabla 8. Resultados
Fuente: Investigador

TPM_TRAN	
COLUMNA	**DUCPLICADOS**
CODIGO	0
CODIGO_CLIENTE	8899
INL_AMT	2024

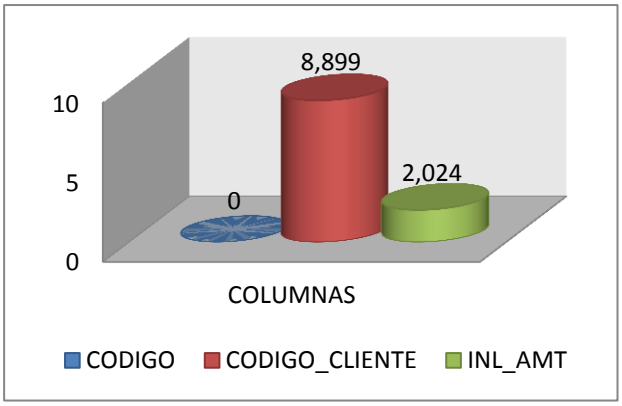


Figura 25. Resultados
Fuente: Investigador

Interpretación de resultados para duplicados.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan registros duplicados a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Total de registros duplicados.

Tabla 9. Resultados
Fuente: Investigador

TOTAL REGISTRO DUPLICADOS	
REGISTROS	72775

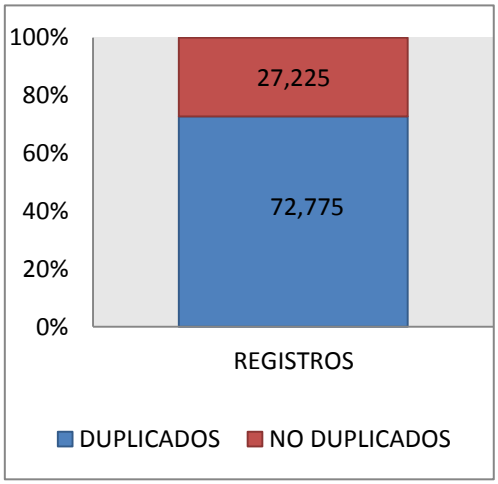


Figura 26 . Resultados
Fuente: Investigador

Interpretación de resultados para total de duplicados.

Como se visualizar en la figura anterior existe un gran número de registros duplicados para la tabla TPM_TRAN los cuales deben ser solventados para obtener una mejor calidad de datos en esta tabla.

Limpieza de los datos.

En la limpieza de los datos esta herramienta presenta una dificultad debido a que los servicios de limpieza son bajo demanda es decir tienen precio por cada dato. Por esta razón este parte de la limpieza de los datos no se tomará en cuenta con esta herramienta.

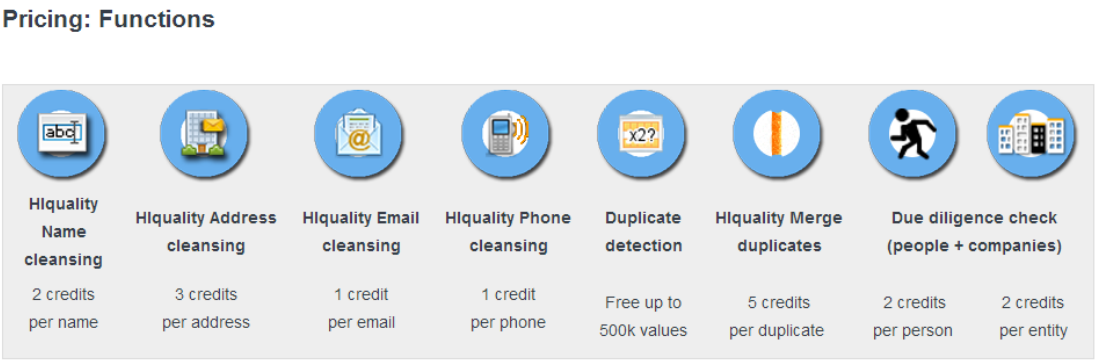


Figura 27. Limpieza de Datos
Fuente: Investigador

SQL Power

Instalación

La instalación solo requiere de ejecutar el archivo .exe después de la descarga y descompresión del archivo esto creará un acceso directo en el menú de inicio.

Para inicial con la herramienta se procede a ejecutarla desde el menú Inicio una vez instalada.

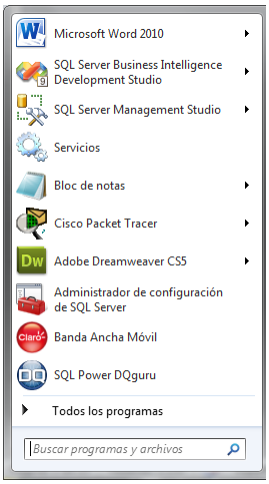


Figura 28. Menú Inicio
Fuente: Investigador

Conexión a los orígenes de datos.

Conexión.

Se procede con la configuración de las conexiones a la base de datos, para ello se configura desde el menú principal de la herramienta

Figura 29. Conexión
Fuente: Investigador

Se selecciona la opción "Connections" del submenú y selecciona la opción administración de conexiones a base de datos.

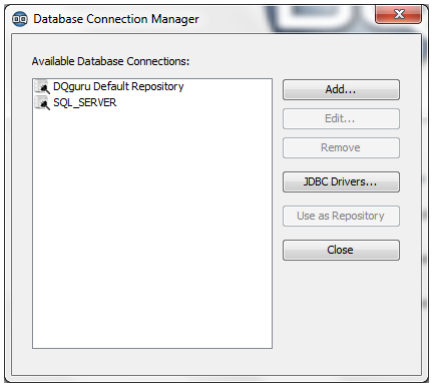


Figura 30. Administración de conexiones
Fuente: Investigador

Selecciona el botón añadir “Add.” para configurar una nueva conexión al SQL

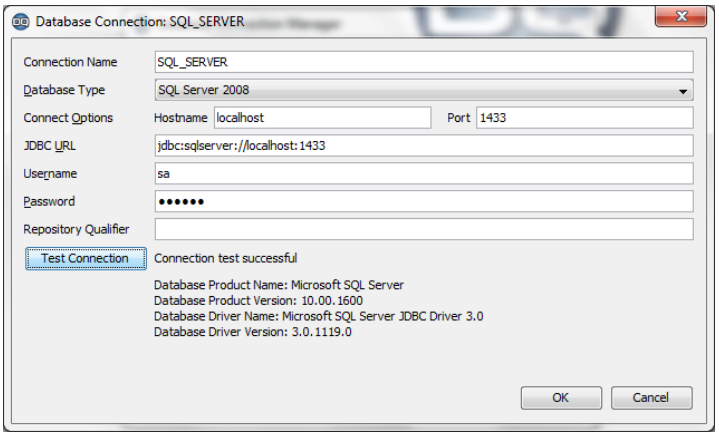


Figura 31. Configuración de nueva conexión
Fuente: Investigador

Llena los datos con la configuración necesaria para crear una conexión al SQL_SERVER con el fin de luego proceder con los trabajos que se desee.

PREFILADO

Para realizar un análisis previo a la corrección de los datos la herramienta SQL Power DQguru no cuenta con opciones de análisis de la información o como se la conoce el Data Profiling. Por esta razón se tomara los resultados obtenidos con la herramienta DataCleaner que fue utilizada con anterioridad.

Los resultados obtenidos de la herramienta DataCleaner se muestran a continuación:

Tabla: TPM_CUS

Los valores nulos.

Tabla 10. Resultados
Fuente: Investigador

TPM_CUS	
COLUMNA	**VALORES NULOS**
CODIGO	0
CODIGO_CLIENTE	0
NOMBRES	0
IDENTIFICACION	0
TIPO_IDENTIFICACION	0

DIRECCION	2652
EMAIL	95118
TELEFONO	382
CELULAR	39794

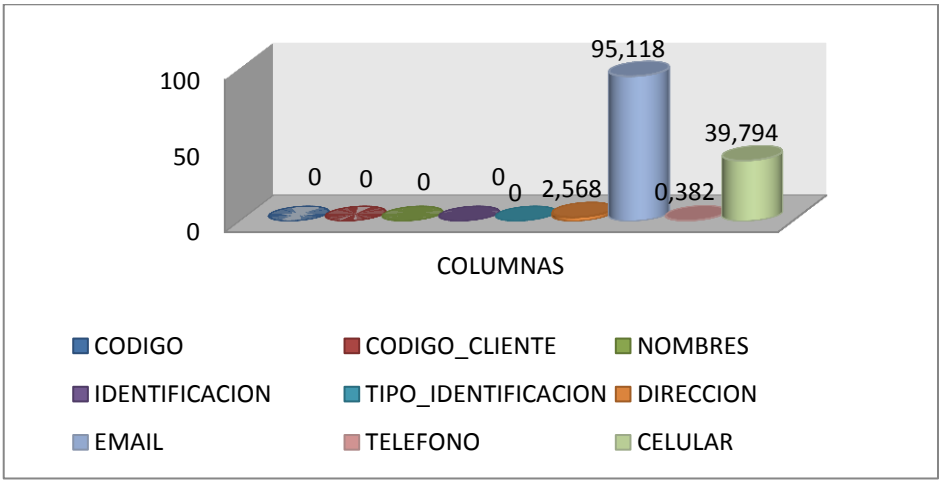


Figura 32. Resultados
Fuente: Investigador

Interpretación de resultados para valores nulos.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan valores nulos a los cuales se les dará su tratamiento de limpieza en instancias posteriores

Caracteres Inválidos.

Tabla 11 . Resultados
Fuente: Investigador

TPM_CUS	
COLUMNA	**CARACTERES INVÁLIDOS**
CODIGO	0
CODIGO_CLIENTE	0
NOMBRES	0
IDENTIFICACION	73
TIPO_IDENTIFICACION	0
DIRECCION	0
EMAIL	0
TELEFONO	2039
CELULAR	0

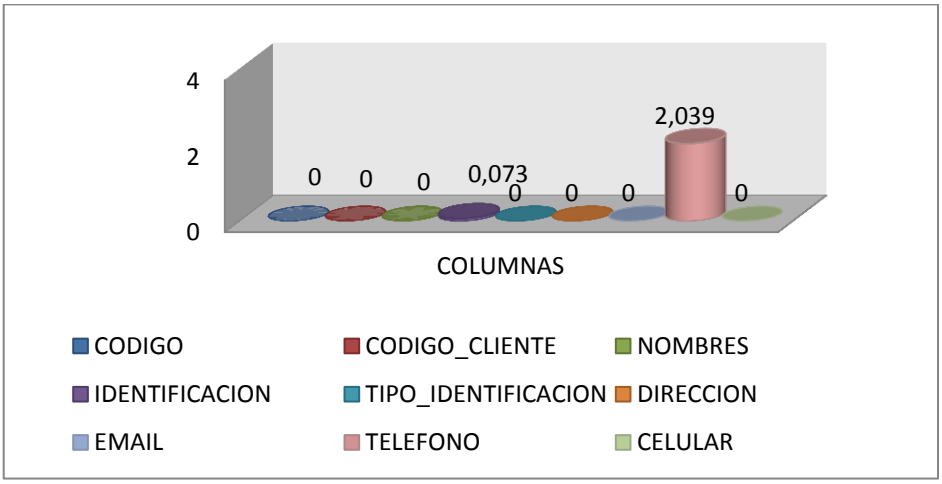


Figura 33. Resultados
Fuente: Investigador

Interpretación de resultados para caracteres inválidos.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan caracteres inválidos a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Registros duplicados.

Tabla 12. Resultados
Fuente: Investigador

TPM_CUS	
COLUMNA	**DUCPLICADOS**
CODIGO	0
CODIGO_CLIENTE	18015
NOMBRES	18015
IDENTIFICACION	18015
TIPO_IDENTIFICACION	18015
DIRECCION	9250
EMAIL	5032
TELEFONO	1355
CELULAR	576

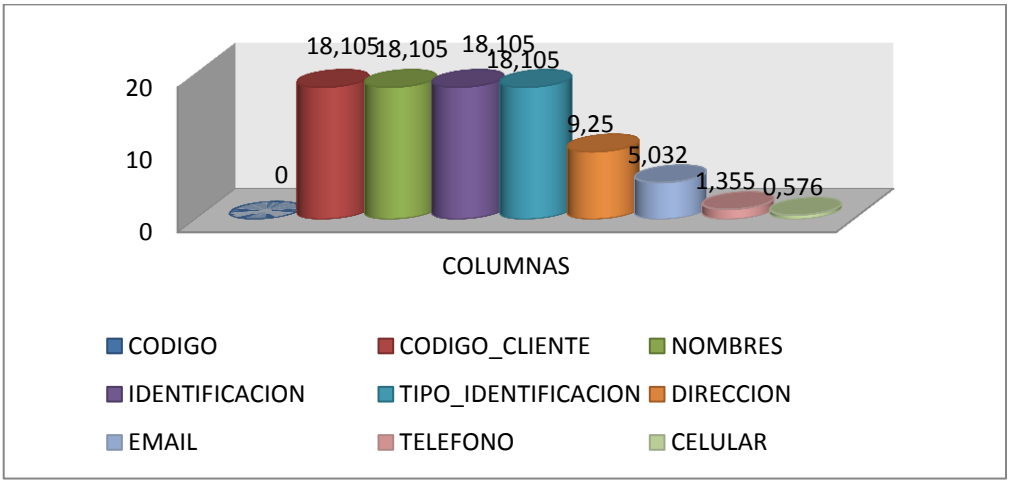


Figura 34. Resultados
Fuente: Investigador

Interpretación de resultados para duplicados.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan registros duplicados a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Total de registros duplicados.

Tabla 13. Resultados
Fuente: Investigador

TOTAL REGISTRO DUPLICADOS	
REGISTROS	46492

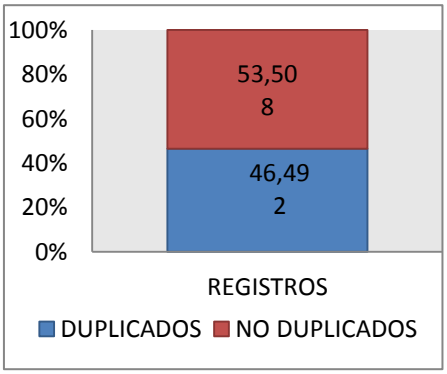


Figura 35 . Resultados
Fuente: Investigador

Interpretación de resultados para total de duplicados.

Como se visualizar en la figura anterior existe un gran número de registros duplicados para la tabla TPM_CUS los cuales deben ser solventados para obtener una mejor calidad de datos en esta tabla.

Tabla: TPM_TRAN

Valores Nulos.

Tabla 14. Resultados
Fuente: Investigador

TPM_TRAN	
COLUMNA	**VALORES NULOS**
CODIGO	0
CODIGO_CLIENTE	0
INL_AMT	81503

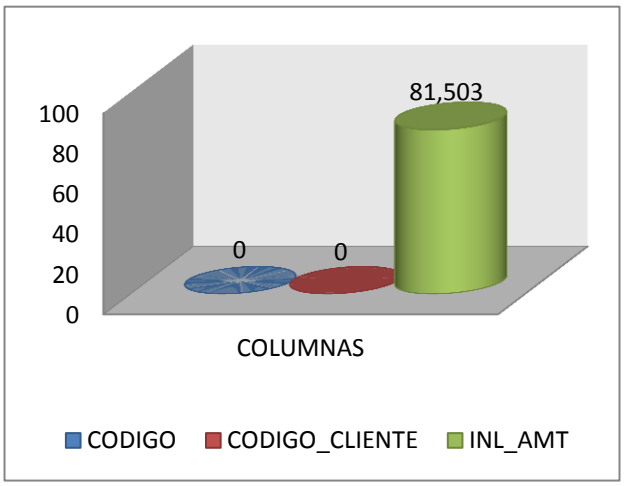


Figura 36. Resultados
Fuente: Investigador

Interpretación de resultados para valores nulos.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan valores nulos a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Caracteres Inválidos.

No se presentaron valores invalidos en esta tabla.
Interpretación de resultados para caracteres inválidos.

Al realizar el análisis de los datos para esta tabla se evidencio que no existían valores inválidos dentro de los campos de la tabla TPM_TRAN.

Registros duplicados.

Tabla 15. Resultados
Fuente: Investigador

TPM_TRAN	
COLUMNA	**DUCPLICADOS**
CODIGO	0
CODIGO_CLIENTE	8899
INL_AMT	2024

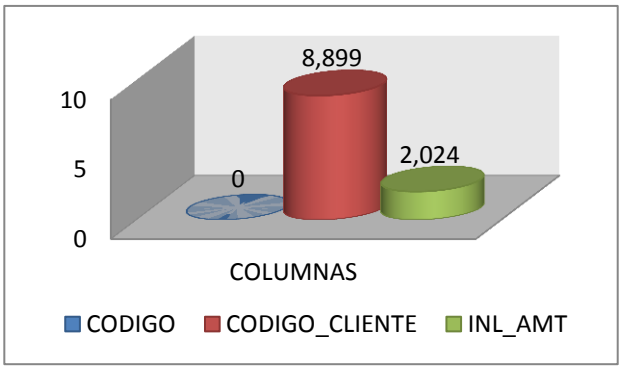


Figura 37 . Resultados
Fuente: Investigador

Interpretación de resultados para duplicados.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan registros duplicados a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Total de registros duplicados.

Tabla 16 Resultados
Fuente: Investigador

TOTAL REGISTRO DUPLICADOS	
REGISTROS	72775

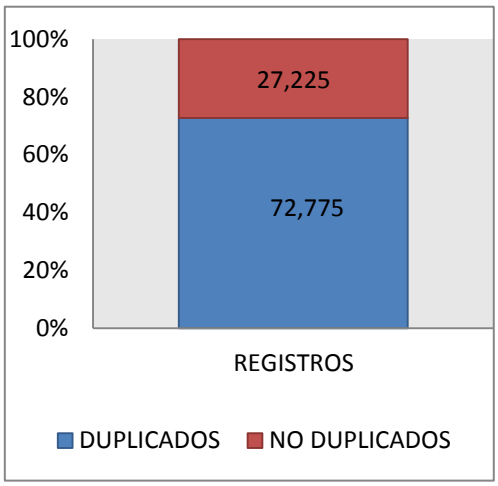


Figura 38. Resultados
Fuente: Investigador

Interpretación de resultados para total de duplicados.

Como se visualizar en la figura anterior existe un gran número de registros duplicados para la tabla TPM_TRAN los cuales deben ser solventados para obtener una mejor calidad de datos en esta tabla.

LIMPIEZA

Los pasos que se siguen para la limpieza con esta herramienta son los siguientes:

5. Crear un nueva carpeta de proyectos
6. Dentro de la carpeta creada, establecer un proyecto de limpieza “Cleasing Project”
7. Seleccionar la carpeta de transformaciones y crear una nueva transformación
8. Utilice los objetos que la herramienta nos provee para la limpieza

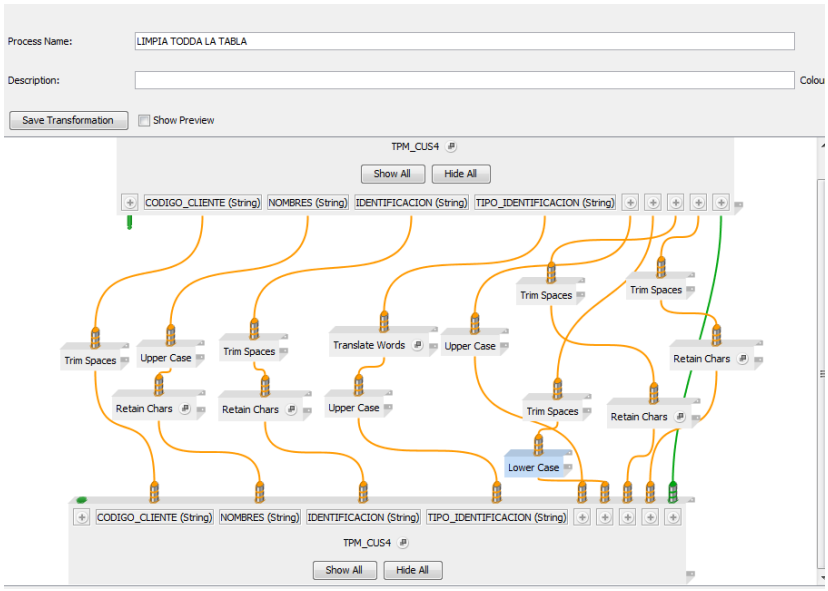


Figura 39. Escenario
Fuente: Investigador

Los objetos que se utilizan para la limpieza son básicamente:

Trim Spaces: Para eliminar espacios entre las palabras.

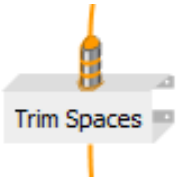


Figura 40. Trim Spaces
Fuente: Investigador

Este objeto no requiere de ninguna configuración.

Upper Case: para convertir las palabras en mayúsculas.

Figura 41. Upper Case
Fuente: Investigador

Este objeto no requiere de ninguna configuración.

Lower Case: para convertir las palabras en minúsculas.

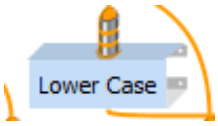


Figura 42. Lower Case
Fuente: Investigador

Este objeto no requiere de ninguna configuración.

Retain Chars: para retener solo ciertos caracteres deseados.

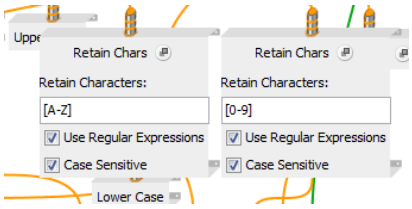


Figura 43. Retain Chars
Fuente: Investigador

Se debe utilizar la expresión de caracteres que deseemos conservar y seleccionar los check box para activar la opción de usar la expresión y uso de la sensibilidad para mayúsculas y minúsculas.

Traslate Words: para cambiar ciertos caracteres por otros.

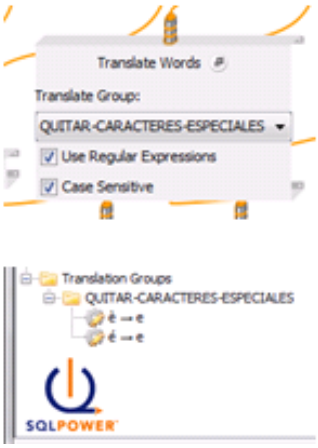


Figura 44. Traslate Words
Fuente: Investigador

Se debe utilizar la expresión de caracteres que desee cambiar y seleccionar los check box para activar la opción de usar la expresión y uso de la sensibilidad para mayúsculas y minúsculas. Los grupos de remplazo se deben personalizar con anterioridad lo cual es muy sencillo.

Quitar Duplicados

Para la limpieza de los datos se deben configurar los siguientes pasos:

Crear un nuevo proyecto de "De-duping project" y configurarlo.

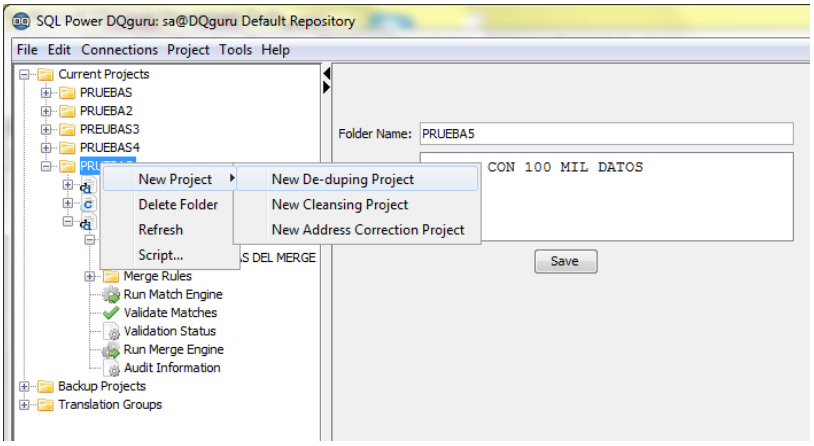


Figura 45. Configuración
Fuente: Investigador

La configura con los siguientes datos:

1. Nombre del proyecto.
2. Ubicación a la carpeta que deseamos que se añada el proyecto.
3. Nuestra conexión al SQL SERVER 2008 antes creada.
4. Selecciona la base de datos.
5. El esquema que se utilizara para la limpieza.
6. La tabla en la cual se aplicaran los cambios.
7. Una columna principal (en el caso de no existir una columna principal se puede escoger la posible PK)
8. Se debe elegir una base de datos donde se almacenara la información de las filas con registros duplicados.
9. De igual manera el esquema que se utilizara.
10. Se debe escribir el nombre de la tabla que almacenara la información de los registros duplicados.
11. Para el punto anterior no es necesario que la tabla exista ya que la herramienta crea la tabla con los campos necesarios de donde tomara la información para los cambios que se vayan a realizar.

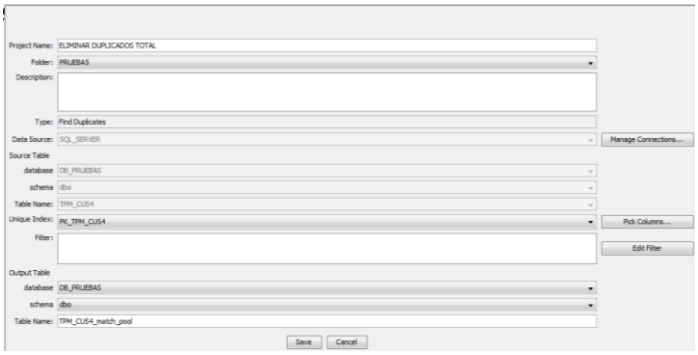


Figura 46. Accesos a la tabla
Fuente: Investigador

Una vez configurado las conexiones, crear las columnas de salida que serán analizadas para obtener los resultados de posibles duplicados.

En este caso se debe crear una transformación de salida en la cual se selecciona una o varias columnas de análisis.

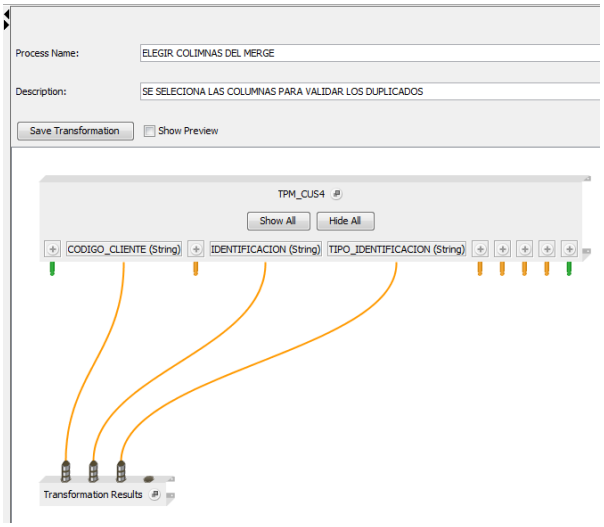


Figura 47. Configuración de Duplicado
Fuente: Investigador

Añadir la tabla TPM_TRAN en el "Merge Rules" y configurar la acción que desea que realice el Merge.

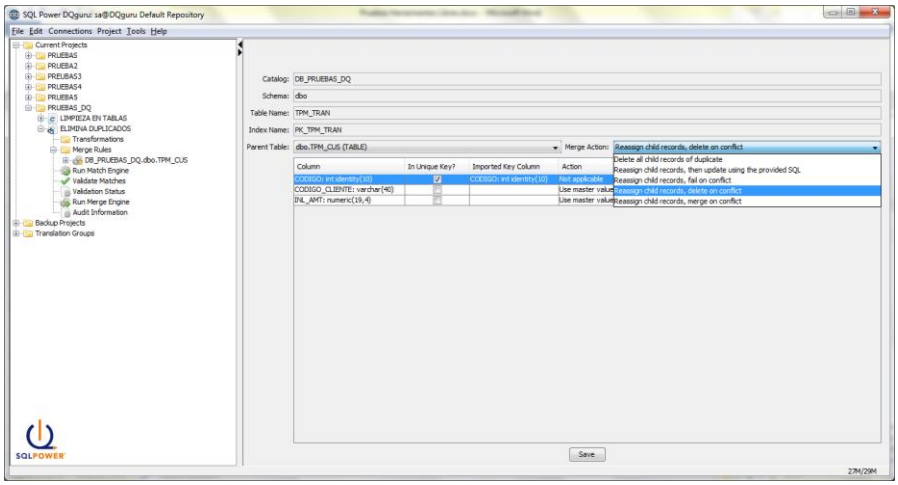


Figura 48. Merge Rules
Fuente: Investigador

Luego se procede a ejecutar la transformación echa anteriormente con el fin de hallar los posibles duplicados para ello se selecciona la opción "Run Match Engine"

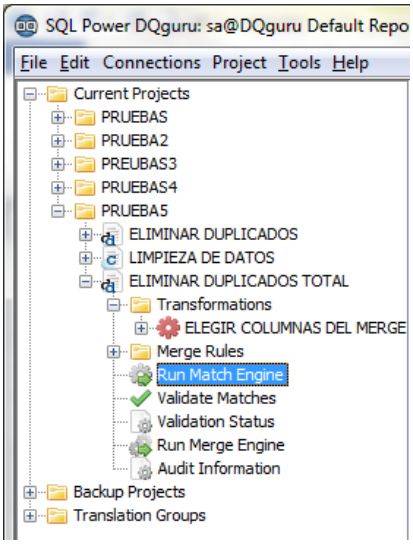


Figura 49. Run Match Engine
Fuente: Investigador

La cual muestra otra pantalla en la cual se pulsa el botón para ejecutar la búsqueda.

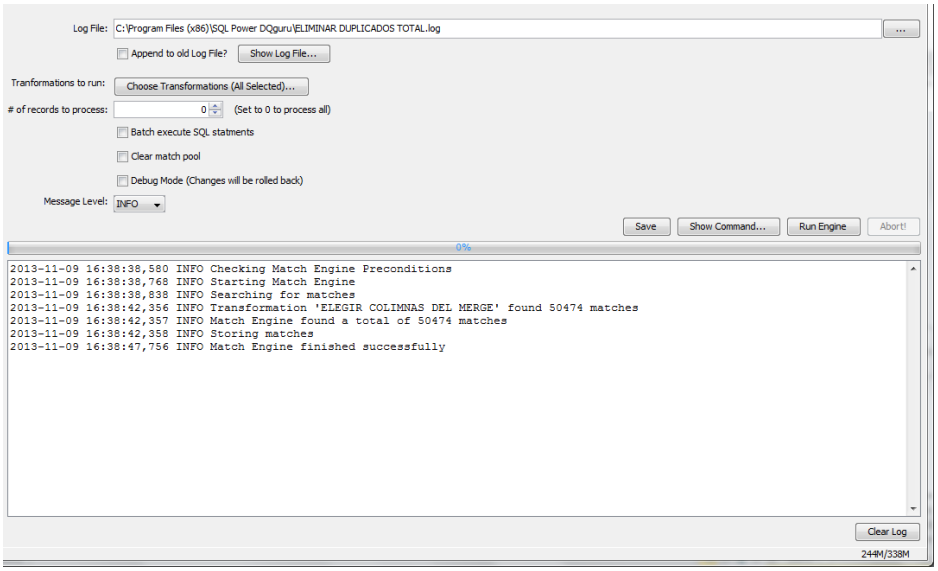


Figura 50. Run Match Engine
Fuente: Investigador

Donde se observa las coincidencias que se han encontrado las cuales se pueden utilizar para la limpieza de duplicados.

Se puede elegir la opción de auto-matching con el fin de que se realice una auto comparación y la herramienta elija la mejor opción.

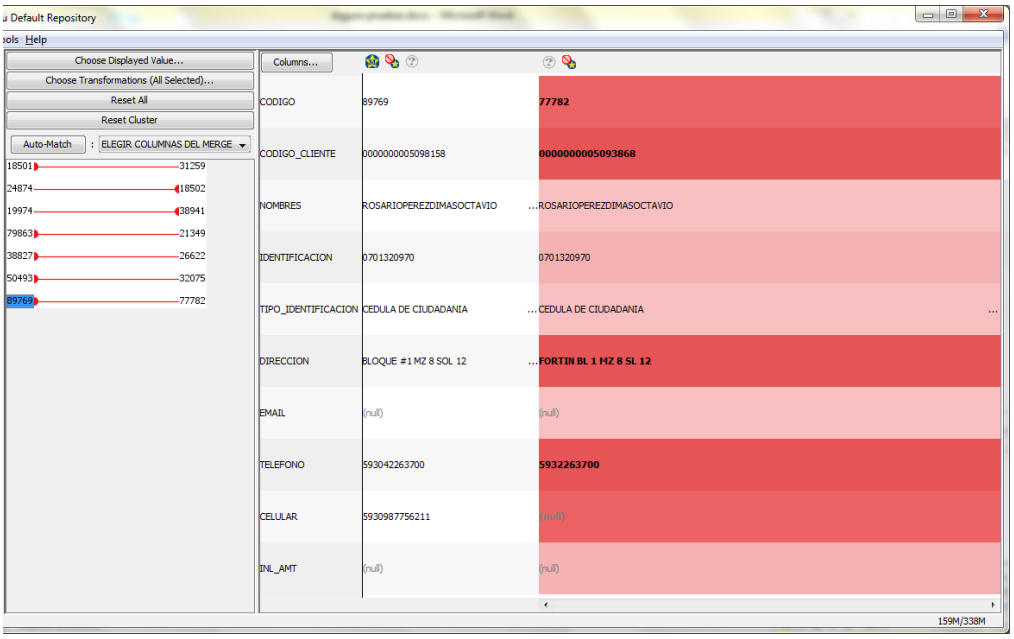


Figura 51. Auto-Matching

Fuente: Investigador

La columna que se muestra con rojo es la columna que esta repetida se puede elegir la columna que se desee que permanezca en nuestra base de datos de acuerdo con las reglas del negocio con las que se está trabajando en mi caso se conserva el registro con mayor número de columnas llenas para todos los casos.

Esta es una advertencia de la herramienta para que se revise si la auto-comparación realizada está de acuerdo con nuestras reglas del negocio.

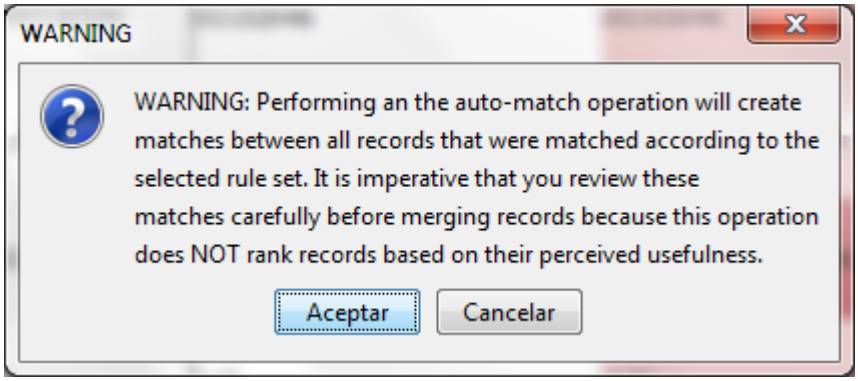


Figura 52. Mensaje de Precaución
Fuente: Investigador

Después de validar las coincidencias en la opción "Validation Matches", se valida el estado de las reglas de coincidencias aplicadas y cuál será el conteo de columnas que van a cambiar

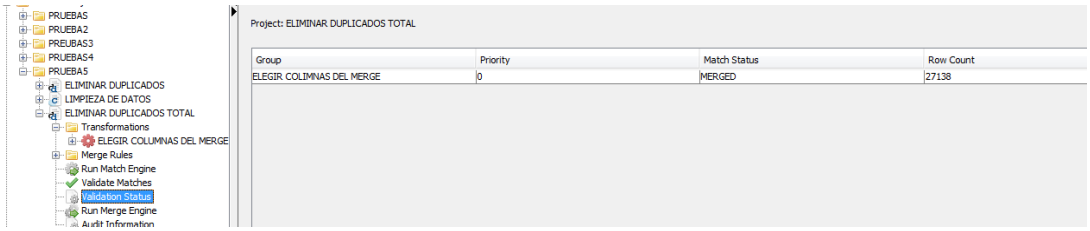


Figura 53. Validation Matches
Fuente: Investigador

El conteo de columnas que van a hacer afectadas se muestra en total y que se realizarán este número de afectaciones en la tabla.

Finalmente se escoge la opción "Run Merge Engine" para ejecutar la limpieza de los datos

Esta es la pantalla que nos muestra que la limpieza de duplicados ha sido un éxito con el número de afectaciones realizadas.

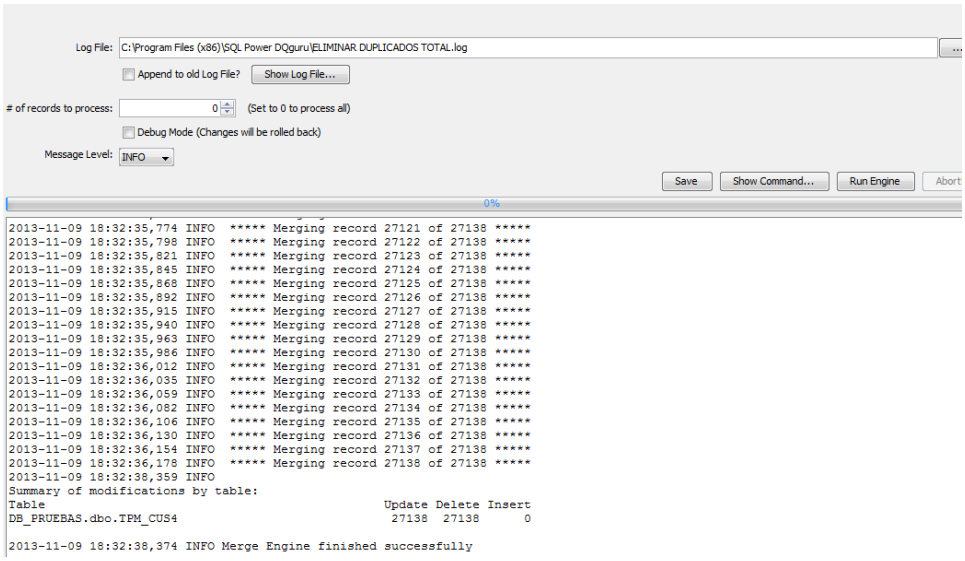
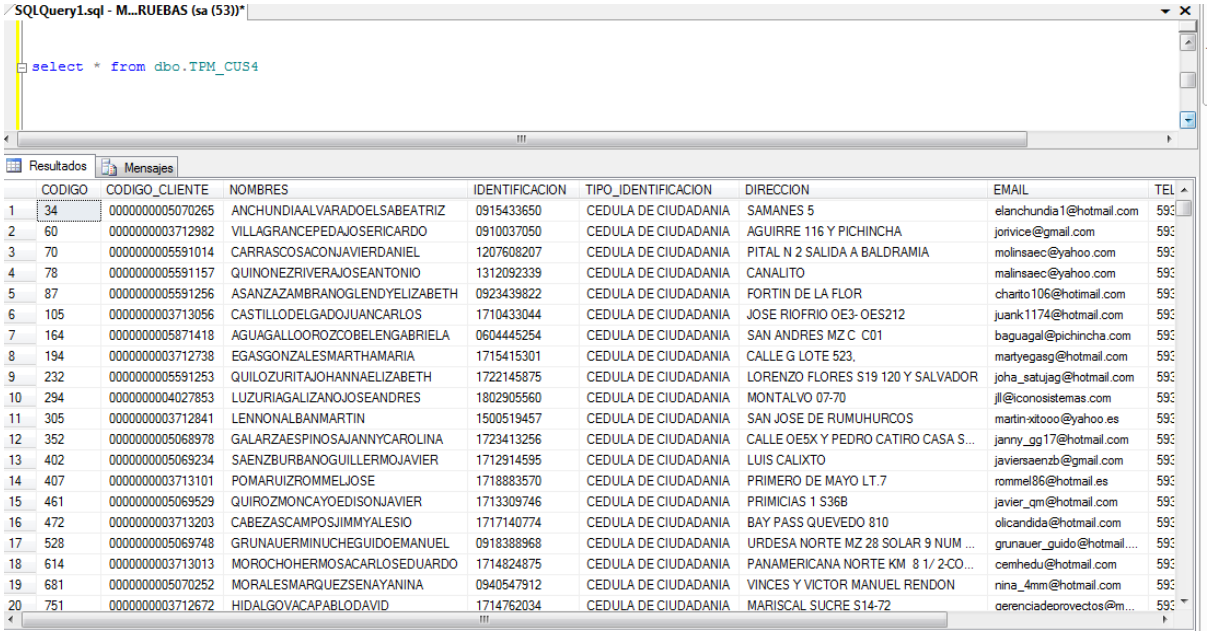


Figura 54. Limpieza de Duplicados
Fuente: Investigador

Para verificar los datos en la tabla se tiene la siguiente imagen que muestra la cantidad de registros que han quedado después de la limpieza.

Este es el resultado de realizar la limpieza de la tabla y dejarla sin duplicados



	CODIGO	CODIGO_CLIENTE	NOMBRES	IDENTIFICACION	TIPO_IDENTIFICACION	DIRECCION	EMAIL	TEL
1	34	0000000005070265	ANCHUNDIAALVARADOELSABEATRIZ	0915433650	CEDULA DE CIUDADANIA	SAMANES 5	elanchundia1@hotmail.com	593
2	60	0000000003712982	VILLAGRANCEPEDAJOSERICARDO	0910037050	CEDULA DE CIUDADANIA	AGUIRRE 116 Y PICHINCHA	jorivice@gmail.com	593
3	70	0000000005591014	CARRASCOSACONJAVIERDANIEL	1207608207	CEDULA DE CIUDADANIA	PITAL N 2 SALIDA A BALDRAMIA	molinsaec@yahoo.com	593
4	78	0000000005591157	QUINONEZRIVERAJOSEANTONIO	1312092339	CEDULA DE CIUDADANIA	CANALITO	malinsaec@yahoo.com	593
5	87	0000000005591256	ASANZAZAMBRANOGLENDYELIZABETH	0923439822	CEDULA DE CIUDADANIA	FORTIN DE LA FLOR	charito106@hotmail.com	593
6	105	0000000003713056	CASTILLODELGADOJUANCARLOS	1710433044	CEDULA DE CIUDADANIA	JOSE RIOFRIO OE3- OES212	juank1174@hotmail.com	593
7	164	0000000005871418	AGUAGALLOOROZCOBELENGABRIELA	0604445254	CEDULA DE CIUDADANIA	SAN ANDRES MZ C C01	baguagal@pichincha.com	593
8	194	0000000003712738	EGASGONZALESMARTHAMARIA	1715415301	CEDULA DE CIUDADANIA	CALLE G LOTE 523,	martyegasg@hotmail.com	593
9	232	0000000005591253	QUILOZURITAJOHANNAELIZABETH	1722145875	CEDULA DE CIUDADANIA	LORENZO FLORES S19 120 Y SALVADOR	joha_satujag@hotmail.com	593
10	294	0000000004027853	LUZURIAGALIZANOJOSEANDRES	1802905560	CEDULA DE CIUDADANIA	MONTALVO 07-70	jll@iconosistemas.com	593
11	305	0000000003712841	LENNONALBANMARTIN	1500519457	CEDULA DE CIUDADANIA	SAN JOSE DE RUMUHURCOS	martin-xitooo@yahoo.es	593
12	352	0000000005068978	GALARZAESPINOSAJANNYCAROLINA	1723413256	CEDULA DE CIUDADANIA	CALLE OE5X Y PEDRO CATIRO CASA S...	janny_gg17@hotmail.com	593
13	402	0000000005069234	SAENZBURBANOGUILLERMOJAVIER	1712914595	CEDULA DE CIUDADANIA	LUIS CALIXTO	javiersaenzb@gmail.com	593
14	407	0000000003713101	POMARUIZROMMELJOSE	1718883570	CEDULA DE CIUDADANIA	PRIMERO DE MAYO LT.7	rommel86@hotmail.es	593
15	461	0000000005069529	QUIROZMONCAYOEDISONJAVIER	1713309746	CEDULA DE CIUDADANIA	PRIMICIAS 1 S36B	javier_qm@hotmail.com	593
16	472	0000000003713203	CABEZASCAMPOSJIMMYALESIO	1717140774	CEDULA DE CIUDADANIA	BAY PASS QUEVEDO 810	olicandida@hotmail.com	593
17	528	0000000005069748	GRUNAUERMINUCHEGUIDOEMANUEL	0918388968	CEDULA DE CIUDADANIA	URDESA NORTE MZ 28 SOLAR 9 NUM ...	grunauer_guido@hotmail....	593
18	614	0000000003713013	MOROCHOHERMOSACARLOSEDUARDO	1714824875	CEDULA DE CIUDADANIA	PANAMERICANA NORTE KM 8 1/ 2-CO...	cemhedu@hotmail.com	593
19	681	0000000005070252	MORALESMARQUEZSENAYANINA	0940547912	CEDULA DE CIUDADANIA	VINCES Y VICTOR MANUEL RENDON	nina_4mm@hotmail.com	593
20	751	0000000003712672	HIDALGOVACAPABLODAVID	1714762034	CEDULA DE CIUDADANIA	MARISCAL SUCRE S14-72	gerenciadeproyectos@m...	593

Figura 55. Limpieza de Duplicados
Fuente: Investigador

Resultados de los datos después de la limpieza de datos.

Los resultados obtenidos de la limpieza con la herramienta se muestran a continuación:

Para análisis de los datos después de la limpieza se utilizó la herramienta DataCleaner debido a que SQL Power no cuenta con esta función.

Los resultados obtenidos de la herramienta DataCleaner se muestran a continuación:

Tabla: TPM_CUS

Los valores nulos.

No se encontraron valores nulos en esta tabla.

Caracteres Inválidos.

Tabla 17 Resultados Tabla TPM_CUS

Fuente: Investigador

TPM_CUS	
COLUMNA	**CARACTERES INVÁLIDOS**
CODIGO	0
CODIGO_CLIENTE	0
NOMBRES	0
IDENTIFICACION	0
TIPO_IDENTIFICACION	0
DIRECCION	1568
EMAIL	5118
TELEFONO	182
CELULAR	9794

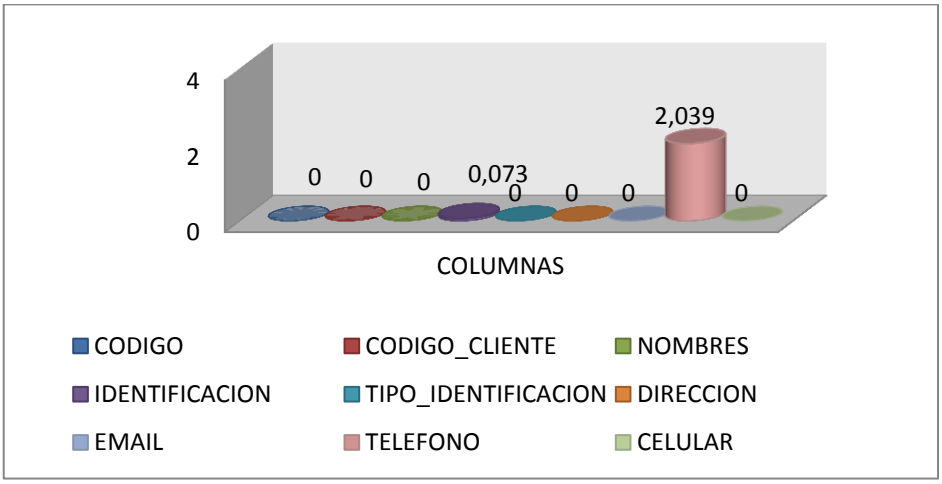


Figura 56. Resultados Tabla TPM_CUS
Fuente: Investigador

Registros duplicados.

No se encontraron valores duplicados en esta tabla.

Interpretación de resultados para la tabla TPM_CUS.

En el análisis de los resultados después de la limpieza en la tabla TPM_CUS observamos que los datos fueron mejorados con el uso de las funciones de limpieza que la herramienta brinda, de esta forma podemos evidenciar que los datos han mejorado en comparación con el análisis preliminar de la tabla.

Tabla: TPM_TRAN

Valores Nulos.

No se encontraron valores nulos en esta tabla.

Caracteres Inválidos.

No se presentaron valores invalidos en esta tabla.

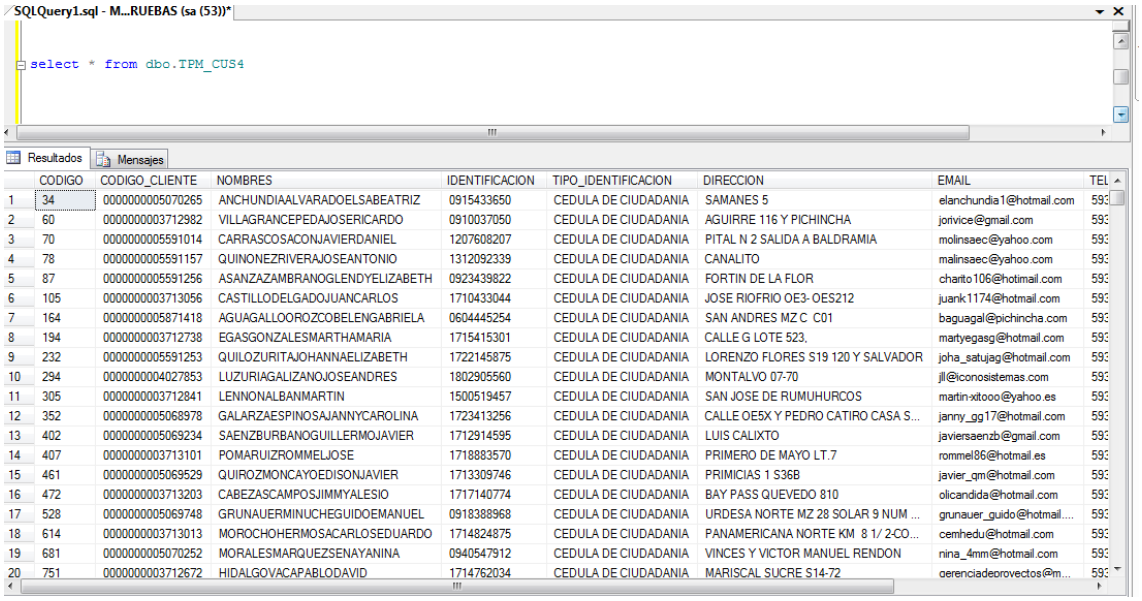
Registros duplicados.

No se presentaron valores duplicados en esta tabla.

Interpretación de resultados para la tabla TPM_TRAN.

En el análisis de los resultados después de la limpieza en la tabla TPM_TRAN observamos que los datos fueron mejorados con el uso de las funciones de limpieza que la herramienta brinda, de esta forma podemos evidenciar que los datos han mejorado en comparación con el análisis preliminar de la tabla.

Los datos obtenidos después del uso de la herramienta



	CODIGO	CODIGO_CLIENTE	NOMBRES	IDENTIFICACION	TIPO_IDENTIFICACION	DIRECCION	EMAIL	TEL
1	34	0000000005070265	ANCHUNDIAALVARADOELSABEATRIZ	0915433650	CEDULA DE CIUDADANIA	SAMANES 5	elanchundia1@hotmail.com	593
2	60	0000000003712982	VILLAGRANCEPEDAJOSERICARDO	0910037050	CEDULA DE CIUDADANIA	AGUIRRE 116 Y PICHINCHA	jorivice@gmail.com	593
3	70	0000000005591014	CARRASCOSACONJAVIERDANIEL	1207608207	CEDULA DE CIUDADANIA	PITAL N 2 SALIDA A BALDRAMIA	molinsaec@yahoo.com	593
4	78	0000000005591157	QUINONEZRIVERAJOSEANTONIO	1312092339	CEDULA DE CIUDADANIA	CANALITO	malinsaec@yahoo.com	593
5	87	0000000005591256	ASANZAZAMBRANOGLENDYELIZABETH	0923439822	CEDULA DE CIUDADANIA	FORTIN DE LA FLOR	charito106@hotmail.com	593
6	105	0000000003713056	CASTILLODELGADOJUANCARLOS	1710433044	CEDULA DE CIUDADANIA	JOSE RIOFRIO OE3- OES212	juank1174@hotmail.com	593
7	164	0000000005871418	AGUAGALLOOROZCOBELENGABRIELA	0604445254	CEDULA DE CIUDADANIA	SAN ANDRES MZ C C01	baguagal@pichincha.com	593
8	194	0000000003712738	EGASGONZALESMARTHAMARIA	1715415301	CEDULA DE CIUDADANIA	CALLE G LOTE 523,	martyegasg@hotmail.com	593
9	232	0000000005591253	QUILOZURITAJOHANNAELIZABETH	1722145875	CEDULA DE CIUDADANIA	LORENZO FLORES S19 120 Y SALVADOR	joha_satujag@hotmail.com	593
10	294	0000000004027853	LUZURIAGALIZANOJOSEANDRES	1802905560	CEDULA DE CIUDADANIA	MONTALVO 07-70	jll@iconosistemas.com	593
11	305	0000000003712841	LENNONALBANMARTIN	1500519457	CEDULA DE CIUDADANIA	SAN JOSE DE RUMUHURCOS	martin-xitooo@yahoo.es	593
12	352	0000000005068978	GALARZAESPINOSAJANNYCAROLINA	1723413256	CEDULA DE CIUDADANIA	CALLE OE5X Y PEDRO CATIRO CASA S...	janny_gg17@hotmail.com	593
13	402	0000000005069234	SAENZBURBANOGUILLERMOJAVIER	1712914595	CEDULA DE CIUDADANIA	LUIS CALIXTO	javiersaenzb@gmail.com	593
14	407	0000000003713101	POMARUIZROMMELJOSE	1718883570	CEDULA DE CIUDADANIA	PRIMERO DE MAYO LT.7	rommel86@hotmail.es	593
15	461	0000000005069529	QUIROZMONCAYOEDISONJAVIER	1713309746	CEDULA DE CIUDADANIA	PRIMICIAS 1 S36B	javier_qm@hotmail.com	593
16	472	0000000003713203	CABEZASCAMPOSJIMMYALESIO	1717140774	CEDULA DE CIUDADANIA	BAY PASS QUEVEDO 810	olicandida@hotmail.com	593
17	528	0000000005069748	GRUNAUERMINUCHEGUIDOEMANUEL	0918388968	CEDULA DE CIUDADANIA	URDESA NORTE MZ 28 SOLAR 9 NUM ...	grunauer_guido@hotmail....	593
18	614	0000000003713013	MOROCHOHERMOSACARLOSEDUARDO	1714824875	CEDULA DE CIUDADANIA	PANAMERICANA NORTE KM 8 1/ 2-CO...	cemhedu@hotmail.com	593
19	681	0000000005070252	MORALESMARQUEZSENAYANINA	0940547912	CEDULA DE CIUDADANIA	VINCES Y VICTOR MANUEL RENDON	nina_4mm@hotmail.com	593
20	751	0000000003712672	HIDALGOVACAPABLODAVID	1714762034	CEDULA DE CIUDADANIA	MARISCAL SUCRE S14-72	gerenciadeproyectos@m...	593

Figura 57 Resultados Limpieza TPM_TRAN

Fuente: Investigador

Resultado del uso de la herramienta.

- La configuración de conexión es muy sencilla e intuitiva.
- La herramienta no cuenta con objetos de análisis de datos previos a la limpieza (profiling), lo cual se presenta como una desventaja.
- La limpieza de los datos es muy fácil y sencilla de configurar ya que el uso de los objetos de manera visual facilita la comprensión.
- La limpieza de duplicados también representa una forma sencilla de su uso y es muy eficiente.
- Los datos son afectados de manera inmediata por lo que se debe tener cuidado con los cambios a realizar en el momento de ejecutarlos, para ello se recomienda realizar back up de los datos.

ORACLE DATA QUALITY

Configurar la conexión con los datos.

Para iniciar el controlador de conexiones (Gestor de Metabase) que debe tener la herramienta previamente instalado y se selecciona del menú de inicio de nuestro equipo.

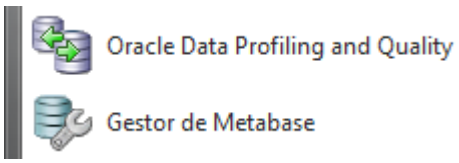


Figura 58. Oracle Data Quality
Fuente: Investigador

Se ingresa las credenciales del administrador para configurar.

- Repositorio
- Usuario
- Contraseña

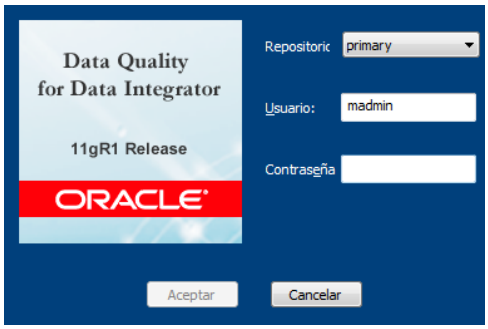


Figura 59. Acceso
Fuente: Investigador

Se presenta la pantalla de inicio

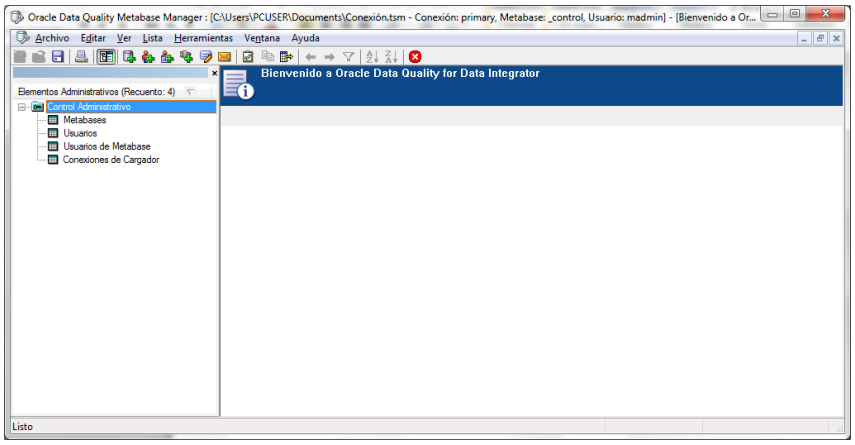


Figura 60. Pantalla de inicio
Fuente: Investigador

Se crea una nueva Metabase seleccionando de la raíz del explorador

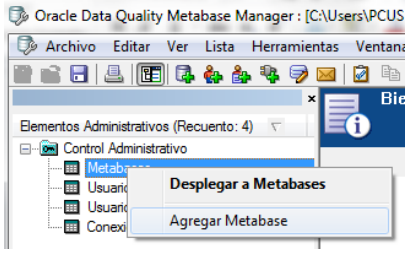


Figura 61. Crear Metabase
Fuente: Investigador

Se selecciona “Agregar Metabase” y llena los parámetros indicados

- Nombre
- Patrón
- Tamaño de la cache pública (el tamaño de la cache publica por defecto es de 16MB pero si disponemos de un buen servidor esta puede aumentar)

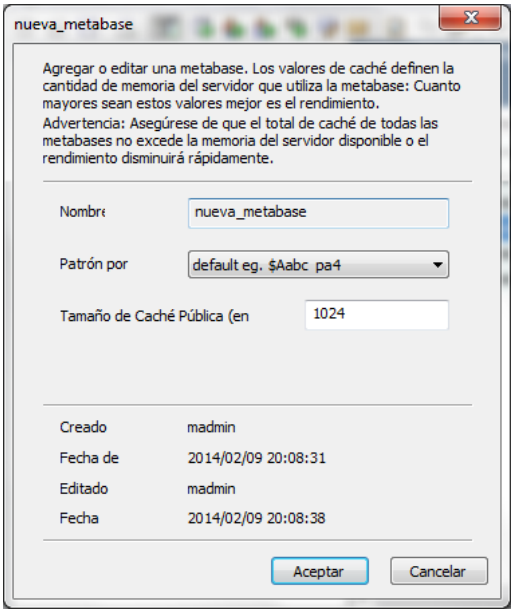


Figura 62. Agregar Metabase
Fuente: Investigador

Se pueden visualizar y reconfigurar los metabases existentes una vez creada.

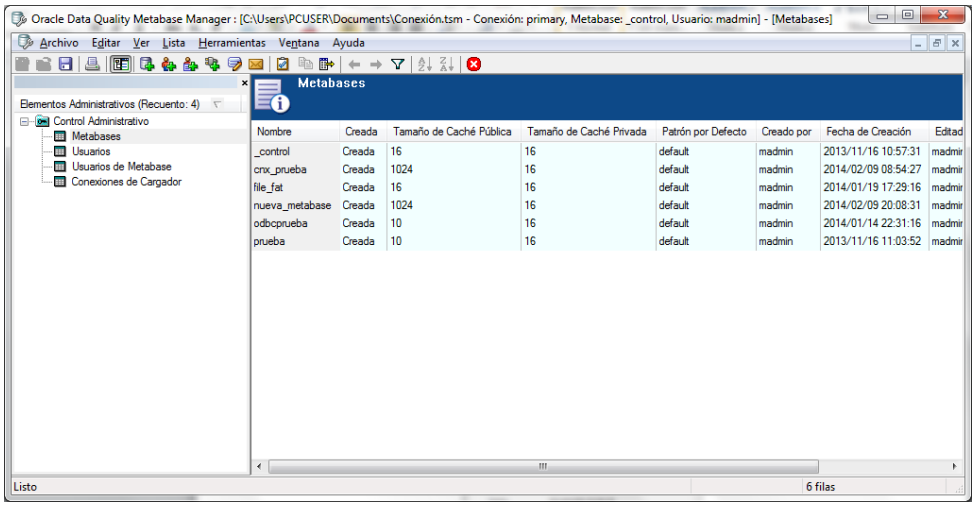


Figura 63 Metabases Creados
Fuente: Investigador

Luego se crea un usuario que sea quien se conecte a la metabase, para lo cual se selecciona de la raíz del explorador "Agregar Usuario…" y se configura las credenciales.

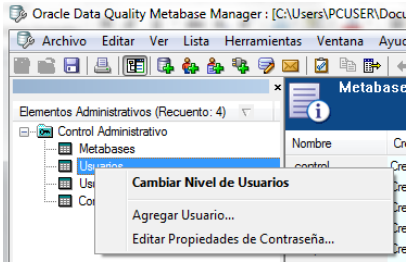


Figura 64 Agregar Usuario
Fuente: Investigador

Se configura:

- Nombre: nuevo_usuario
- Forzar vencimiento (no seleccionamos esta opción)
- Contraseña
- Confirmación

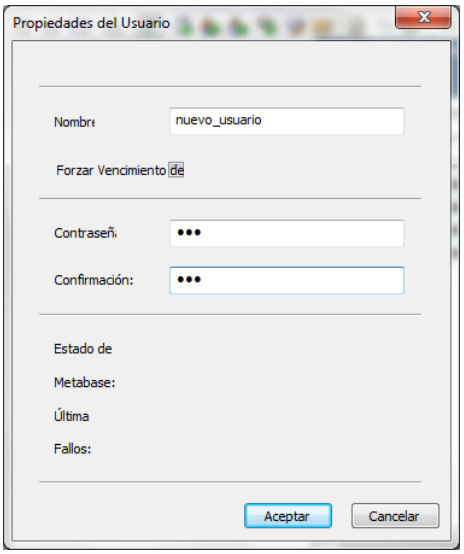


Figura 65 Propiedades de Usuario
Fuente: Investigador

Se revisa la creación del usuario.

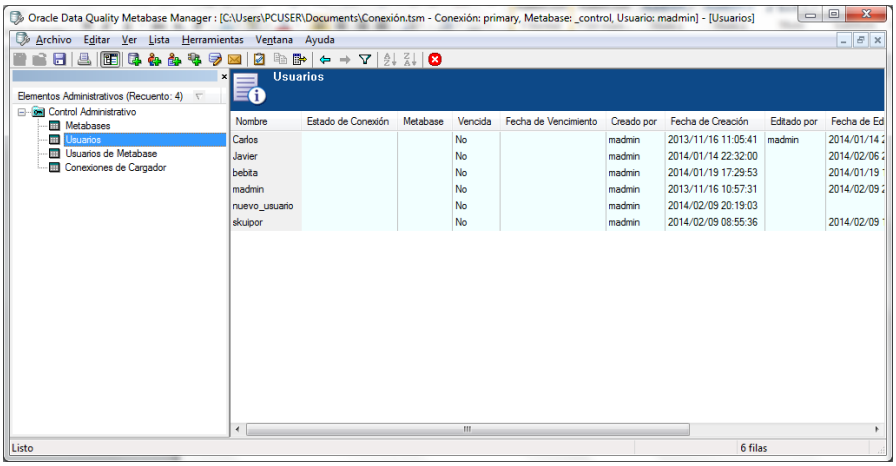


Figura 66. Vista Usuario
Fuente: Investigador

Ahora se asigna el usuario creado para que pueda conectarse a la metabase creada con anterioridad, para ello seleccionamos "Agregar usuario a metabase" del explorador raíz.

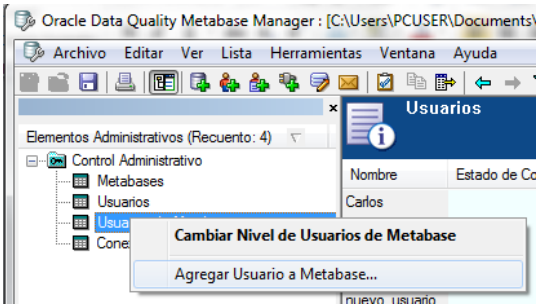


Figura 67. Agregar usuario
Fuente: Investigador

En la ventana que se despliega se selecciona al usuario y la metabase creada con anterioridad

- Nombre de usuario
- Metabase
- Usuario limitado (no seleccionamos esta opción ya que esto disminuirá los permisos de acceso a la metabase)

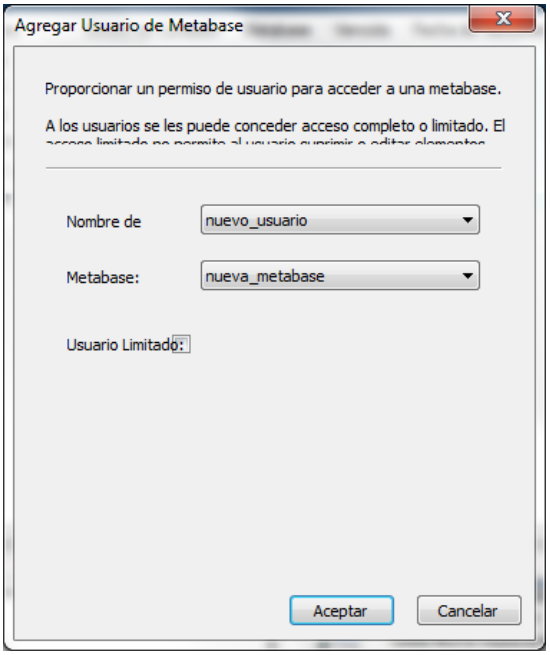


Figura 68. Agregar Usuario Metabase
Fuente: Investigador

Para finalizar se configura la conexión con los datos, para ello se selecciona "Agregar Conexión de Cargador"

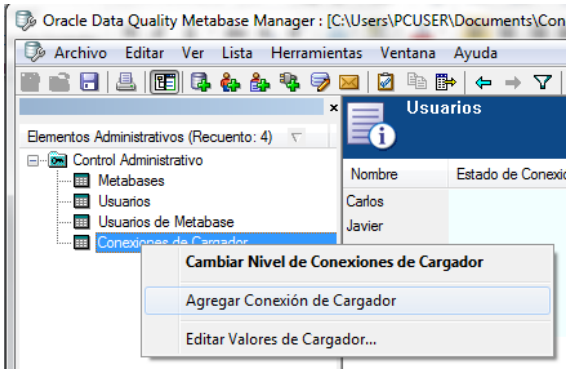


Figura 69. Agregar Conexión de Cargador
Fuente: Investigador

Se llenan los datos con la conexión que se va a realizar en nuestro caso usaremos una conexión ODBC

- Nombre
- Tipo de conexión

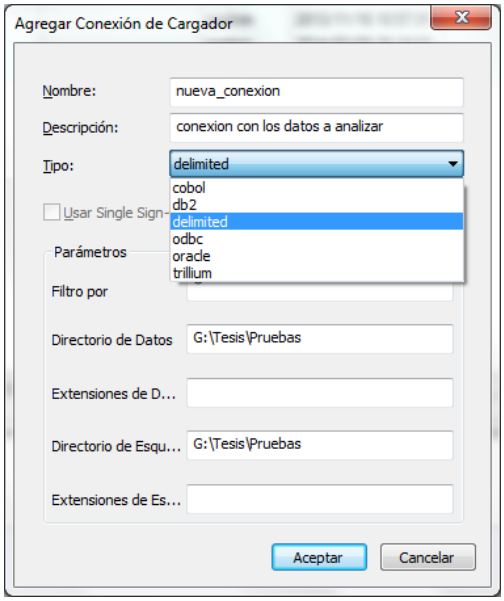


Figura 70. Pantalla Agregar Conexión
Fuente: Investigador

Antes de finalizar se definimos el acceso a la metabase o metabases que se inicialicen con esta conexión y se presiona aceptar.

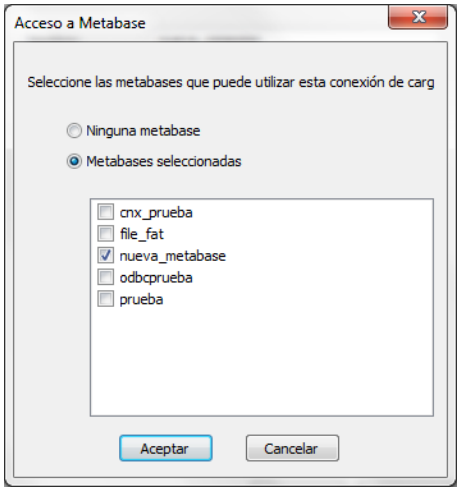


Figura 71. Acceso a Metabase
Fuente: Investigador

Luego se presiona aceptar para finalizar y revise la creación del cargador de objetos

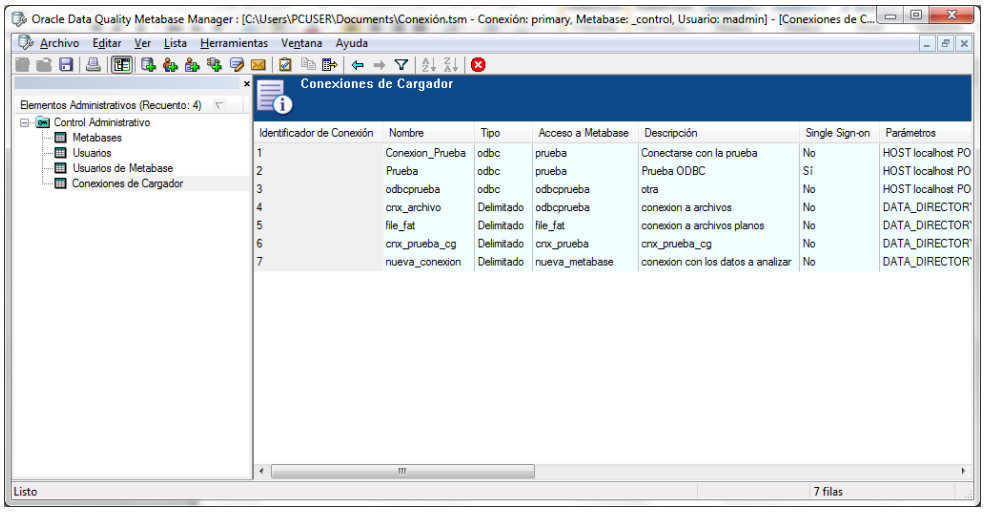


Figura 72. Creación del cargador
Fuente: Investigador

Una vez configurado el acceso a la metabase, abrir del menú inicio la herramienta de Oracle Data Profiling and Quality.

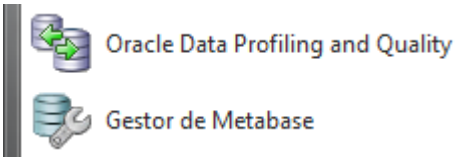


Figura 73. Menú Inicio
Fuente: Investigador

Ingresar con los datos de las credenciales creadas con anterioridad

Figura 74. Pantalla Acceso
Fuente: Investigador

La herramienta muestra la pantalla de inicio

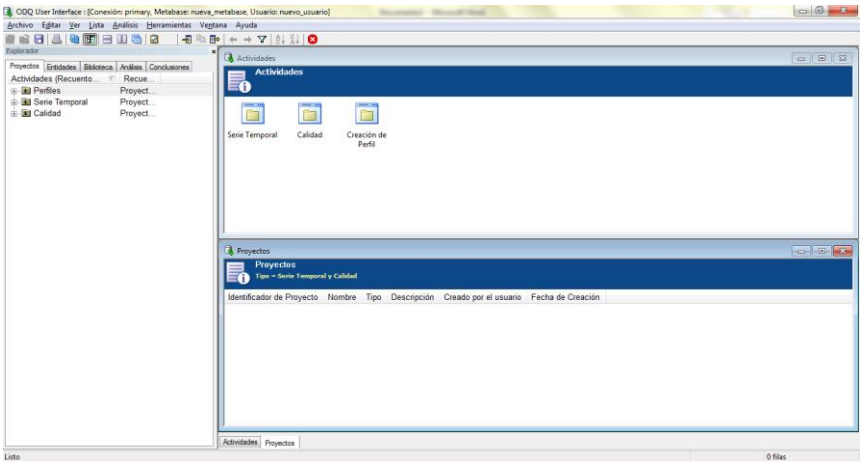


Figura 75. Pantalla de Inicio
Fuente: Investigador

Para acceder a los datos que se tienen configurados, se selecciona la opción "Análisis" del menú principal.

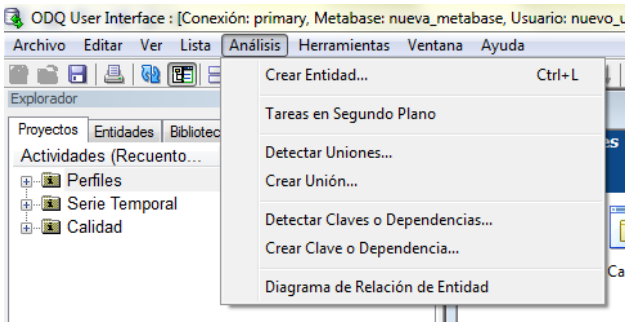


Figura 76. Pestaña de Análisis
Fuente: Investigador

Crear la entidad para extraer las tablas que ahí existen

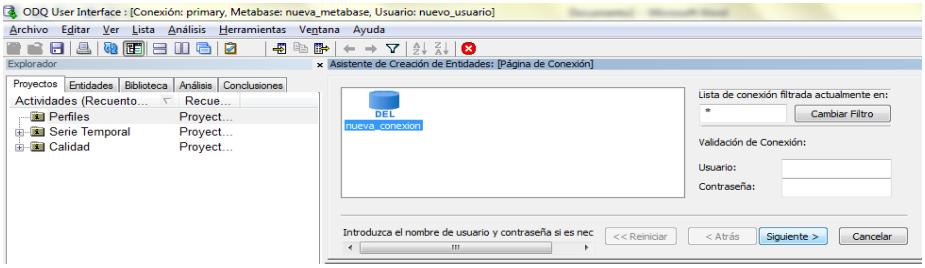


Figura 77. Crear Entidad
Fuente: Investigador

Aquí presenta la conexión a los datos que fueron configurados con anterioridad, seleccione “nueva conexión” y presione siguiente.

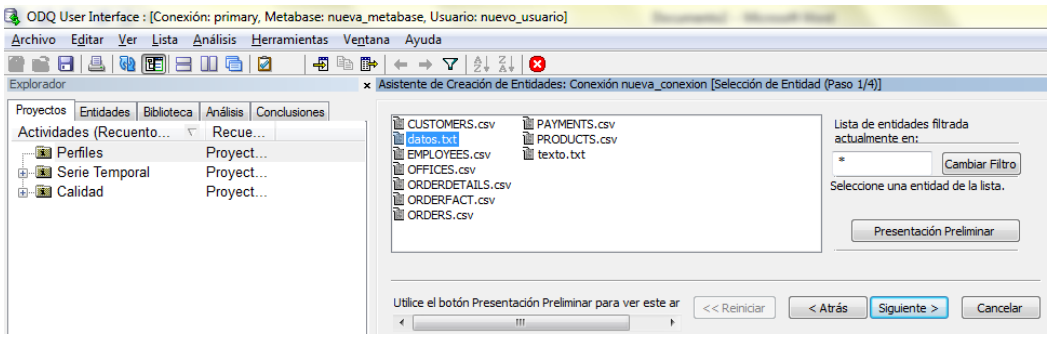


Figura 78. Nueva conexión
Fuente: Investigador

Esta es una vista previa de los datos: opción “Presentación preliminar”.

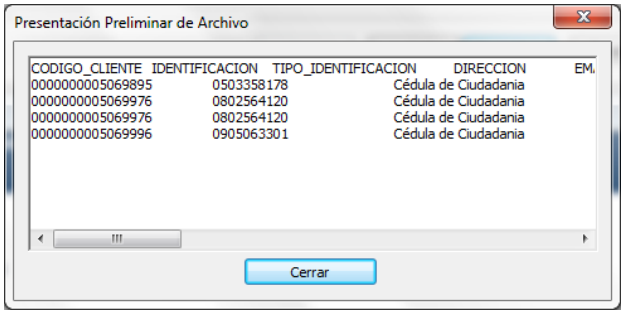


Figura 79. Presentación preliminar
Fuente: Investigador

Carga los datos de la fuente

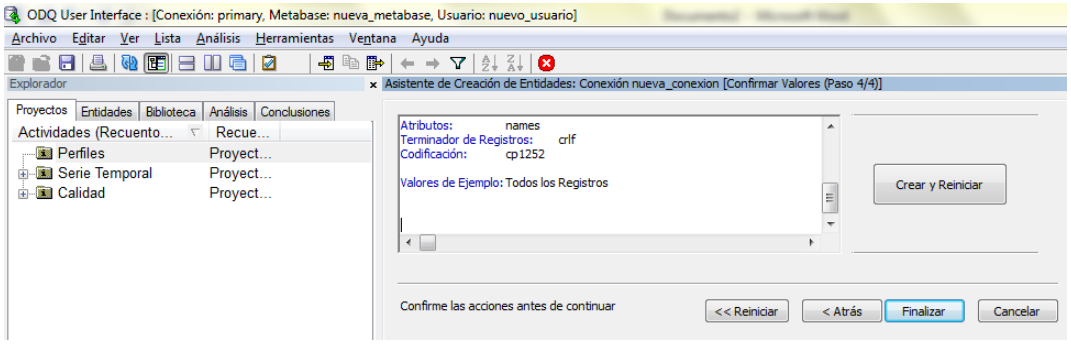


Figura 80. Cargar datos
Fuente: Investigador

Para finalizar presenta un resumen de la configuración y tenemos la opción de Crear y Reiniciar si se desea crear otra entidad o finalizar para continuar con otros trabajos.

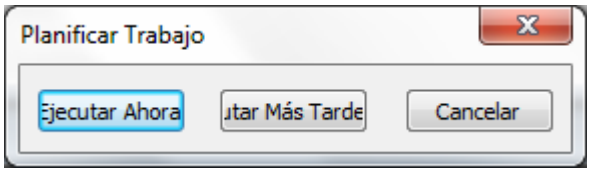


Figura 81. Opciones
Fuente: Investigador

Se planifica el trabajo para que se ejecute ahora y proceder con los otros trabajos.

Para revisar el trabajo realizado y si la creación de la entidad fue exitosa se selecciona del explorador de objetos la pestaña de entidades

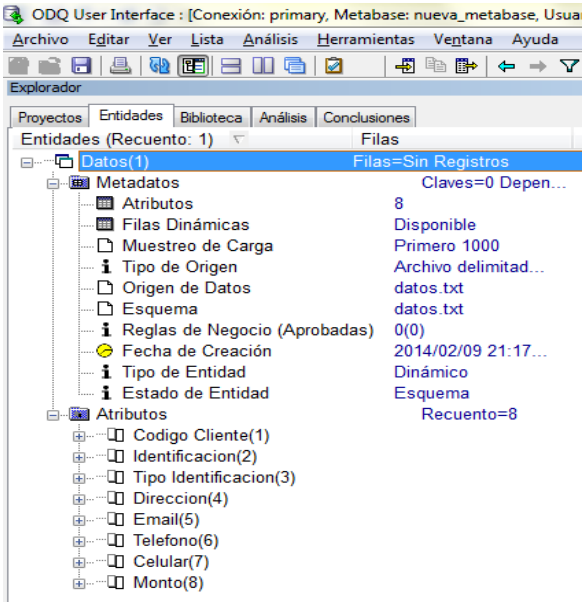


Figura 82. Pestaña de Entidades
Fuente: Investigador

Perfil de datos

Una vez creada la entidad, crear un proyecto de Profiling o de Calidad empezaremos con el proyecto de perfiles, para crear un proyecto de perfilado de datos se debe seleccionar la pestaña de proyectos y seleccionar Crear Proyecto.

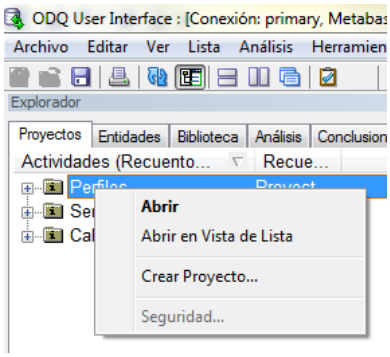


Figura 83. Opción Crear Proyecto
Fuente: Investigador

Se llenamos los datos que pide

- Nombre
- Descripción
- Y selecciona la entidad creada con anterioridad.

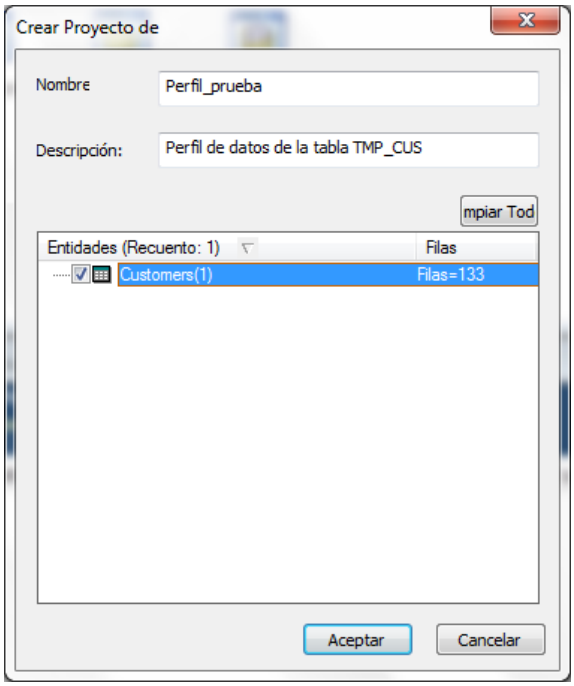


Figura 84. Ingresar datos
Fuente: Investigador

El proyecto creado aparecerá de inmediato con el análisis de la entidad seleccionada

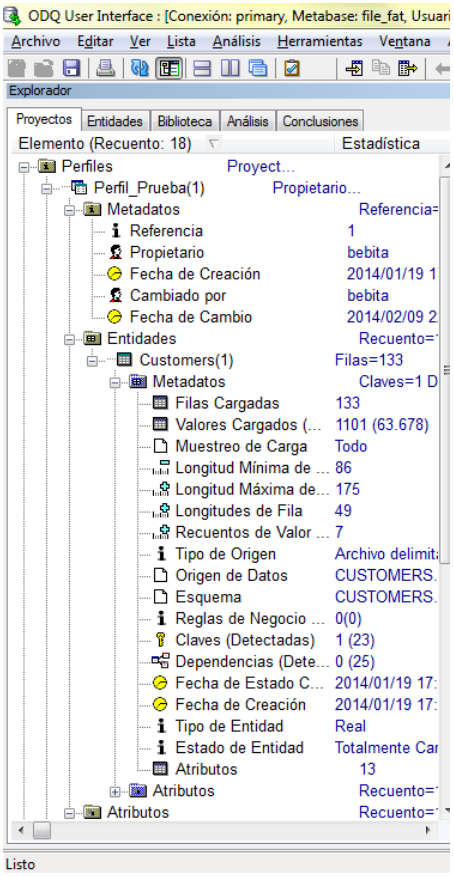


Figura 85. Proyecto Creado
Fuente: Investigador

En el resumen muestra todos los datos que han sido encontrados por la herramienta tales como números de fila, longitud máxima y mínima de cada campo estado de la entidad, etc.

El resultado de los datos analizados es de fácil entendimiento y comprensión para conocer cuales el estado de nuestros datos

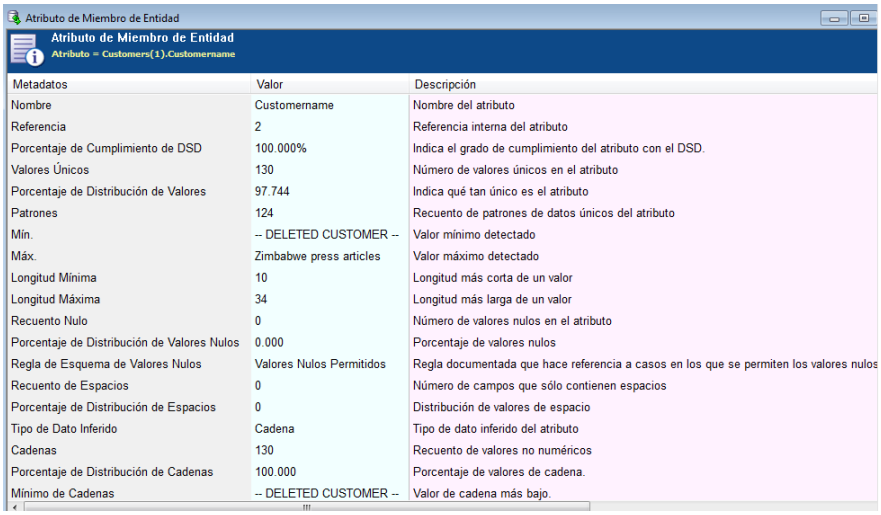


Figura 86. Resultados de Datos
Fuente: Investigador

Los resultados obtenidos de la herramienta Oracle Data Integrator Data Quality se muestran a continuación:

Tabla: TPM_CUS

Los valores nulos.

Tabla 18. Valores Nulos de la tabla TPM_CUS
Fuente: Investigador

TPM_CUS	
COLUMNA	**VALORES NULOS**
CODIGO	0
CODIGO_CLIENTE	0
NOMBRES	0
IDENTIFICACION	0
TIPO_IDENTIFICACION	0
DIRECCION	2652
EMAIL	95118
TELEFONO	382
CELULAR	39794

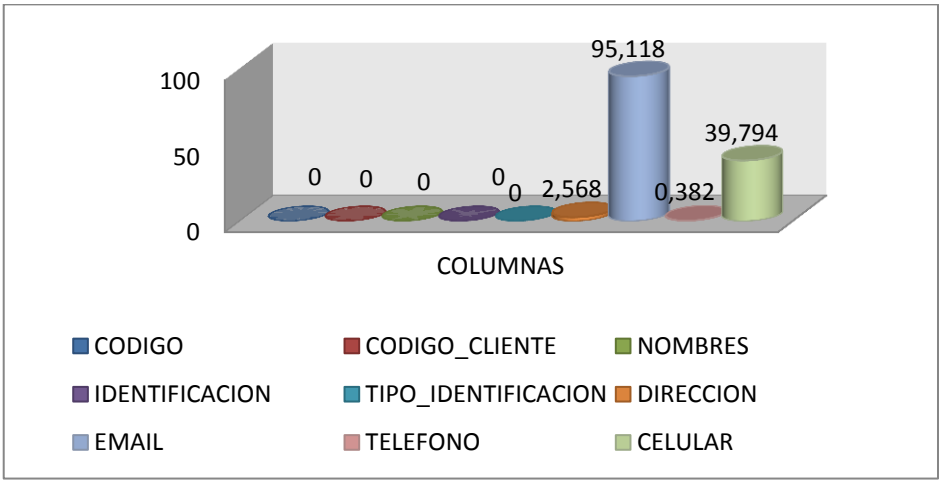


Figura 87. Valores Nulos de la tabla TPM_CUS
Fuente: Investigador

Interpretación de resultados para valores nulos.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan valores nulos a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Caracteres Inválidos.

Tabla 19. Caracteres Inválidos de la tabla TPM_CUS
Fuente: Investigador

TPM_CUS	
COLUMNA	**CARACTERES INVÁLIDOS**
CODIGO	0
CODIGO_CLIENTE	0
NOMBRES	0
IDENTIFICACION	73
TIPO_IDENTIFICACION	0
DIRECCION	0
EMAIL	0
TELEFONO	2039
CELULAR	0

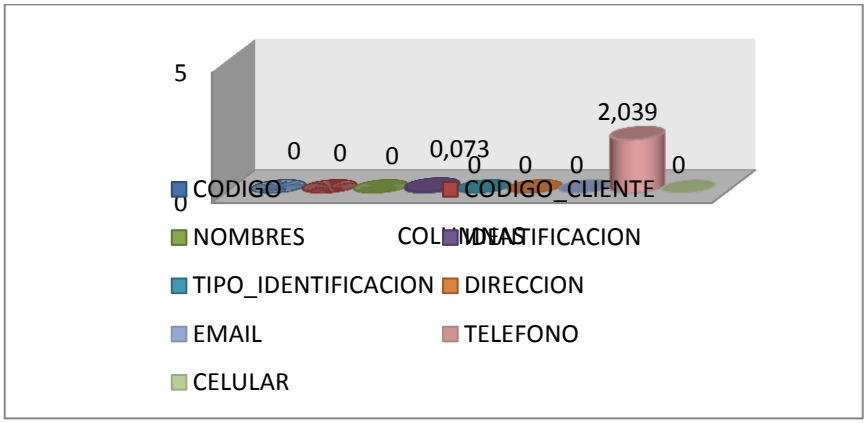


Figura 88. Caracteres Inválidos de la tabla TPM_CUS
Fuente: Investigador

Interpretación de resultados para caracteres inválidos.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan caracteres inválidos a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Registros duplicados.

Tabla 20. Registros Duplicados de la tabla TPM_CUS
Fuente: Investigador

TPM_CUS	
COLUMNA	**DUCPLICADOS**
CODIGO	0
CODIGO_CLIENTE	18015
NOMBRES	18015
IDENTIFICACION	18015
TIPO_IDENTIFICACION	18015
DIRECCION	9250
EMAIL	5032
TELEFONO	1355
CELULAR	576

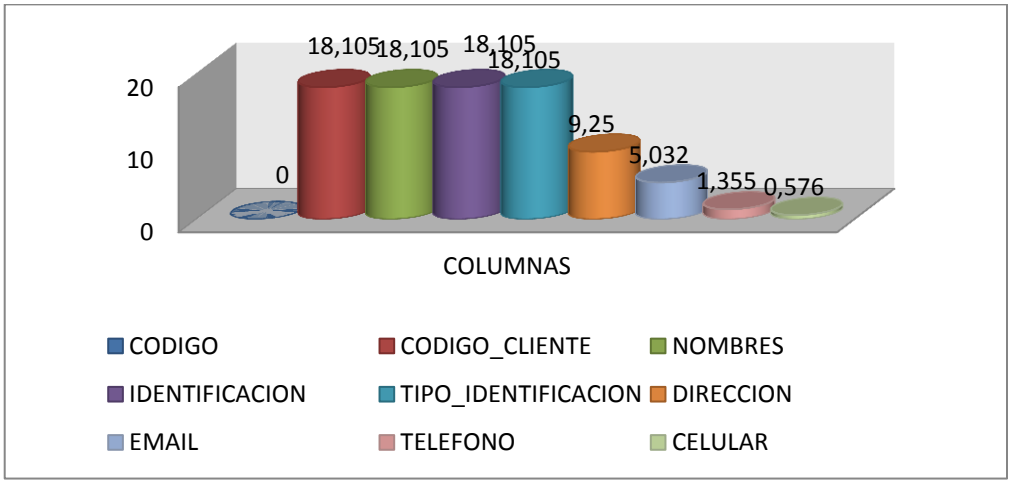


Figura 89. Registros Duplicados de la tabla TPM_CUS
Fuente: Investigador

Interpretación de resultados para duplicados.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan registros duplicados a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Total de registros duplicados.

Tabla 21. Total registros duplicados
Fuente: Investigador

TOTAL REGISTRO DUPLICADOS	
REGISTROS	46492

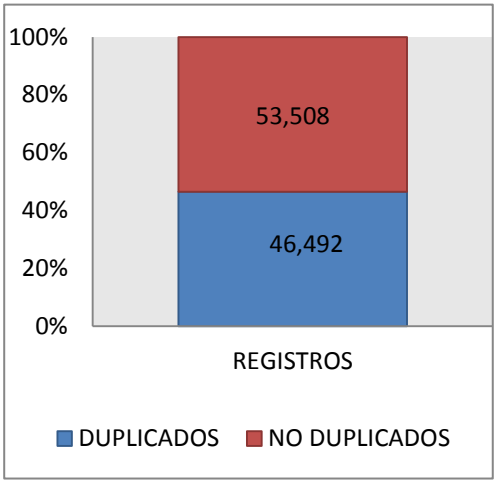


Figura 90. Total registros duplicados
Fuente: Investigador

Interpretación de resultados para total de duplicados.

Como se visualizar en la figura anterior existe un gran número de registros duplicados para la tabla TPM_CUS los cuales deben ser solventados para obtener una mejor calidad de datos en esta tabla.

Tabla: TPM_TRAN

Valores Nulos.

Tabla 22. Total registros duplicados
Fuente: Investigador

TPM_TRAN	
COLUMNA	**VALORES NULOS**
CODIGO	0
CODIGO_CLIENTE	0
INL_AMT	81503

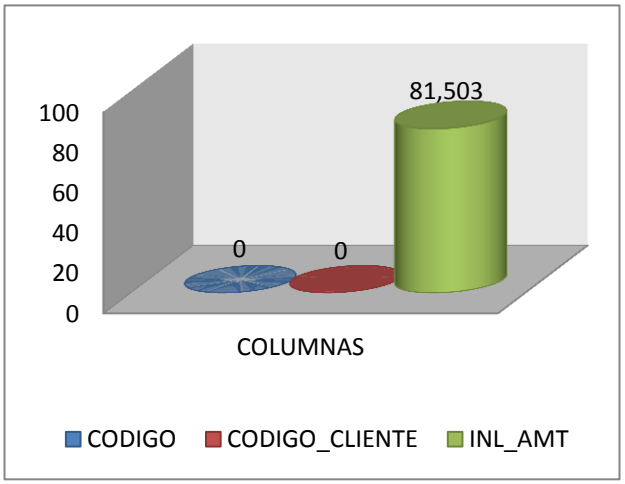


Figura 91. Total registros duplicados
Fuente: Investigador

Interpretación de resultados para valores nulos.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan valores nulos a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Caracteres Inválidos.

No se presentaron valores invalidos en esta tabla.

Interpretación de resultados para caracteres inválidos.

Al realizar el análisis de los datos para esta tabla se evidencio que no existían valores inválidos dentro de los campos de la tabla TPM_TRAN.

Registros duplicados.

Tabla 23. Total registros duplicados

Fuente: Investigador

TPM_TRAN	
COLUMNA	**DUCPLICADOS**
CODIGO	0
CODIGO_CLIENTE	8899
INL_AMT	2024

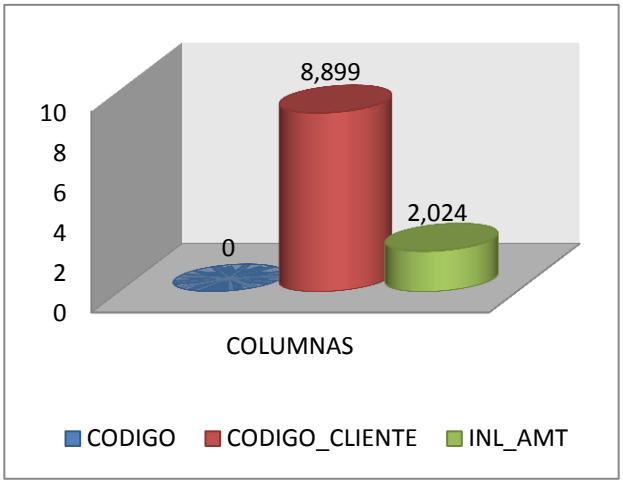


Figura 92. Total registros duplicados
Fuente: Investigador

Interpretación de resultados para duplicados.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan registros duplicados a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Total de registros duplicados.

Tabla 24. Total registros duplicados
Fuente: Investigador

TOTAL REGISTRO DUPLICADOS	
REGISTROS	72775

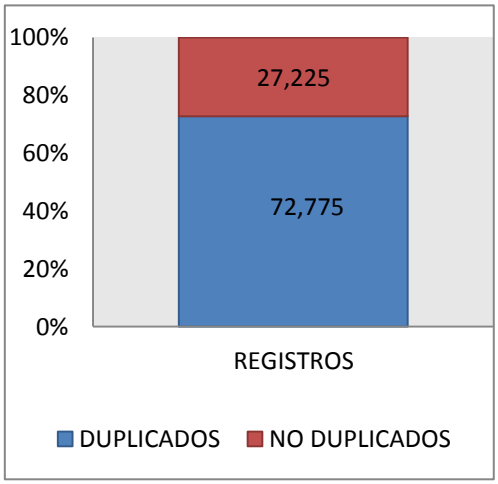


Figura 93. Total registros duplicados
Fuente: Investigador

Interpretación de resultados para total de duplicados.

Como se visualizar en la figura anterior existe un gran número de registros duplicados para la tabla TPM_TRAN los cuales deben ser solventados para obtener una mejor calidad de datos en esta tabla.

Limpieza

Pasos que se deben seguir para el uso de la herramienta:

- Identificar la tabla destino
- Identificar las tablas fuentes
- Identificar las tablas de Referencia (Lookup)
- Verificar los pareos de campos (mapping)
    - Pareos Automáticos
    - Columnas no nulas
    - Añadir columnas adicionales
- Probar regularmente la extracción
- En las transformaciones
    - Identificar, verificar y validar las condiciones
    - Verificar y validar campos y funciones para convertir formatos de fecha
    - Verificar tamaños de columnas para no truncar los datos extraídos que de algún tipo de error
    - Verificar los tipos de datos (Datatype)
    - Verificar las secuencias

Seleccione la entidad que a la cual se aplica las funciones de calidad

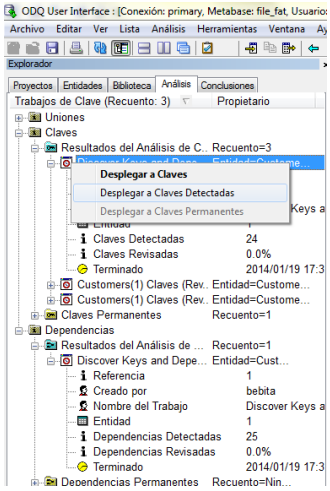


Figura 94. Entidad
Fuente: Investigador

Se crea el proceso de limpieza con las funciones que se detallaron con anterioridad

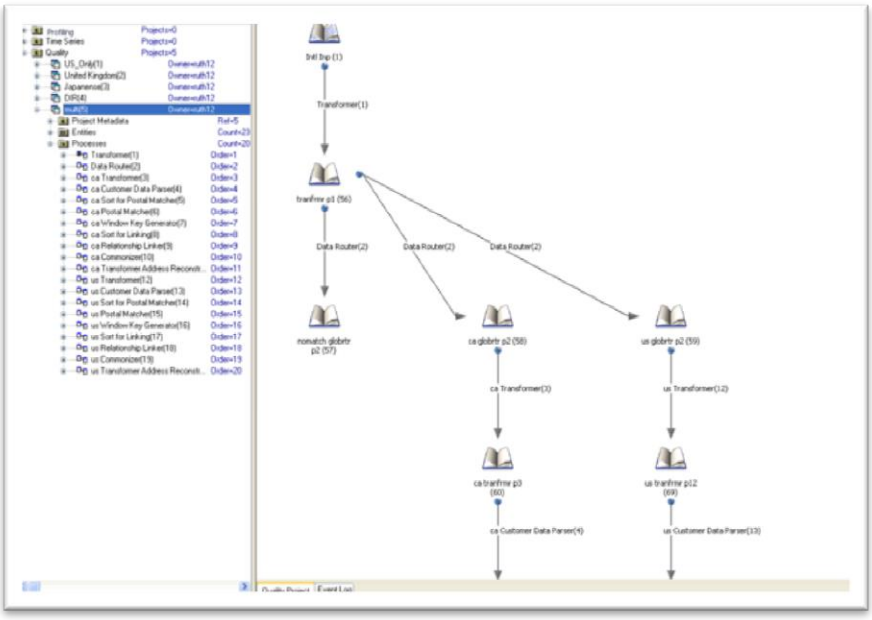


Figura 95. Limpieza
Fuente: Investigador

Los resultados obtenidos de la herramienta Oracle Data Quality después de la limpieza se muestran a continuación:

Tabla: TPM_CUS

Los valores nulos.

No se encontraron valores nulos en esta tabla.

Caracteres Inválidos.

No se encontraron valores inválidos en esta tabla.

Registros duplicados.

No se encontraron valores duplicados en esta tabla.

Interpretación de resultados para la tabla TPM_CUS.

En el análisis de los resultados después de la limpieza en la tabla TPM_CUS observamos que los datos fueron mejorados con el uso de las funciones de limpieza que las herramientas nos brinda, de esta forma podemos evidenciar que los datos han mejorado en comparación con el análisis preliminar de la tabla.

Tabla: TPM_TRAN

Valores Nulos.

No se encontraron valores nulos en esta tabla.

Caracteres Inválidos.

No se presentaron valores invalidos en esta tabla.

Registros duplicados.

No se presentaron valores duplicados en esta tabla.

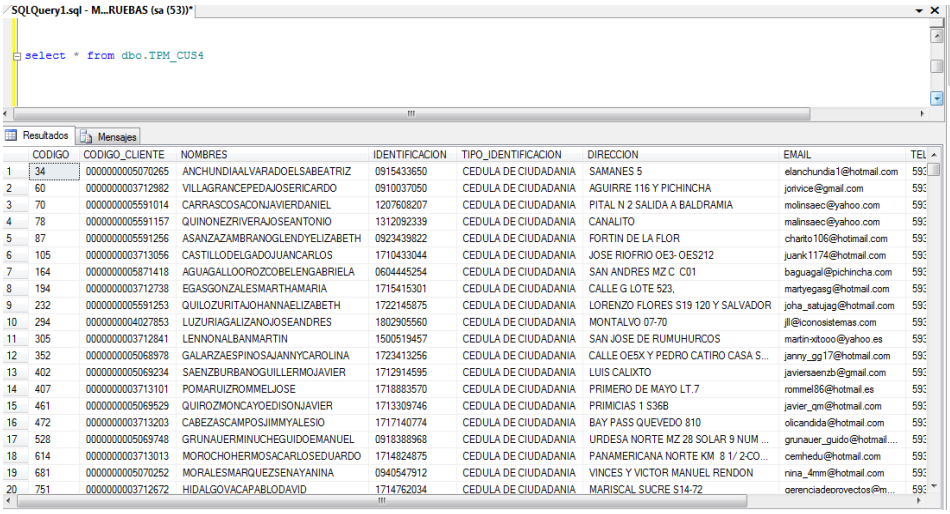
Interpretación de resultados para la tabla TPM_TRAN.

En el análisis de los resultados después de la limpieza en la tabla TPM_TRAN observamos que los datos fueron mejorados con el uso de las funciones de limpieza que las herramientas nos brinda, de esta forma podemos evidenciar que los datos han mejorado en comparación con el análisis preliminar de la tabla.

Resultado del uso de la herramienta

- La herramienta presenta dificultades al momento de la instalación y configuración ya que esto lleva mucho tiempo y un gran conocimiento de la herramienta.
- La exploración de los datos muestra grandes resultados y su generación es muy sencilla y fácil de entender.
- Los resultados mostrados por el perfilado son excelentes y muy fáciles de entender y la limpieza de los datos se la realiza en muy corto tiempo.

Los datos se obtenidos después de la limpieza con la herramienta Oracle Data Quality.



SQLQuery1.sql - M...RUEBAS (sa (53))*

```
select * from dbo.TPM_CUS4
```

Resultados | Mensajes

	CODIGO	CODIGO_CLIENTE	NOMBRES	IDENTIFICACION	TIPO_IDENTIFICACION	DIRECCION	EMAIL	TEL
1	34	0000000005070265	ANCHUNDIAALVARADOELSABEATRIZ	0915433650	CEDULA DE CIUDADANIA	SAMANES 5	elanchundia1@hotmail.com	593
2	60	0000000003712982	VILLAGRANCEPEDAJOSERICARDO	0910037050	CEDULA DE CIUDADANIA	AGUIRRE 116 Y PICHINCHA	jorivice@gmail.com	593
3	70	0000000005591014	CARRASCOSACONJAVIERDANIEL	1207608207	CEDULA DE CIUDADANIA	PITAL N 2 SALIDA A BALDRAMIA	molinsaec@yahoo.com	593
4	78	0000000005591157	QUINONEZRIVERAJOSEANTONIO	1312092339	CEDULA DE CIUDADANIA	CANALITO	malinsaec@yahoo.com	593
5	87	0000000005591256	ASANZAZAMBRANOGLENDYELIZABETH	0923439822	CEDULA DE CIUDADANIA	FORTIN DE LA FLOR	charito106@hotmail.com	593
6	105	0000000003713056	CASTILLODELGADOJUANCARLOS	1710433044	CEDULA DE CIUDADANIA	JOSE RIOFRIO OE3- OES212	juank1174@hotmail.com	593
7	164	0000000005871418	AGUAGALLOOROZCOBELENGABRIELA	0604445254	CEDULA DE CIUDADANIA	SAN ANDRES MZ C C01	baguagal@pichincha.com	593
8	194	0000000003712738	EGASGONZALESMARTHAMARIA	1715415301	CEDULA DE CIUDADANIA	CALLE G LOTE 523,	martyegasg@hotmail.com	593
9	232	0000000005591253	QUILOZURITAJOHANNAELIZABETH	1722145875	CEDULA DE CIUDADANIA	LORENZO FLORES S19 120 Y SALVADOR	joha_satujag@hotmail.com	593
10	294	0000000004027853	LUZURIAGALIZANOJOSEANDRES	1802905560	CEDULA DE CIUDADANIA	MONTALVO 07-70	jll@iconosistemas.com	593
11	305	0000000003712841	LENNONALBANMARTIN	1500519457	CEDULA DE CIUDADANIA	SAN JOSE DE RUMUHURCOS	martin-xtooo@yahoo.es	593
12	352	0000000005068978	GALARZAESPINOSAJANNYCAROLINA	1723413256	CEDULA DE CIUDADANIA	CALLE OE5X Y PEDRO CATIRO CASA S...	janny_gg17@hotmail.com	593
13	402	0000000005069234	SAENZBURBANOGUILLERMOJAVIER	1712914595	CEDULA DE CIUDADANIA	LUIS CALIXTO	javiersaenzb@gmail.com	593
14	407	0000000003713101	POMARUIZROMMELJOSE	1718883570	CEDULA DE CIUDADANIA	PRIMERO DE MAYO LT.7	rommel86@hotmail.es	593
15	461	0000000005069529	QUIROZMONCAYOEDISONJAVIER	1713309746	CEDULA DE CIUDADANIA	PRIMICIAS 1 S36B	javier_qm@hotmail.com	593
16	472	0000000003713203	CABEZASCAMPOSJIMMYALESIO	1717140774	CEDULA DE CIUDADANIA	BAY PASS QUEVEDO 810	olicandida@hotmail.com	593
17	528	0000000005069748	GRUNAUERMINUCHEGUIDOEMANUEL	0918388968	CEDULA DE CIUDADANIA	URDESA NORTE MZ 28 SOLAR 9 NUM ...	grunauer_guido@hotmail....	593
18	614	0000000003713013	MOROCHOHERMOSACARLOSEDUARDO	1714824875	CEDULA DE CIUDADANIA	PANAMERICANA NORTE KM 8 1/ 2-CO...	cemhedu@hotmail.com	593
19	681	0000000005070252	MORALESMARQUEZSENAYANINA	0940547912	CEDULA DE CIUDADANIA	VINCES Y VICTOR MANUEL RENDON	nina_4mm@hotmail.com	593
20	751	0000000003712672	HIDALGOVACAPABLODAVID	1714762034	CEDULA DE CIUDADANIA	MARISCAL SUCRE S14-72	gerenciadeproyectos@m...	593

Figura 96. Datos después de la limpieza
Fuente: Investigador

INFORMATICA

Para iniciar con la prueba de uso de la herramienta se debe configurar las conexiones con los datos que se van analizar.

Se empieza con la configuración del servidor, para ello: abrir la consola de administración de este, seleccione “Informatica Administrator Home Page” de la carpeta que se creó en el menú de inicio al momento de la instalación.

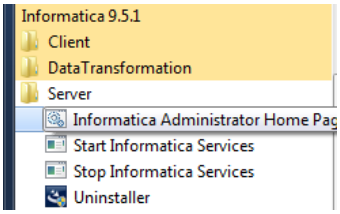


Figura 97. Informatica Administrator Home Page
Fuente: Investigador

La página se abrirá en el navegador predeterminado de nuestro sistema.

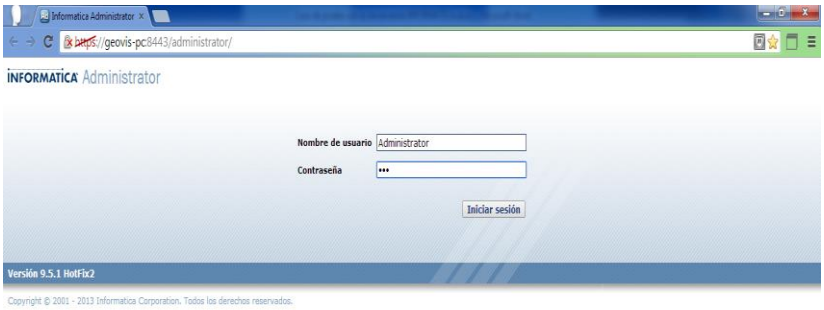


Figura 98. Pantalla de acceso

Fuente: Investigador

Se inicia sesión con nuestras credenciales y se despliega la pantalla de inicio de la herramienta.

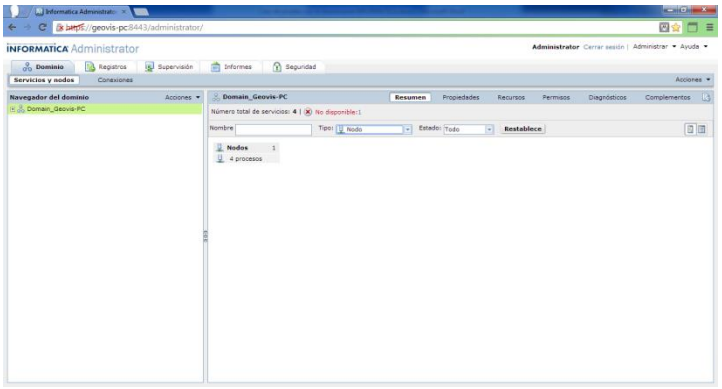


Figura 99. Pantalla de inicio
Fuente: Investigador

Se procede con la creación del servicio de conexión con el repositorio, seleccione el dominio y presione en la opción de “Acciones”, luego se crea un nuevo servicio de repositorio de modelo de datos.

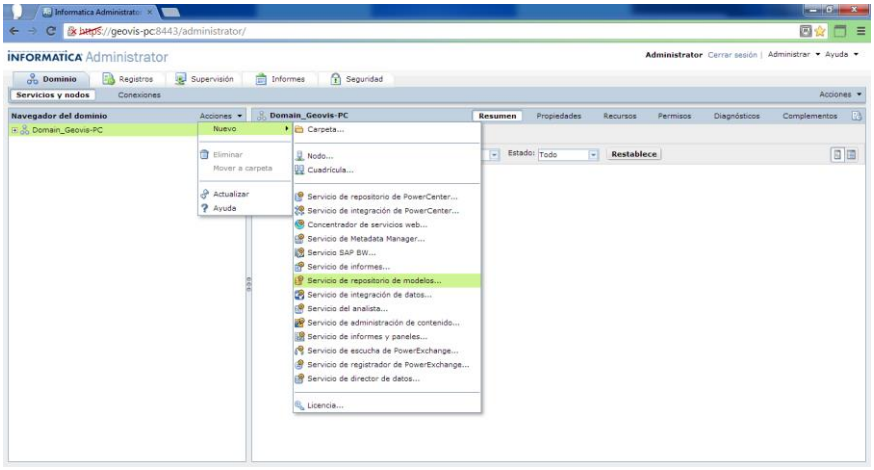


Figura 100. Acciones
Fuente: Investigador

Luego nos aparecerá una ventana emergente en la cual se debe llenar los datos del servicio que va a crear:

- Nombre
- Descripción
- Ubicación
- Licencia
- Nodo

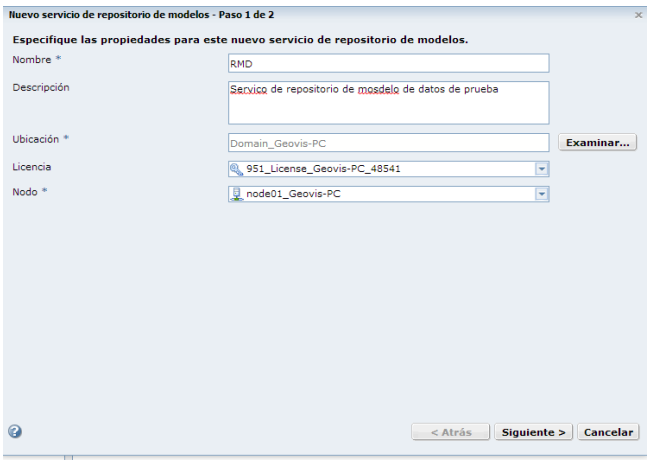


Figura 101. Datos del servicio
Fuente: Investigador

Presione siguiente para continuar con la configuración.

- Tipo de base de datos: seleccionamos SQLSERVER
- Nombre de usuario
- Contraseña
- Cadena de conexión

Esquema de la base datos: podemos dejar en blanco.

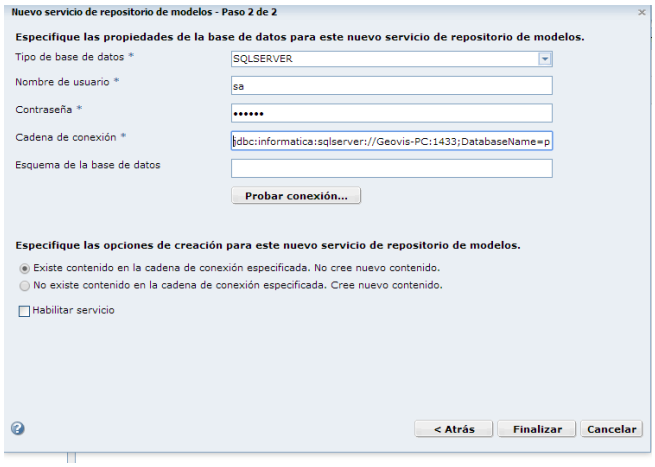


Figura 102. Esquema de la base
Fuente: Investigador

Antes de finalizar la creación del servicio probe la conexión a la base de datos

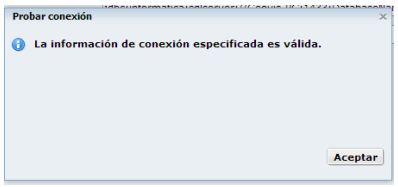


Figura 103. Creación del servicio
Fuente: Investigador

Una vez probada la conexión se finaliza con la creación del servicio

Si todo se configuró bien el servicio debe haber iniciado y estar disponible.

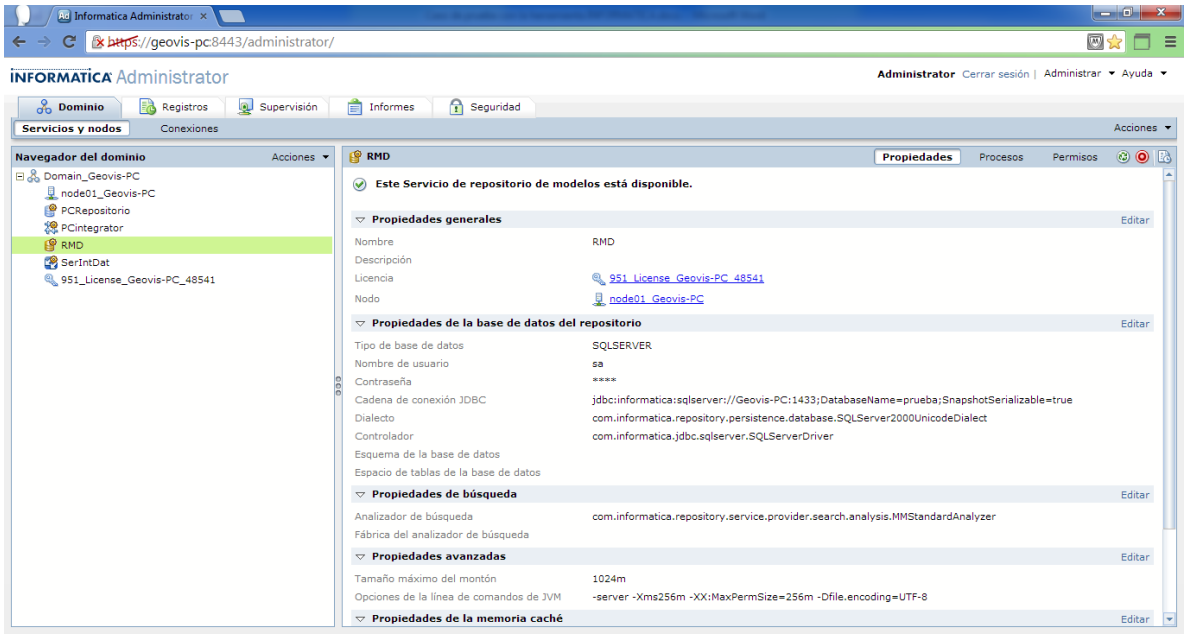


Figura 104. Servicio disponible
Fuente: Investigador

Ahora vamos a configurar el servicio de Integración de datos el cual nos permitirá ver los datos del servicio de repositorio de datos.

Para ello volvemos a seleccionar el dominio y presionar “Acciones” y creamos un nuevo servicio de integración de datos

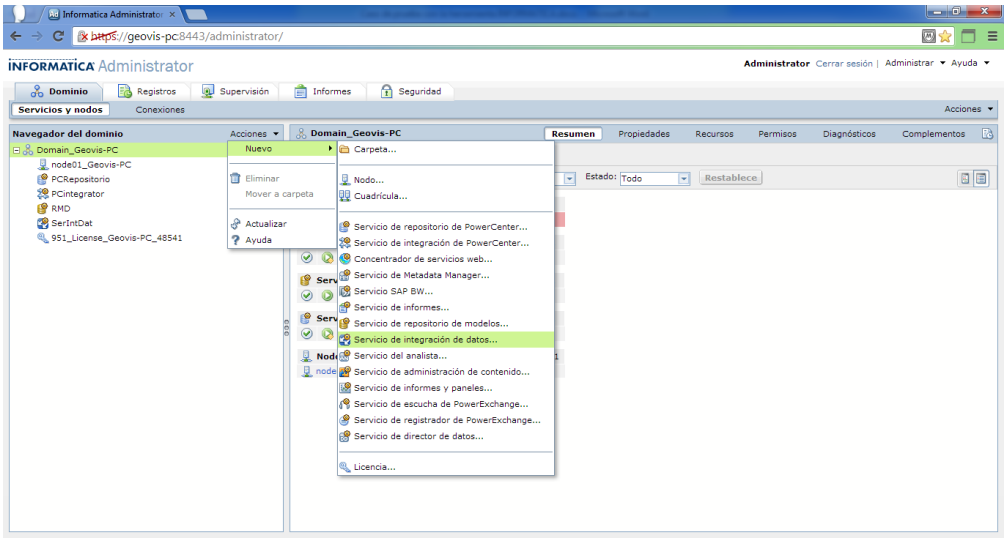


Figura 105. Acciones
Fuente: Investigador

Se abre una nueva ventana donde se configura las preferencias de creación del servicio

- Nombre
- Descripción
- Ubicación
- Asignar
- Nodo
- Servicio de repositorio de modelos
- Usuario
- Contraseña

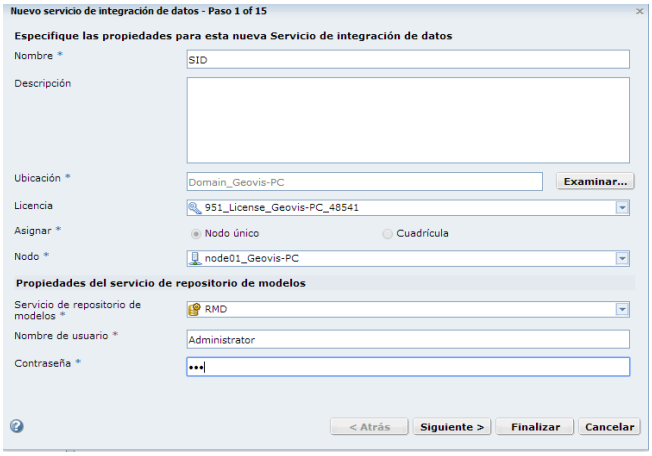


Figura 106. Datos del servicio
Fuente: Investigador

Se presiona siguiente para continuar con la configuración

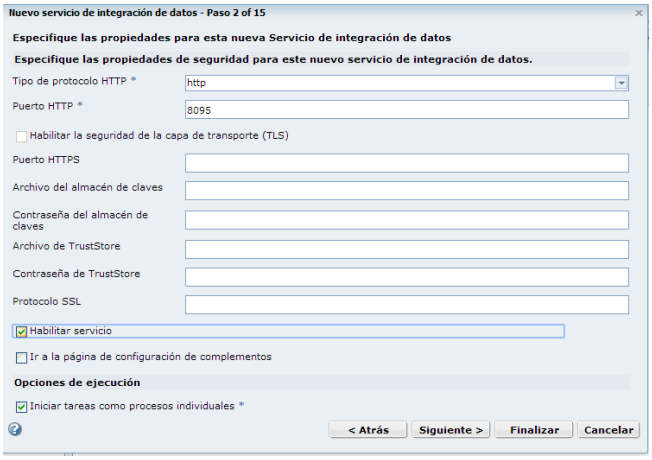


Figura 107. Configuración
Fuente: Investigador

Finalizar la configuración para continuar

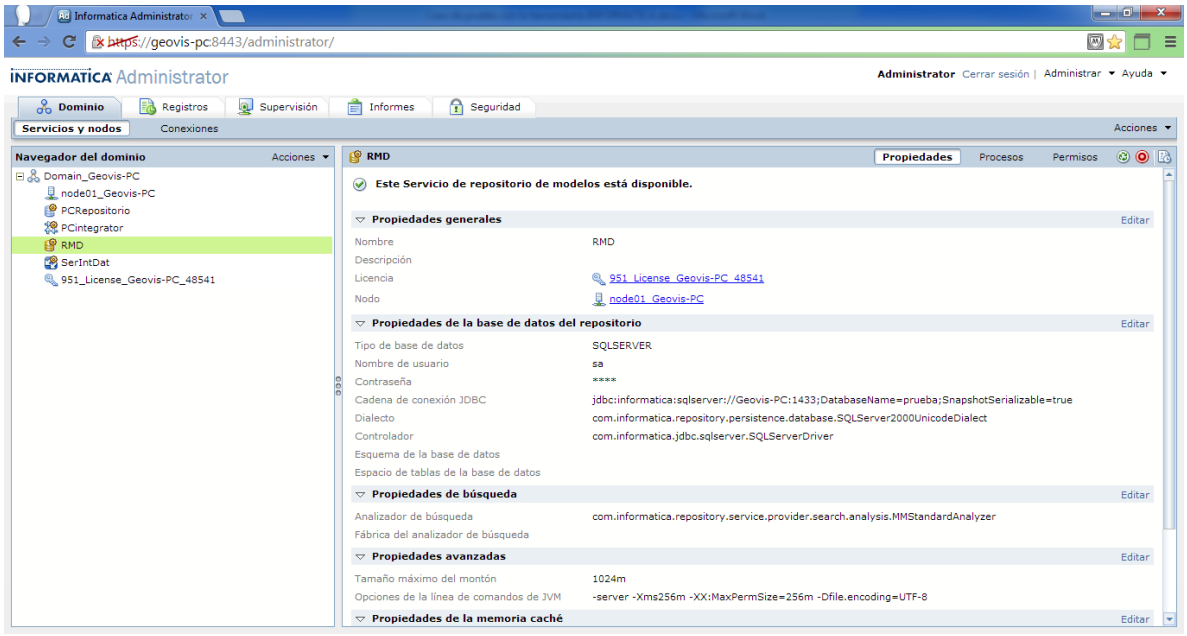


Figura 108. Finalizar Configuración
Fuente: Investigador

Una vez configurado el servidor para conectarse con el repositorio de modelo de datos se conecta con la herramienta de desarrollo de INFORMATICA DEVELOPER.

Primero se debe configurar el repositorio al cual se desea conectar para ello seleccione del menú principal “Archivo” y del sub menú la opción “Conectarse a Repositorio”

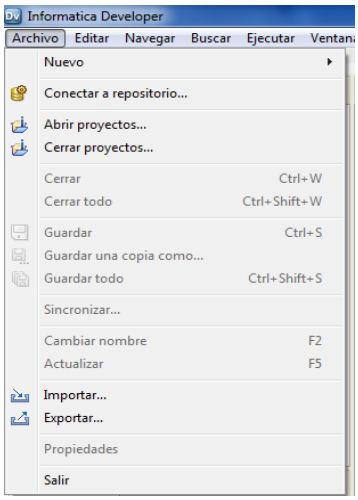


Figura 109. Archivo
Fuente: Investigador

Se abrirá una pantalla emergente que permite la selección de los objetos que se encuentran en nuestro dominio antes configurado.

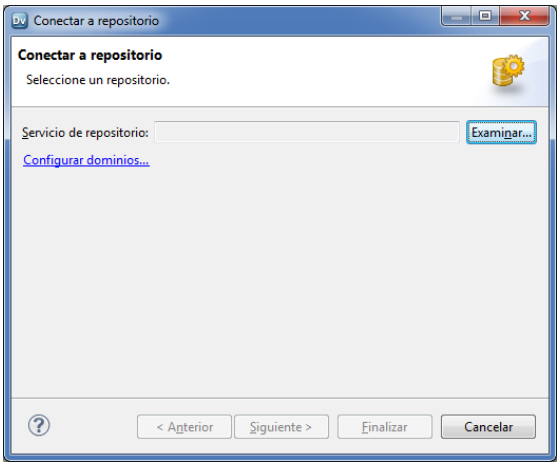


Figura 110. Pantalla emergente
Fuente: Investigador

Si no se ha configurado un dominio aun, se selecciona el link “Configurar dominios” el cual despliega una nueva ventana, añadir uno nuevo con los valores establecidos en la instalación del servidor.

- Nombre del dominio
- Nombre del host
- Puerto

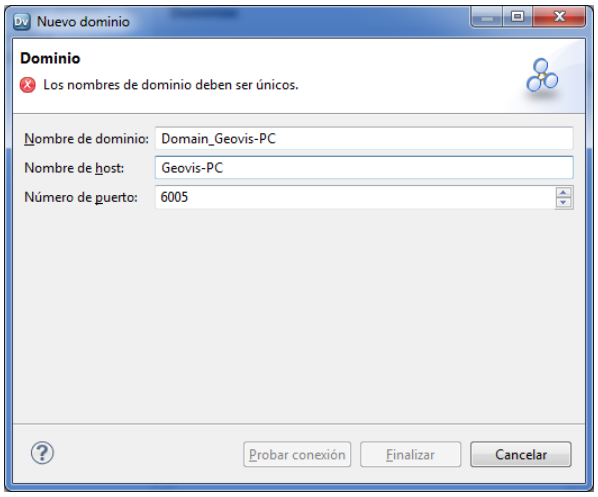


Figura 111. Configurar dominios
Fuente: Investigador

Una vez añadido el dominio seleccione y pruebe la conexión a este.

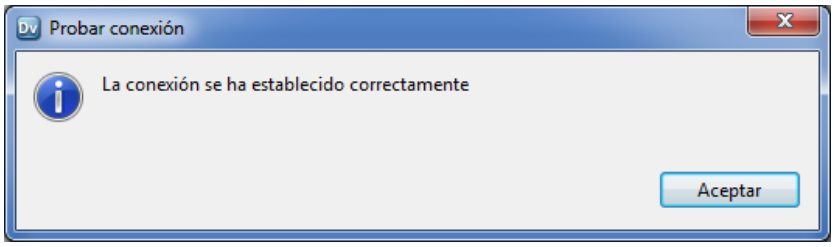


Figura 112. Mensaje
Fuente: Investigador

Ya hecha y probada la conexión al dominio regrese a la ventana de conexión con el repositorio en el cual se puede seleccionar el repositorio al cual se desee que la herramienta se conecte.

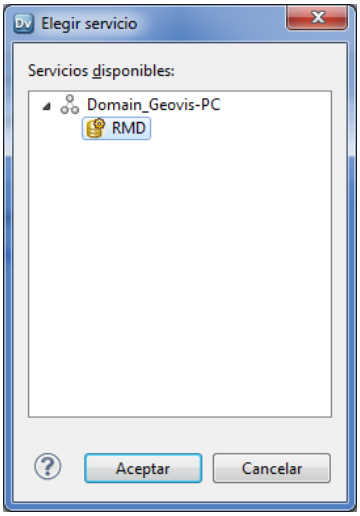


Figura 113. Ventana de conexión
Fuente: Investigador

Al seleccionar el servicio adecuado se presiona aceptar

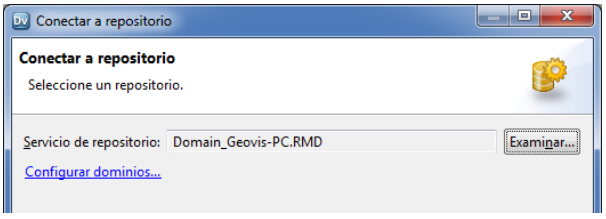


Figura 114. Seleccionar servicio
Fuente: Investigador

Para finalizar la herramienta preguntara las credenciales con las cuales desea ingresar al servicio de repositorio de datos

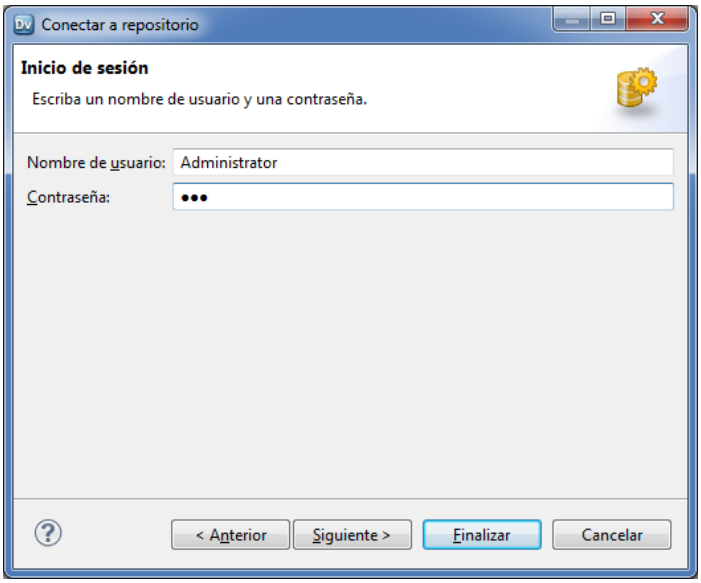


Figura 115. Inicio de Sesión
Fuente: Investigador

La conexión al repositorio se ha establecido y muestra los objetos creados.

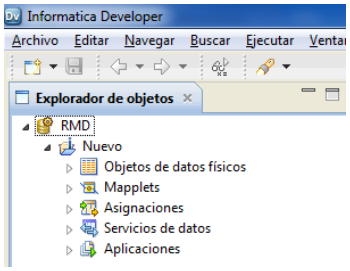


Figura 116. Objetos del repositorio
Fuente: Investigador

Se crea un nuevo proyecto dentro del repositorio para ello seleccione el repositorio y damos un clic izquierdo para que se muestre las opciones de creación

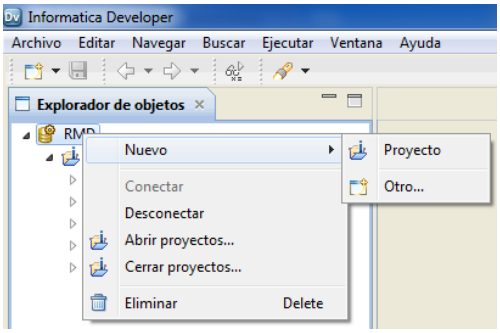


Figura 117. Nuevo proyecto
Fuente: Investigador

Se crea un nuevo proyecto

- Nombre
- Servicio de repositorio

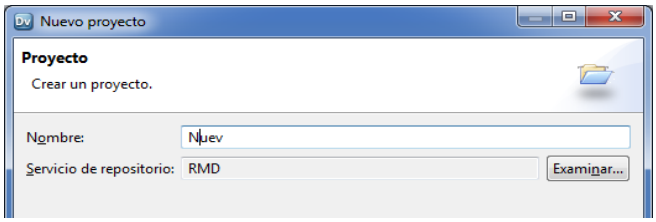


Figura 118. Nombre del proyecto
Fuente: Investigador

Presione siguiente para configurar los permisos del proyecto

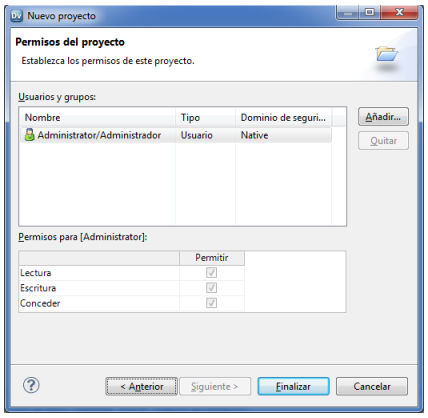


Figura 119. Permisos del proyecto
Fuente: Investigador

Para crear un proyecto de perfilado de datos se selecciona en el proyecto creado el servicio de repositorio anterior y presionamos nuevo “Perfil”

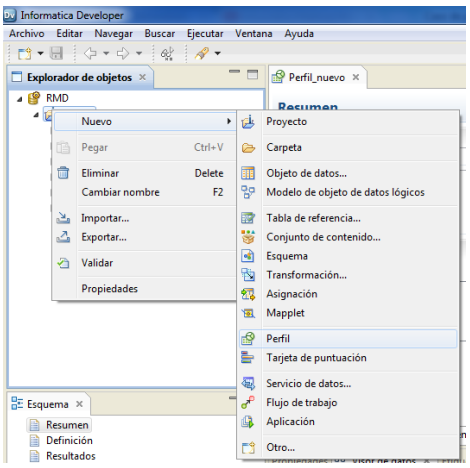


Figura 120. Perfil

Fuente: Investigador

El cual muestra una nueva ventana en la cual se debe configurar las preferencias.

- Nombre del perfil
- Ubicación
- Descripción
- Objetos de Datos

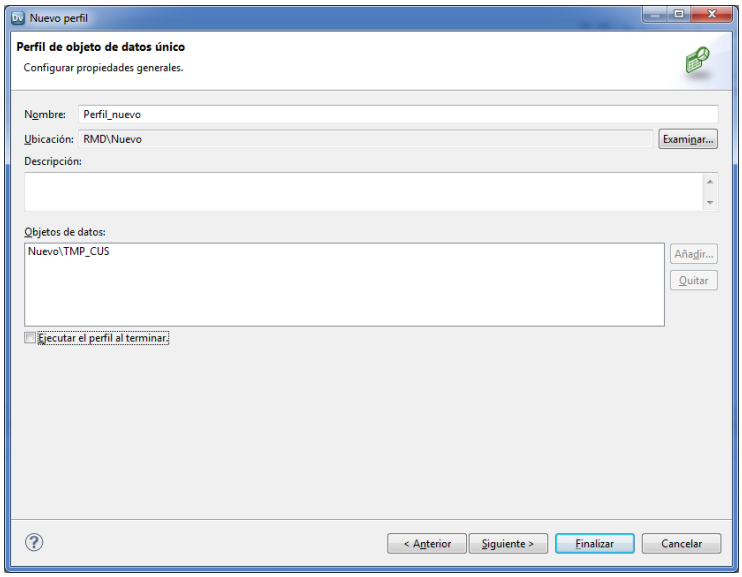


Figura 121. Ventana de Configuración

Fuente: Investigador

Siguiente para continuar, luego seleccione las preferencias de nuestro trabajo tales como las columnas para el trabajo, si se desea filtrar los datos, etc.

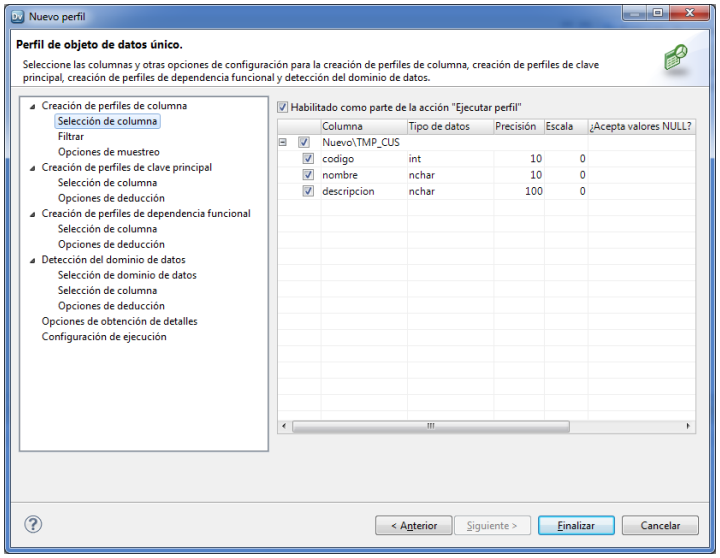


Figura 122. Ventana II de configuración
Fuente: Investigador

Una vez creado se muestra la pantalla de resumen.

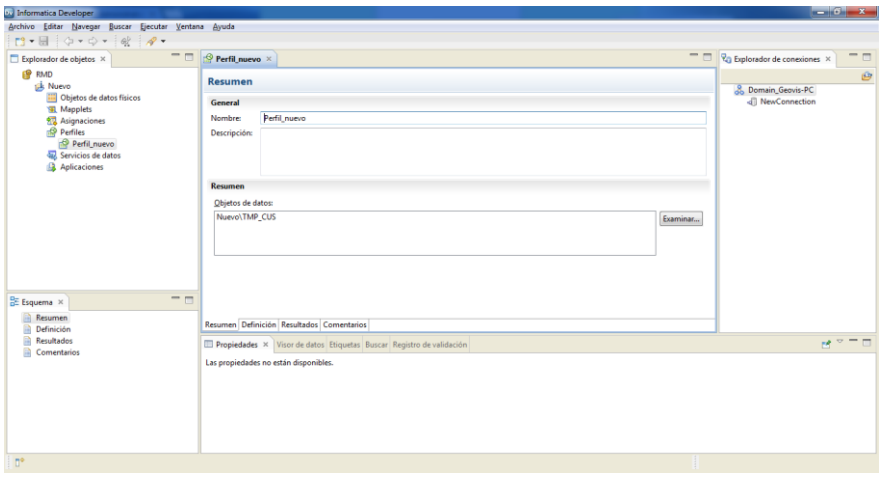


Figura 123. Pantalla de resumen
Fuente: Investigador

Para ejecutar el trabajo se selecciona la pestaña Resultados de la parte inferior del resumen y elegimos el botón de ejecutar

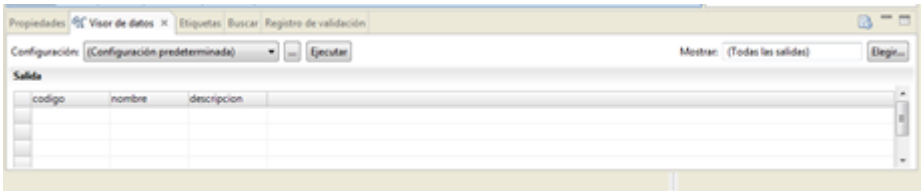


Figura 124. Pestaña de Resultados
Fuente: Investigador

En la misma ubicación se encuentra los resultados del trabajo.

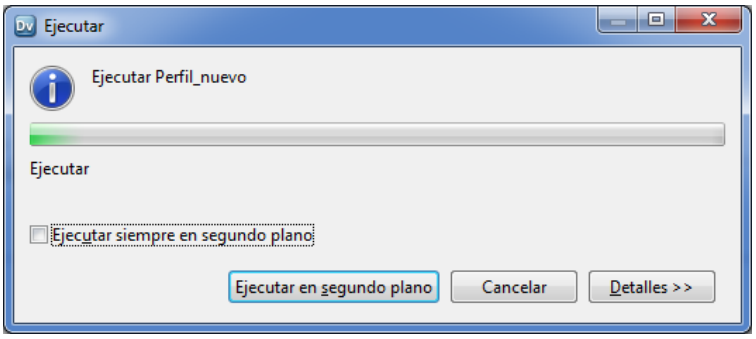


Figura 125. Ejecutar el perfil
Fuente: Investigador

Los resultados obtenidos de la herramienta Informatica Data Quality se muestra a continuación:

Tabla: TPM_CUS

Los valores nulos.

Tabla 25. Valores nulos de la tabla TPM_CUS
Fuente: Investigador

TPM_CUS	
COLUMNA	**VALORES NULOS**
CODIGO	0
CODIGO_CLIENTE	0
NOMBRES	0
IDENTIFICACION	0
TIPO_IDENTIFICACION	0
DIRECCION	2652
EMAIL	95118
TELEFONO	382
CELULAR	39794

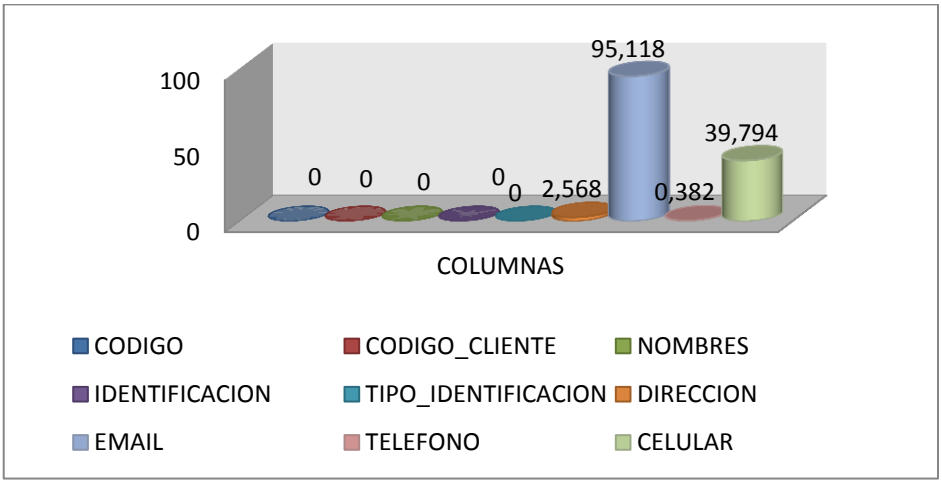


Figura 126. Valores nulos de la tabla TPM_CUS
Fuente: Investigador

Interpretación de resultados para valores nulos.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan valores nulos a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Caracteres Inválidos.

Tabla 26. Caracteres Inválidos de la tabla TPM_CUS
Fuente: Investigador

TPM_CUS	
COLUMNA	**CARACTERES INVÁLIDOS**
CODIGO	0
CODIGO_CLIENTE	0
NOMBRES	0
IDENTIFICACION	73
TIPO_IDENTIFICACION	0
DIRECCION	0
EMAIL	0
TELEFONO	2039
CELULAR	0

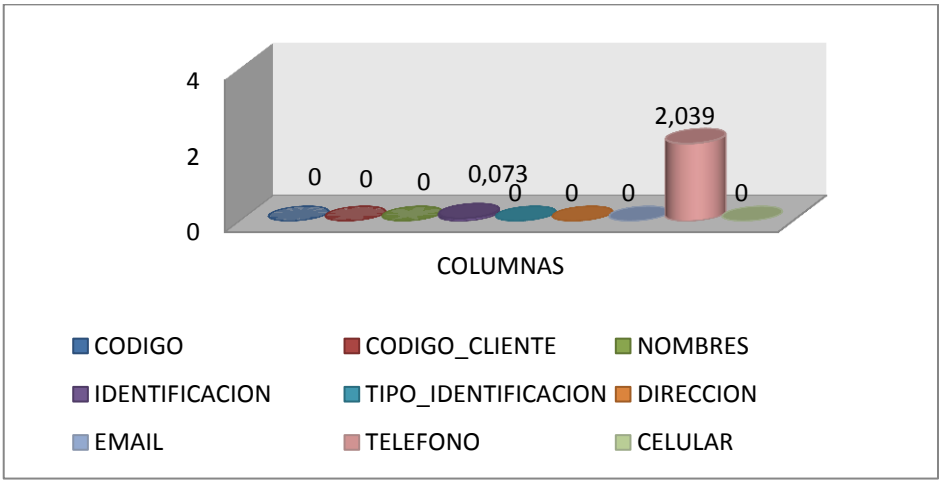


Figura 127. Caracteres Inválidos de la tabla TPM_CUS
Fuente: Investigador

Interpretación de resultados para caracteres inválidos.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan caracteres inválidos a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Registros duplicados.

Tabla 27. Registros duplicados de la tabla TPM_CUS
Fuente: Investigador

TPM_CUS	
COLUMNA	**DUCPLICADOS**
CODIGO	0
CODIGO_CLIENTE	18015
NOMBRES	18015
IDENTIFICACION	18015
TIPO_IDENTIFICACION	18015
DIRECCION	9250
EMAIL	5032
TELEFONO	1355
CELULAR	576

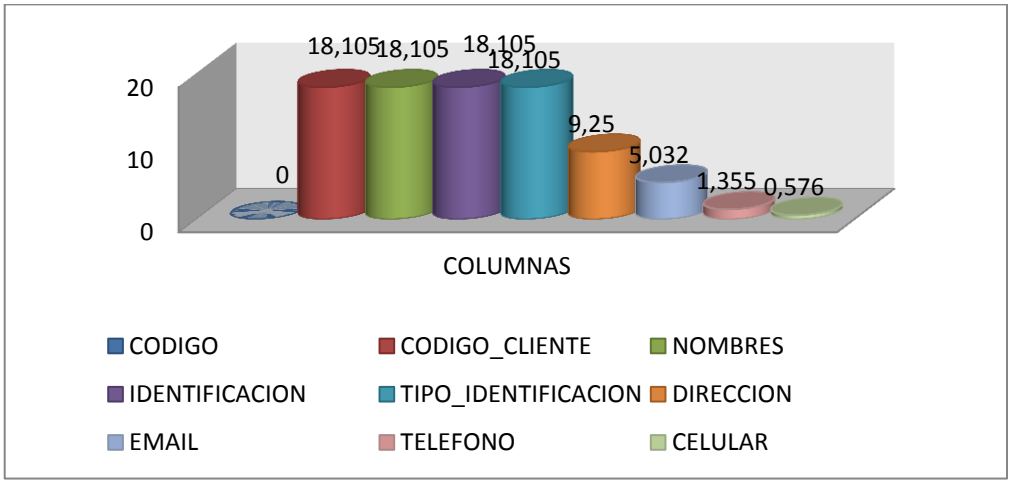


Figura 128. Registros duplicados de la tabla TPM_CUS
Fuente: Investigador

Interpretación de resultados para duplicados.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan registros duplicados a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Total de registros duplicados.

Tabla 28. Total Registros
Fuente: Investigador

TOTAL REGISTRO DUPLICADOS	
REGISTROS	46492

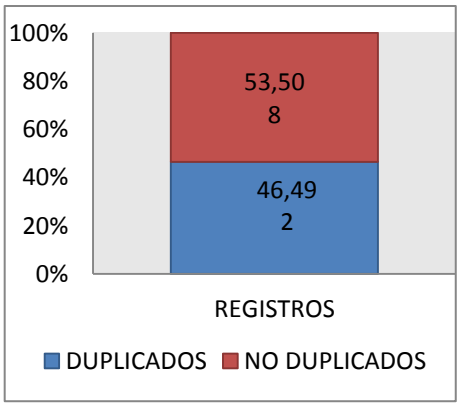


Figura 129. Total Registros
Fuente: Investigador

Interpretación de resultados para total de duplicados.

Como se visualizar en la figura anterior existe un gran número de registros duplicados para la tabla TPM_CUS los cuales deben ser solventados para obtener una mejor calidad de datos en esta tabla.

Tabla: TPM_TRAN

Valores Nulos.

Tabla 29. Valores Nulos
Fuente: Investigador

TPM_TRAN	
COLUMNA	**VALORES NULOS**
CODIGO	0
CODIGO_CLIENTE	0
INL_AMT	81503

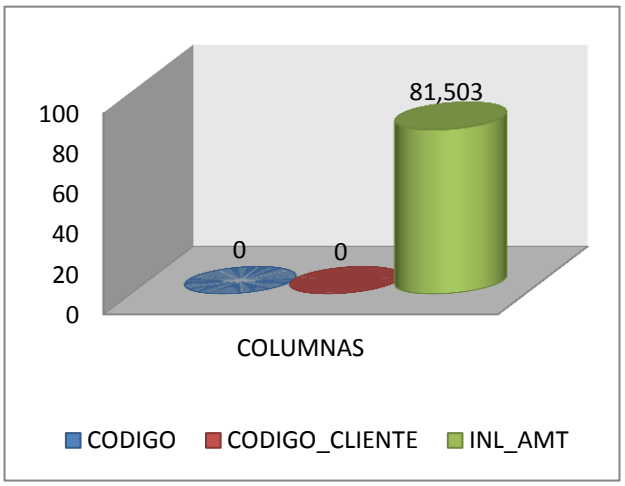


Figura 130. Valores Nulos
Fuente: Investigador

Interpretación de resultados para valores nulos.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan valores nulos a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Caracteres Inválidos.

No se presentaron valores invalidos en esta tabla.

Interpretación de resultados para caracteres inválidos.

Al realizar el análisis de los datos para esta tabla se evidencio que no existían valores inválidos dentro de los campos de la tabla TPM_TRAN.

Registros duplicados.

Tabla 30. Registros Duplicados
Fuente: Investigador

TPM_TRAN	
COLUMNA	**DUCPLICADOS**
CODIGO	0
CODIGO_CLIENTE	8899
INL_AMT	2024

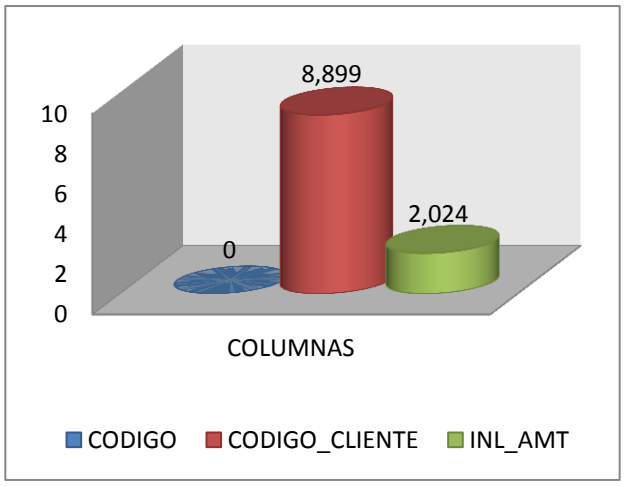


Figura 131. Registros Duplicados
Fuente: Investigador

Interpretación de resultados para duplicados.

Como se visualizar en la figura anterior existen algunas columnas las cuales presentan registros duplicados a los cuales se les dará su tratamiento de limpieza en instancias posteriores.

Total de registros duplicados.

Tabla 31. Total Registros Duplicados
Fuente: Investigador

TOTAL REGISTRO DUPLICADOS	
REGISTROS	72775

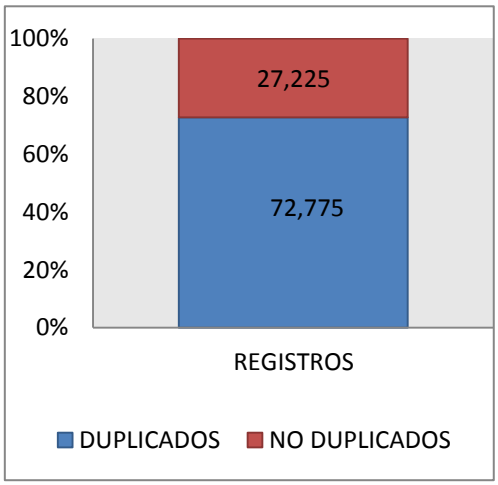


Figura 132. Total Registros Duplicados
Fuente: Investigador

Interpretación de resultados para total de duplicados.

Como se visualizar en la figura anterior existe un gran número de registros duplicados para la tabla TPM_TRAN los cuales deben ser solventados para obtener una mejor calidad de datos en esta tabla.

LIMPIEZA
Los pasos que se siguen son los siguientes:

4. Se crea un nueva carpeta de proyectos
5. Dentro de la carpeta creada se arrastra todos los componentes necesarios para la limpieza de los datos.
6. Se selecciona todos los componentes de transformación y limpieza necesarios y ejecutamos la limpieza.

Se utiliza los objetos que la herramienta que provee para la limpieza

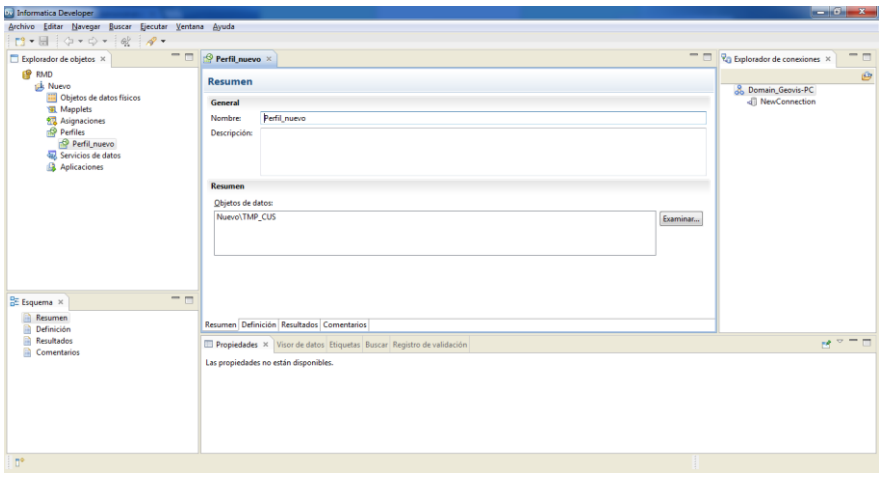


Figura 133. Limpieza
Fuente: Investigador

Los resultados obtenidos de la herramienta Informatica Data Quality después de la limpieza se muestran a continuación:

Tabla: TPM_CUS

Los valores nulos.

No se encontraron valores nulos en esta tabla.

Caracteres Inválidos.

No se encontraron valores inválidos en esta tabla.

Registros duplicados.

No se encontraron valores duplicados en esta tabla.

Interpretación de resultados para la tabla TPM_CUS.

En el análisis de los resultados después de la limpieza en la tabla TPM_CUS observamos que los datos fueron mejorados con el uso de las funciones de limpieza que las herramientas nos brinda, de esta forma podemos evidenciar que los datos han mejorado en comparación con el análisis preliminar de la tabla.

Tabla: TPM_TRAN

Valores Nulos.

No se encontraron valores nulos en esta tabla.

Caracteres Inválidos.

No se presentaron valores invalidos en esta tabla.

Registros duplicados.

No se presentaron valores duplicados en esta tabla.

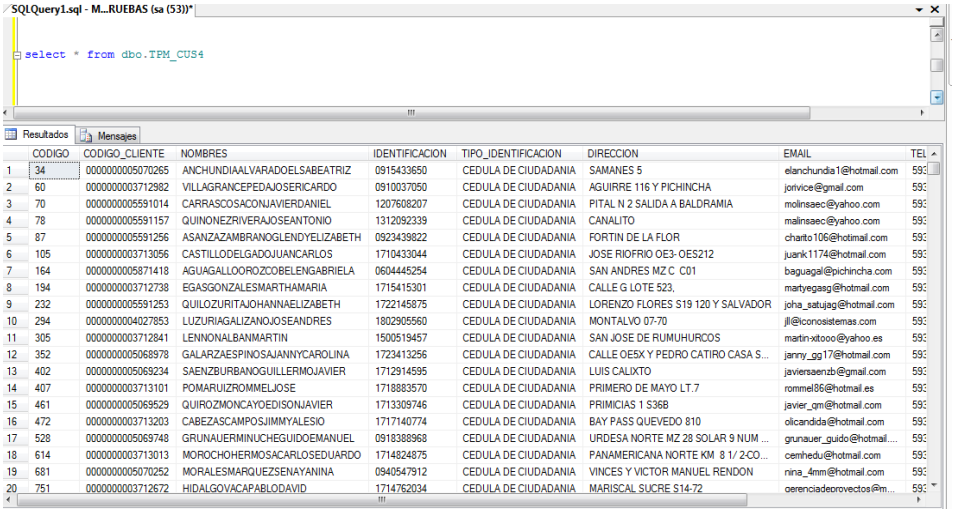
Interpretación de resultados para la tabla TPM_TRAN.

En el análisis de los resultados después de la limpieza en la tabla TPM_TRAN observamos que los datos fueron mejorados con el uso de las funciones de limpieza que las herramientas nos brinda, de esta forma podemos evidenciar que los datos han mejorado en comparación con el análisis preliminar de la tabla.

Resultado del uso de la herramienta

- La herramienta presenta mucha dificultad al momento de la instalación y configuración. Además que se debe tener un alto conocimiento de la utilización de esta.
- Los resultados del perfilado son muy sencillos de interpretar y de fácil comprensión.
- Para la limpieza de los datos se debe conocer bien los componentes que se piensan utilizar porque esto llevara a conseguir mejores resultados.

Los datos se obtenidos después de la limpieza con la herramienta Informatica Data Quality.



	CODIGO	CODIGO_CLIENTE	NOMBRES	IDENTIFICACION	TIPO_IDENTIFICACION	DIRECCION	EMAIL	TEL
1	34	0000000005070265	ANCHUNDIAALVARADOELSABEATRIZ	0915433650	CEDULA DE CIUDADANIA	SAMANES 5	elanchundia1@hotmail.com	593
2	60	0000000003712982	VILLAGRANCEPEDAJOSERICARDO	0910037050	CEDULA DE CIUDADANIA	AGUIRRE 116 Y PICHINCHA	jorivice@gmail.com	593
3	70	0000000005591014	CARRASCOSACONJAVIERDANIEL	1207608207	CEDULA DE CIUDADANIA	PITAL N 2 SALIDA A BALDRAMIA	molinsaec@yahoo.com	593
4	78	0000000005591157	QUINONEZRIVERAJOSEANTONIO	1312092339	CEDULA DE CIUDADANIA	CANALITO	malinsaec@yahoo.com	593
5	87	0000000005591256	ASANZAZAMBRANOGLENDYELIZABETH	0923439822	CEDULA DE CIUDADANIA	FORTIN DE LA FLOR	charito106@hotimail.com	593
6	105	0000000003713056	CASTILLODELGADOJUANCARLOS	1710433044	CEDULA DE CIUDADANIA	JOSE RIOFRIO OE3- OES212	juank1174@hotmail.com	593
7	164	0000000005871418	AGUAGALLOOROZCOBELENGABRIELA	0604445254	CEDULA DE CIUDADANIA	SAN ANDRES MZ C C01	baguagal@pichincha.com	593
8	194	0000000003712738	EGASGONZALESMARTHAMARIA	1715415301	CEDULA DE CIUDADANIA	CALLE G LOTE 523,	martyegasg@hotmail.com	593
9	232	0000000005591253	QUILOZURITAJOHANNAELIZABETH	1722145875	CEDULA DE CIUDADANIA	LORENZO FLORES S19 120 Y SALVADOR	joha_satujag@hotmail.com	593
10	294	0000000004027853	LUZURIAGALIZANOJOSEANDRES	1802905560	CEDULA DE CIUDADANIA	MONTALVO 07-70	jll@iconosistemas.com	593
11	305	0000000003712841	LENNONALBANMARTIN	1500519457	CEDULA DE CIUDADANIA	SAN JOSE DE RUMUHURCOS	martin-xitooo@yahoo.es	593
12	352	0000000005068978	GALARZAESPINOSAJANNYCAROLINA	1723413256	CEDULA DE CIUDADANIA	CALLE OE5X Y PEDRO CATIRO CASA S...	janny_gg17@hotmail.com	593
13	402	0000000005069234	SAENZBURBANOGUILLERMOJAVIER	1712914595	CEDULA DE CIUDADANIA	LUIS CALIXTO	javiersaenzb@gmail.com	593
14	407	0000000003713101	POMARUIZROMMELJOSE	1718883570	CEDULA DE CIUDADANIA	PRIMERO DE MAYO LT.7	rommel86@hotmail.es	593
15	461	0000000005069529	QUIROZMONCAYOEDISONJAVIER	1713309746	CEDULA DE CIUDADANIA	PRIMICIAS 1 S36B	javier_qm@hotmail.com	593
16	472	0000000003713203	CABEZASCAMPOSJIMMYALESIO	1717140774	CEDULA DE CIUDADANIA	BAY PASS QUEVEDO 810	olicandida@hotmail.com	593
17	528	0000000005069748	GRUNAUERMINUCHEGUIDOEMANUEL	0918388968	CEDULA DE CIUDADANIA	URDESA NORTE MZ 28 SOLAR 9 NUM ...	grunauer_guido@hotmail....	593
18	614	0000000003713013	MOROCHOHERMOSACARLOSEDUARDO	1714824875	CEDULA DE CIUDADANIA	PANAMERICANA NORTE KM 8 1/ 2-CO...	cemhedu@hotmail.com	593
19	681	0000000005070252	MORALESMARQUEZSENAYANINA	0940547912	CEDULA DE CIUDADANIA	VINCES Y VICTOR MANUEL RENDON	nina_4mm@hotmail.com	593
20	751	0000000003712672	HIDALGOVACAPABLODAVID	1714762034	CEDULA DE CIUDADANIA	MARISCAL SUCRE S14-72	gerenciadeproyectos@m...	593

Figura 134. Datos Después De La Limpieza Herramienta Informática
Fuente: Investigador

ANEXO II

LIMPIEZA DE LOS DATOS UTILIZANDO LA HERRAMIENTA INFORMATICA DATA QUALITY

CONEXION CON LA BASE DE DATOS OASIS

Para la conexión con la base de datos OASIS se debe configurar el servicio de repositorio de modelo de datos y el servicio de integración de datos en el servidor.

Repositorio de Modelo de Datos.

Paso 1. Propiedades del repositorio

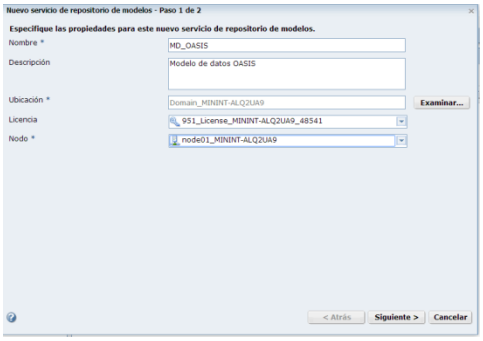


Figura 135. Propiedades del repositorio
Fuente: Investigador

Paso 2. Propiedades del repositorio y drivers

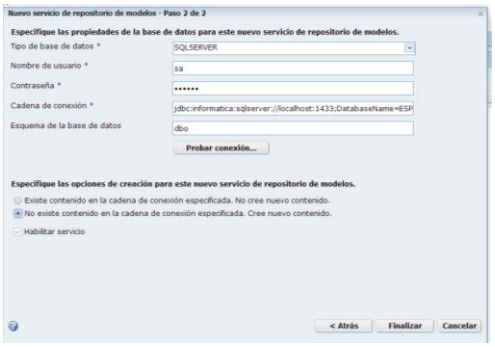


Figura 136. Propiedades del repositorio y drivers
Fuente: Investigador

Integración de Datos

Paso 1. Propiedades del servicio de integración de datos

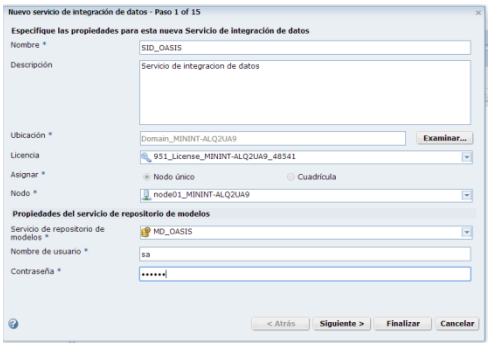


Figura 137. Propiedades del servicio
Fuente: Investigador

Paso 2. Propiedades de seguridad.

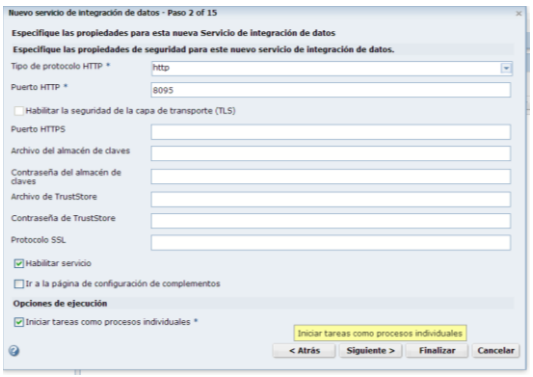


Figura 138. Propiedades de seguridad
Fuente: Investigador

Una vez configurados los servicios se utilizó la herramienta cliente para el paso de limpieza.

Figura 139. Herramienta: Informática
Fuente: Investigador

LIMPIEZA

Conexión con el repositorio desde el cliente.

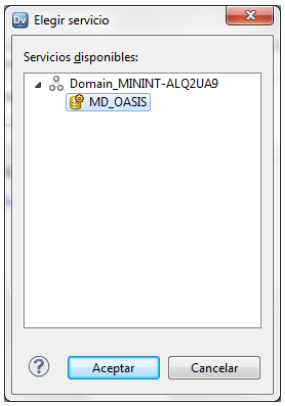


Figura 140. Conexión
Fuente: Investigador

Iniciar sesión con el usuario administrador.

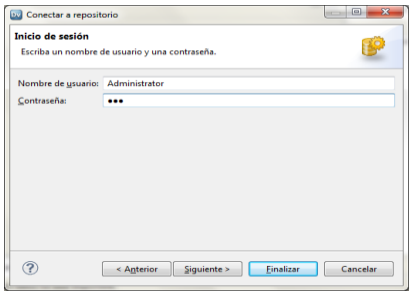


Figura 141. Inicio de sesión.
Fuente: Investigador

Se crea un nuevo proyecto para la limpieza de los datos.

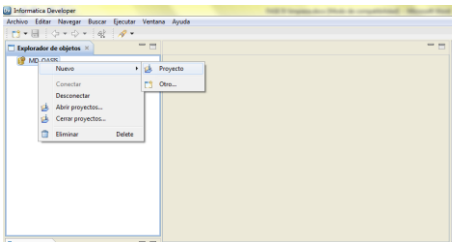


Figura 142. Crear proyecto
Fuente: Investigador

Se coloca el nombre y las propiedades de conexión

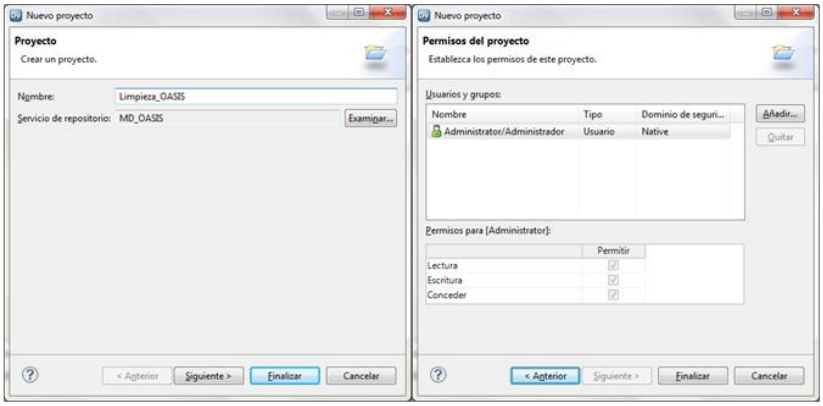


Figura 143. Nombre y propiedades de conexión
Fuente: Investigador

Se crea la conexión con la base de datos para traer las tablas.

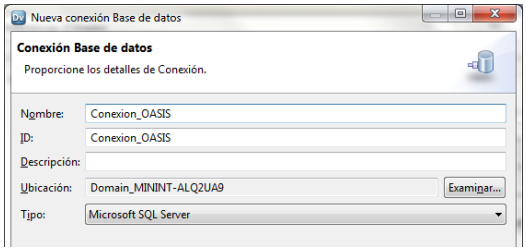


Figura 144. Conexión con la base de datos
Fuente: Investigador

Configuramos los detalles de la conexión.

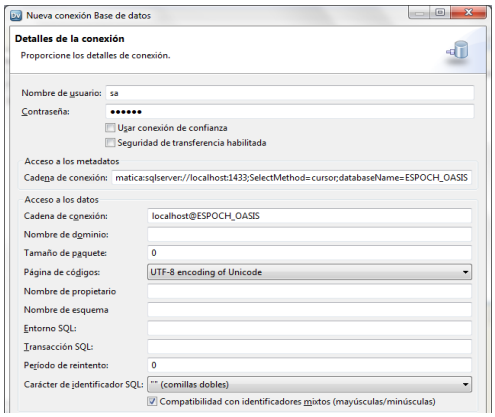


Figura 145. Configuración
Fuente: Investigador

Se secciona las tablas con las vamos a trabajar y procedemos con la limpieza.

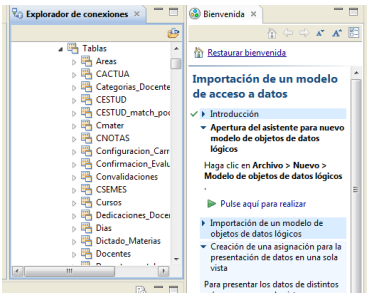


Figura 146. Selección de tablas de la Base de Datos
Fuente: Investigador

Se seleccionan los analizadores y transformadores necesarios para la limpieza

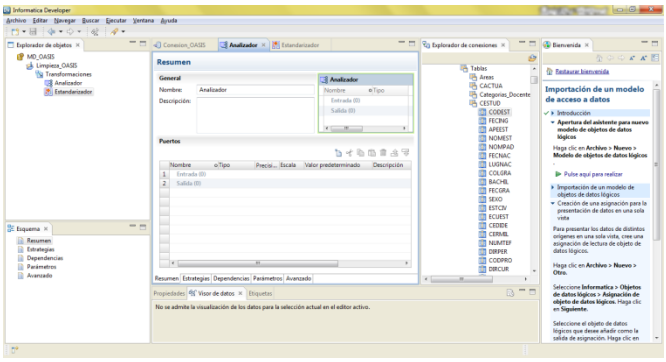


Figura 147. Analizadores y Transformadores
Fuente: Investigador

Los resultados obtenidos después de la limpieza con la herramienta se muestran a continuación, para este análisis se utilizó la herramienta DataCleaner:

- **CESTUD**

Análisis final de la tabla CESTUD después de la limpieza de los datos

Valores NULL

La tabla no contiene valores nulos en sus columnas.

Valores Inválidos

La tabla contiene los valores aceptables esperados según los requerimientos del negocio.

Valores Duplicados

La tabla no contiene valores duplicados en sus registros.

Análisis de resultados

En el análisis realizado para la tabla CESTUD en la cual observamos que las inconsistencias tales como valores nulos para algunas columnas así como valores inválidos dentro de otras columnas y los registros duplicados que habíamos encontrado en el análisis preliminar se han solventado con la limpieza de los datos utilizando la herramienta Informatica Data Quality dejando la tabla de la mejor forma posible.

- **ESTUDIANTES**

Análisis final de la tabla Estudiantes después de la limpieza de los datos

Valores NULL

La tabla no contiene valores nulos en sus columnas.

Valores Inválidos

La tabla contiene los valores aceptables esperados según los requerimientos del negocio.

Valores Duplicados

No se encontraron valores duplicados para esta tabla.

Análisis de resultados

En el análisis realizado para la tabla Estudiantes en la cual observamos que las inconsistencias tales como valores nulos para algunas columnas así como valores inválidos dentro de otras columnas y los registros duplicados que habíamos encontrado en el análisis preliminar se han solventado con la limpieza de los datos utilizando la herramienta Informatica Data Quality dejando la tabla de la mejor forma posible.

- **DOCENTES**

Análisis final de la tabla Docentes después de la limpieza de los datos

Valores NULL

La tabla no contiene valores nulos en sus columnas.

Valores Inválidos

La tabla contiene los valores aceptables esperados según los requerimientos del negocio.

Valores Duplicados

No se encontraron valores duplicados para esta tabla.

Análisis de resultados

En el análisis realizado para la tabla Docentes en la cual observamos que las inconsistencias tales como valores nulos para algunas columnas así como valores inválidos dentro de otras

columnas y los registros duplicados que habíamos encontrado en el análisis preliminar se han solventado con la limpieza de los datos utilizando la herramienta Informatica Data Quality dejando la tabla de la mejor forma posible.

- **MATERIAS**

Análisis final de la tabla Materias después de la limpieza de los datos

Valores NULL

La tabla no contiene valores nulos en sus columnas.

Valores Inválidos

La tabla contiene los valores aceptables esperados según los requerimientos del negocio.

Valores Duplicados

No se encontraron valores duplicados para esta tabla.

Análisis de resultados

En el análisis realizado para la tabla Materias en la cual observamos que las inconsistencias tales como valores nulos para algunas columnas así como valores inválidos dentro de otras columnas y los registros duplicados que habíamos encontrado en el análisis preliminar se han solventado con la limpieza de los datos utilizando la herramienta Informatica Data Quality dejando la tabla de la mejor forma posible.

- **MATRICULAS**

Análisis final de la tabla Matriculas después de la limpieza de los datos

Valores NULL

La tabla no contiene valores nulos en sus columnas.

Valores Inválidos

La tabla contiene los valores aceptables esperados según los requerimientos del negocio.

Valores Duplicados

No se encontraron valores duplicados para esta tabla.

Análisis de resultados

En el análisis realizado para la tabla Matriculas en la cual observamos que las inconsistencias tales como valores nulos para algunas columnas así como valores inválidos dentro de otras columnas y los registros duplicados que habíamos encontrado en el análisis preliminar se han solventado con la limpieza de los datos utilizando la herramienta Informatica Data Quality dejando la tabla de la mejor forma posible.

- **EVALUACIONES**

Análisis final de la tabla Evaluaciones después de la limpieza de los datos

Valores NULL

La tabla no contiene valores nulos en sus columnas.

Valores Inválidos

La tabla contiene los valores aceptables esperados según los requerimientos del negocio.

Valores Duplicados

No se encontraron valores duplicados para esta tabla.

Análisis de resultados

En el análisis realizado para la tabla Evaluaciones en la cual observamos que las inconsistencias tales como valores nulos para algunas columnas así como valores inválidos dentro de otras columnas y los registros duplicados que habíamos encontrado en el análisis preliminar se han solventado con la limpieza de los datos utilizando la herramienta Informatica Data Quality dejando la tabla de la mejor forma posible.

- **NOTAS_EXAMENES**

Análisis final de la tabla Notas_Examenes después de la limpieza de los datos

Valores NULL

La tabla no contiene valores nulos en sus columnas.

Valores Inválidos

La tabla contiene los valores aceptables esperados según los requerimientos del negocio.

Valores Duplicados

No se encontraron valores duplicados para esta tabla.

Análisis de resultados

En el análisis realizado para la tabla Notas_Examenes en la cual observamos que las inconsistencias tales como valores nulos para algunas columnas así como valores inválidos dentro de otras columnas y los registros duplicados que habíamos encontrado en el análisis preliminar se han solventado con la limpieza de los datos utilizando la herramienta Informatica Data Quality dejando la tabla de la mejor forma posible.

- **PERIODOS**

Análisis final de la tabla Periodos después de la limpieza de los datos

Valores NULL

La tabla no contiene valores nulos en sus columnas.

Valores Inválidos

La tabla contiene los valores aceptables esperados según los requerimientos del negocio.

Valores Duplicados

No se encontraron valores duplicados para esta tabla.

Análisis de resultados

En el análisis realizado para la tabla Periodos en la cual observamos que las inconsistencias tales como valores nulos para algunas columnas así como valores inválidos dentro de otras columnas y los registros duplicados que habíamos encontrado en el análisis preliminar se han solventado con la limpieza de los datos utilizando la herramienta Informatica Data Quality dejando la tabla de la mejor forma posible.

ANÁLISIS DE LAS DIMENSIONES DE CALIDAD

Análisis de la calidad de los datos para la tabla CESTUD:

En los datos encontrados en el analisis para la tabla CESTUD despues de la limpieza se evidencio:

Parametro de Precisión:
En la tabla no se han encontrado registros nulos.

Análisis de resultados para la precisión.

En el análisis de los datos para la tabla CESTUD de la base de datos OASIS en el parámetro de Precisión se puede evidenciar que no se ha encontrado registros nulos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Valores Aceptables

En la tabla no se han encontrado registros inválidos.

Análisis de resultados para los Valores Aceptables.

En el análisis de los datos para la tabla CESTUD de la base de datos OASIS en el parámetro de Valores Aceptables se puede evidenciar que no se ha encontrado registros inválidos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Análisis de resultados para los valores válidos.

En el análisis realizado se ha evidenciado que en algunas columnas de la tabla CESTUD de la base de datos OASIS los valores ahí guardados se han mejorado con el uso de la herramienta.

Parámetro de Duplicidad

En la tabla no se han encontrado registros duplicados para esta tabla.

Análisis de resultados para la Duplicidad.

En el análisis de los datos para la tabla CESTUD de la base de datos OASIS en el parámetro de duplicidad se puede evidenciar que se ha solventado este problema con la eliminación de los registros duplicados por esa razón para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Tabla 32. Parámetro de Confianza
Fuente: Investigador

TABLA CESTUD			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
APEEST	100%	100%	100%
FECING	100%	100%	100%
NOMPAD	100%	100%	100%
LUGNAC	100%	100%	100%
COLGRA	100%	100%	100%
BACHIL	100%	100%	100%
FECGRA	100%	100%	100%
CEDIDE	100%	100%	100%
CERMIL	100%	100%	100%
NUMTEF	100%	100%	100%
DIRPER	100%	100%	100%
DIRCUR	100%	100%	100%
DOCUME	100%	100%	100%
CODEST	100%	100%	100%

Resultados finales.

Tabla 33 Resultados Finales
Fuente: Investigador

CESTUD	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

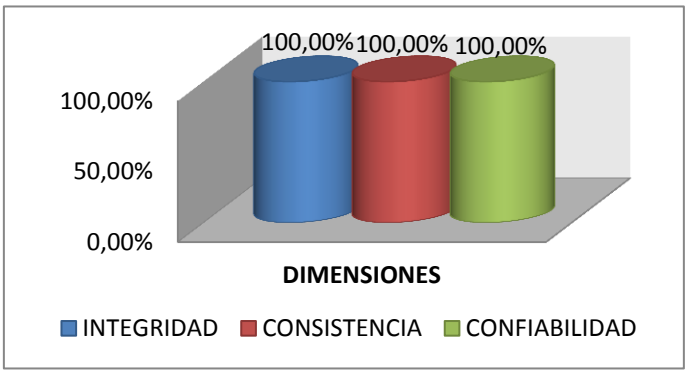


Figura 148 Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.
Como se puede observar en la tabla y en la gráfica anterior las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla CESTUD, con un nivel de integridad del 100% en comparación al análisis preliminar de los datos en el cual se obtuvo 99,03% mejorándolo en un 1,97%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación al 84,84% obtenido en el análisis preliminar mejorando a esta dimensión en un 15,16% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 93,35% obtenido en el análisis preliminar mejorando esta dimensión en 6,65% dejando a la tabla con un mejor nivel de conformidad de datos.

Análisis de la calidad de los datos para la tabla Docentes:
En los datos encontrados en el analisis para la tabla Docentes despues de la limpieza se evidencio:

Parametro de Precisión
En la tabla no se han encontrado registros nulos.

Análisis de resultados para la precisión.

En el análisis de los datos para la tabla Docentes de la base de datos OASIS en el parámetro de Precisión se puede evidenciar que no se ha encontrado registros nulos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Valores Aceptables

En la tabla no se han encontrado registros inválidos.

Análisis de resultados para los Valores Aceptables.

En el análisis de los datos para la tabla Docentes de la base de datos OASIS en el parámetro de Valores Aceptables se puede evidenciar que no se ha encontrado registros inválidos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro Duplicidad

En esta tabla no se encontraron registros duplicados.

Análisis de resultados para la Duplicidad.

En el análisis de los datos para la tabla Docentes de la base de datos OASIS en el parámetro de duplicidad se puede evidenciar que no se ha encontrado registros duplicados lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Tabla 34. Parámetro de Confianza
Fuente: Investigador

TABLA DOCENTES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
strCedulaMil	100%	100%	100%
strCarnetSeg	100%	100%	100%
strDireccion	100%	100%	100%
strTel	100%	100%	100%
strMail	100%	100%	100%
strWww	100%	100%	100%
strCodTipoSan	100%	100%	100%
strCodEstCiv	100%	100%	100%
strTitulos	100%	100%	100%
strCargos	100%	100%	100%
strCodTipTit	100%	100%	100%
strNacionalidad	100%	100%	100%
strNombres	100%	100%	100%
strApellidos	100%	100%	100%

Resultados finales.

Tabla 35. Resultados Finales
Fuente: Investigador

TABLA DOCENTES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

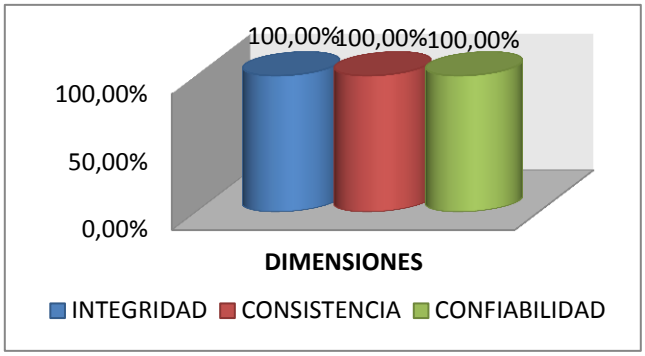


Figura 149 . Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.
Como se puede observar en la tabla y en la gráfica anterior las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla Docentes, con un nivel de integridad del 100% al igual que en el análisis preliminar de los datos en el cual se obtuvo 100%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación al 75,41% obtenido en el análisis preliminar mejorando a esta dimensión en un 24,59% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 90,16% obtenido en el análisis preliminar mejorando esta dimensión en 9,84% dejando a la tabla con un mejor nivel de conformidad de datos.

Análisis de la tabla Estudiantes:
En los datos encontrados en el analisis para la tabla Estudiantes despues de la limpieza de los datos se evidencio:

Parametro de Precisión
En la tabla no se han encontrado registros nulos.

Análisis de resultados para la precisión.

En el análisis de los datos para la tabla Estudiantes de la base de datos OASIS en el parámetro de Precisión se puede evidenciar que no se ha encontrado registros nulos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Valores Aceptables

En la tabla no se han encontrado registros inválidos.

Análisis de resultados para los Valores Aceptables.

En el análisis de los datos para la tabla Estudiantes de la base de datos OASIS en el parámetro de Valores Aceptables se puede evidenciar que no se ha encontrado registros inválidos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro Duplicidad

En esta tabla no se encontraron registros duplicados.

Análisis de resultados para la Duplicidad.

En el análisis preliminar de los datos para la tabla Estudiantes de la base de datos OASIS en el parámetro de duplicidad se puede evidenciar que no se ha encontrado registros duplicados lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Tabla 36. Parámetro de Confianza para Duplicidad
Fuente: Investigador

TABLA ESTUDIANTES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
FECING	100%	100%	100%
strMail	100%	100%	100%
strDocumentacion	100%	100%	100%
strCodTit	100%	100%	100%
strCedulaMil	100%	100%	100%
strCodInt	100%	100%	100%
strNacionalidad	100%	100%	100%
strNombres	100%	100%	100%
strApellidos	100%	100%	100%

Resultados finales.

Tabla 37. Resultados Finales
Fuente: Investigador

TABLA ESTUDIANTES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

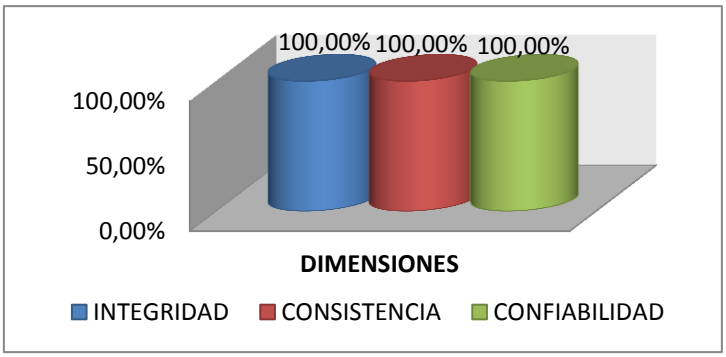


Figura 150 Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.
Como se puede observar en la tabla y en la gráfica anterior las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla Estudiantes, con un nivel de integridad del 100% al igual que en el análisis preliminar de los datos en el cual se obtuvo 100%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación al 85,76% obtenido en el análisis preliminar mejorando a esta dimensión en un 14,24% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 94,30% obtenido en el análisis preliminar mejorando esta dimensión en 5,7% dejando a la tabla con un mejor nivel de conformidad de datos.

Análisis de la calidad de datos para tabla Materias:
En los datos encontrados en el analisis preliminar para la tabla Materias se evidencio:

Parametro de Precisión:
En esta tabla no se encontraron registros con valores nulos.

Análisis de resultados para la precisión.

En el análisis de los datos para la tabla Materias de la base de datos OASIS en el parámetro de Precisión se puede evidenciar que no se ha encontrado registros nulos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Valores Aceptables

En esta tabla no se encontraron registros con valores inválidos.

Análisis de resultados para los valores válidos.

En el análisis de los datos para la tabla Materias de la base de datos OASIS en el parámetro de Valores Aceptables se puede evidenciar que no se ha encontrado registros inválidos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Duplicidad

En esta tabla no se encontraron registros duplicados para esta tabla.

Análisis de resultados para la Duplicidad.

En el análisis de los datos para la tabla Materias de la base de datos OASIS en el parámetro de Duplicidad se puede evidenciar que no se ha encontrado registros inválidos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Aquí utilizamos las formulas planteadas con anterioridad.

Tabla 38. Parámetro de Confianza
Fuente: Investigador

TABLA MATERIAS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
srtCodigo	100%	100%	100%
strNombre	100%	100%	100%
dtFechaCreada	100%	100%	100%
dtFechaElim	100%	100%	100%
blnActiva	100%	100%	100%

Resultados finales.

Tabla 39. Resultados Finales
Fuente: Investigador

TABLA MATERIAS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

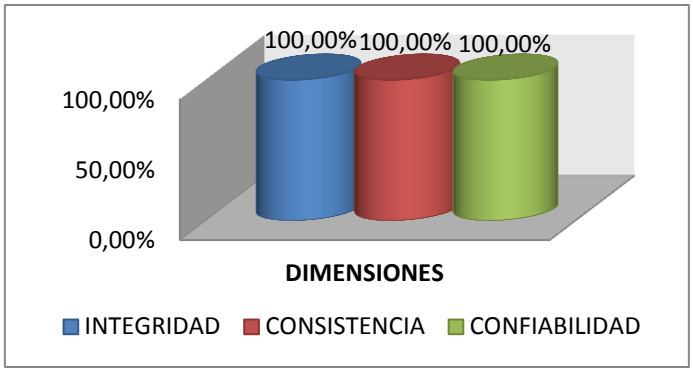


Figura 151. Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.
Como se puede observar en la tabla y en la gráfica anterior las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla Materias, con un nivel de integridad del 100% al igual que en el análisis preliminar de los datos en el cual se obtuvo 100%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación al 94,14% obtenido en el análisis preliminar mejorando a esta dimensión en un 5,86% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 97,66% obtenido en el análisis preliminar mejorando esta dimensión en 2,34% dejando a la tabla con un mejor nivel de conformidad de datos.

Análisis de la calidad de datos para la tabla Matriculas:
En los datos encontrados en el analisis preliminar para la tabla Matriculas se evidencio:

Parametro de Precisión:
En esta tabla no se encontraron registros con valores nulos.

Análisis de resultados para la precisión.

En el análisis de los datos para la tabla Matriculas de la base de datos OASIS en el parámetro de Precisión se puede evidenciar que no se ha encontrado registros nulos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Valores Aceptables

En esta tabla no se encontraron registros con valores inválidos.

Análisis de resultados para los Valores Aceptables.

En el análisis de los datos para la tabla Matriculas de la base de datos OASIS en el parámetro de Valores Aceptables se puede evidenciar que no se ha encontrado registros inválidos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Duplicidad

En esta tabla no se encontraron registros duplicados para esta tabla.

Análisis de resultados para la Duplicidad.

En el análisis de los datos para la tabla Matriculas de la base de datos OASIS en el parámetro de Duplicidad se puede evidenciar que no se ha encontrado registros inválidos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Aquí utilizamos las formulas planteadas con anterioridad.

Tabla 40. Parámetro de Confianza
Fuente: Investigador

TABLA MATRICULAS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
sintCodigo	100%	100%	100%
strCodPeriodo	100%	100%	100%
strCodEstud	100%	100%	100%
strCodNivel	100%	100%	100%
strAutorizadaPor	100%	100%	100%
dtFechaAutorizada	100%	100%	100%
strCreadaPor	100%	100%	100%
dtFechaCreada	100%	100%	100%
strCodEstado	100%	100%	100%
strObservaciones	100%	100%	100%

Resultados finales.

Tabla 41. Resultados Finales
Fuente: Investigador

TABLA MATRICULAS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

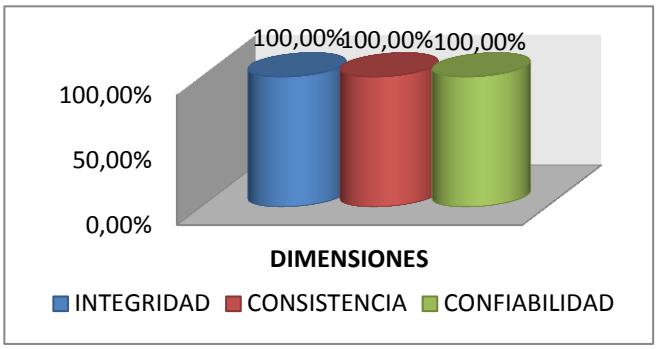


Figura 152. Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.
Como se puede observar en la tabla y en la gráfica anterior las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla Matriculas, con un nivel de integridad del 100% al igual que en el análisis preliminar de los datos en el cual se obtuvo 100%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación al 94,36% obtenido en el análisis preliminar mejorando a esta dimensión en un 5,64% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 97,75% obtenido en el análisis preliminar mejorando esta dimensión en 2,25% dejando a la tabla con un mejor nivel de conformidad de datos.

Análisis de la calidad de datos para la tabla Evaluaciones:
En los datos encontrados en el analisis preliminar para la tabla Evaluaciones se evidencio:

Parametro de Precisión:
En esta tabla no se encontraron registros con valores nulos.

Análisis de resultados para la precisión.

En el análisis de los datos para la tabla Evaluaciones de la base de datos OASIS en el parámetro de Precisión se puede evidenciar que no se ha encontrado registros nulos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Valores Aceptables

En esta tabla no se encontraron registros con valores inválidos.

Análisis de resultados para los Valores Aceptables

En el análisis de los datos para la tabla Evaluaciones de la base de datos OASIS en el parámetro de Valores Aceptables se puede evidenciar que no se ha encontrado registros inválidos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Duplicidad

En esta tabla no se encontraron registros duplicados para esta tabla.

Análisis de resultados para la Duplicidad.

En el análisis de los datos para la tabla Evaluaciones de la base de datos OASIS en el parámetro de Duplicidad se puede evidenciar que no se ha encontrado registros inválidos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Aquí utilizamos las formulas planteadas con anterioridad.

Tabla 42. Parámetro de Confianza
Fuente: Investigador

TABLA EVALUACIONES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
sintCodMatricula	100%	100%	100%
strCodPeriodo	100%	100%	100%
strCodMateria	100%	100%	100%
bytNota1	100%	100%	100%
bytNota2	100%	100%	100%
bytNota3	100%	100%	100%
strObservaciones	100%	100%	100%

Resultados finales.

Tabla 43. Resultados Finales
Fuente: Investigador

TABLA EVALUACIONES	
DIMENSIÓN	TOTAL
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

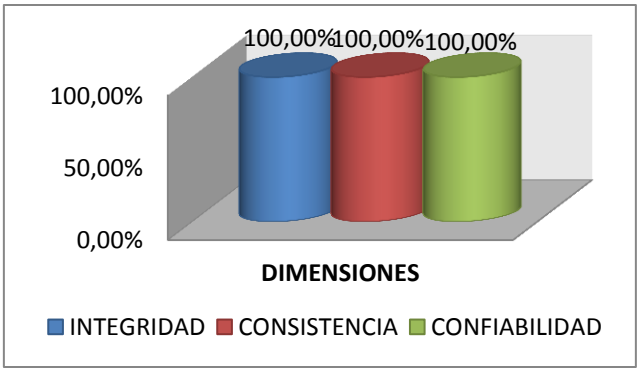


Figura 153. Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.

Como se puede observar en la tabla y en la gráfica anterior las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla Evaluaciones, con un nivel de integridad del 100% al igual que en el análisis preliminar de los datos en el cual se obtuvo 100%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación al 92,86% obtenido en el análisis preliminar mejorando a esta dimensión en un 7,14% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 96,43% obtenido en el análisis preliminar mejorando esta dimensión en 3,57% dejando a la tabla con un mejor nivel de conformidad de datos.

Análisis de la calidad de datos para la tabla Notas_Examenes:

En los datos encontrados en el analisis preliminar para la tabla Notas_Examenes se evidencio:

Parametro de Precisión:

En esta tabla no se encontraron registros con valores nulos.

Análisis de resultados para la precisión.

En el análisis de los datos para la tabla Notas_Examenes de la base de datos OASIS en el parámetro de Precisión se puede evidenciar que no se ha encontrado registros nulos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Valores Aceptables

En esta tabla no se encontraron registros con valores inválidos.

Análisis de resultados para los Valores Aceptables.

En el análisis de los datos para la tabla Notas_Examenes de la base de datos OASIS en el parámetro de Valores Aceptables se puede evidenciar que no se ha encontrado registros inválidos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Duplicidad

En esta tabla no se encontraron registros duplicados para esta tabla.

Análisis de resultados para la Duplicidad.

En el análisis de los datos para la tabla Notas_Examenes de la base de datos OASIS en el parámetro de Duplicidad se puede evidenciar que no se ha encontrado registros inválidos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Aquí utilizamos las formulas planteadas con anterioridad.

Tabla 44. Parámetro de Confianza
Fuente: Investigador

TABLA NOTAS_EXAMENES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
sintCodMatricula	100%	100%	100%
strCodPeriodo	100%	100%	100%
strCodMateria	100%	100%	100%
strCodTipoExamen	100%	100%	100%
bytAcumulado	100%	100%	100%
bytNota	100%	100%	100%
strCodEquiv	100%	100%	100%
strObservaciones	100%	100%	100%

Resultados finales.

Tabla 45. Resultados Finales
Fuente: Investigador

TABLA NOTAS_EXAMENES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

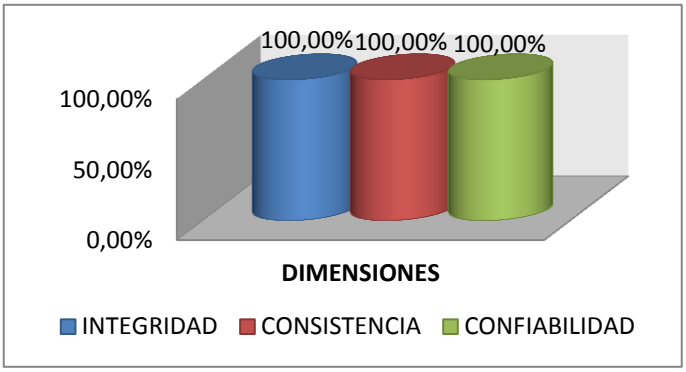


Figura 154 . Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.

Como se puede observar en la tabla y en la gráfica anterior las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla Notas_Examenes, con un nivel de integridad del 100% al igual que en el análisis preliminar de los datos en el cual se obtuvo 100%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación al 93,98% obtenido en el análisis preliminar mejorando a esta dimensión en un 6,02% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 97,59% obtenido en el análisis preliminar mejorando esta dimensión en 2,41% dejando a la tabla con un mejor nivel de conformidad de datos.

Análisis de la calidad de datos para la tabla Periodos:

En los datos encontrados en el analisis preliminar para la tabla Periodos se evidencio:

Parametro de Precisión:

En esta tabla no se encontraron registros con valores nulos.

Análisis de resultados para la precisión.

En el análisis de los datos para la tabla Periodos de la base de datos OASIS en el parámetro de Precisión se puede evidenciar que no se ha encontrado registros nulos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Valores Aceptables

En esta tabla no se encontraron registros con valores inválidos.

Análisis de resultados para los Valores Aceptables.

En el análisis de los datos para la tabla Periodos de la base de datos OASIS en el parámetro de Valores Aceptables se puede evidenciar que no se ha encontrado registros inválidos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Duplicidad

En esta tabla no se encontraron registros duplicados para esta tabla.

Análisis de resultados para la Duplicidad.

En el análisis de los datos para la tabla Periodos de la base de datos OASIS en el parámetro de Duplicidad se puede evidenciar que no se ha encontrado registros inválidos lo cual nos indica que para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Aquí utilizamos las formulas planteadas con anterioridad.

Tabla 46. Parámetro de Confianza
Fuente: Investigador

TABLA PERIODOS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
strCodigo	100%	100%	100%
strDescripcion	100%	100%	100%
dtFechaInic	100%	100%	100%
dtFechaFin	100%	100%	100%
sintUltNumMat	100%	100%	100%
strCodPensum	100%	100%	100%
blnTransicion	100%	100%	100%
blnVigente	100%	100%	100%
dtFechaTopeMatOrd	100%	100%	100%
dtFechaTopeMatExt	100%	100%	100%
dtFechaTopeMatPro	100%	100%	100%
dtFechaTopeRetMat	100%	100%	100%
strCodReglamento	100%	100%	100%

Resultados finales.

Tabla 47 Resultados Finales
Fuente: Investigador

TABLA PERIODOS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	100%
CONFIABILIDAD	100%

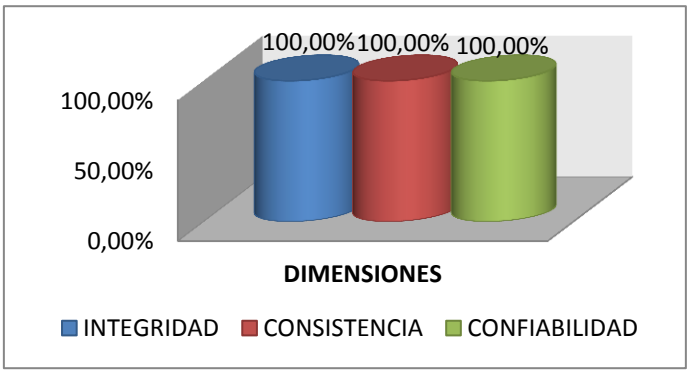


Figura 155. Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.

Como se puede observar en la tabla y en la gráfica anterior las dimensiones de calidad de datos fueron afectadas de manera favorable dándole un nivel de conformidad mucho más alta a la tabla Periodos, con un nivel de integridad del 100% al igual que en el análisis preliminar de los datos en el cual también se obtuvo 100%, además de la dimensión de consistencia la cual se obtuvo un 100% después de la limpieza en comparación con el 87,36% obtenido en el análisis preliminar mejorando a esta dimensión en un 12,64% y en la dimensión de confiabilidad de esta tabla obteniendo 100% después de la limpieza en comparación con el 94,94% obtenido en el análisis preliminar mejorando esta dimensión en 5,06% dejando a la tabla con un mejor nivel de conformidad de datos.

ANEXO III

ANALISIS PRELIMINAR DE LA BASE DE DATOS OASIS

ANALISIS PRELIMINAR DE LA CALIDAD DE DATOS DEL OASIS.

- **CESTUD**

Análisis preliminar de la tabla CESTUD usando la herramienta DataCleaner.

Valores NULL

Tabla 48 Valores Nulos
Fuente: Investigador

TABLA CESTUD		
Columnas	**Cantidad de NULL**	**Porcentaje de NULL**
FECING	16	0,44%
NOMPAD	44	1,22%
LUGNAC	39	1.08%
COLGRA	35	0.97%
BACHIL	39	1.08%
FECGRA	239	6,64%
CEDIDE	86	2,39%
CERMIL	1996	55,49%
NUMTEF	2157	59,97%
DIRPER	931	25,88%
DIRCUR	3108	86,41%
DOCUME	3589	99,78%

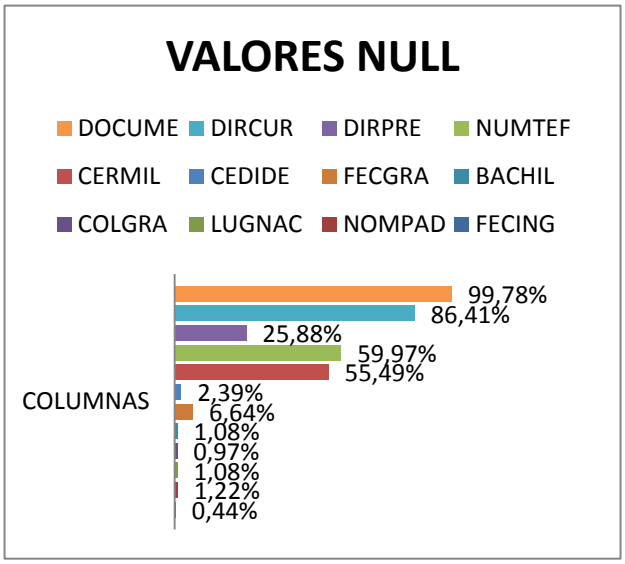


Figura 156 Valores Nulos
Fuente: Investigador

Valores Inválidos

Tabla 49 Valores Inválidos
Fuente: Investigador

TABLA CESTUD		
Columnas	**Cantidad de Inválidos**	**Porcentaje de Inválidos**
APEEST	14	0,39%
BACHIL	1	0,03%
CEDIDE	8	0,22%

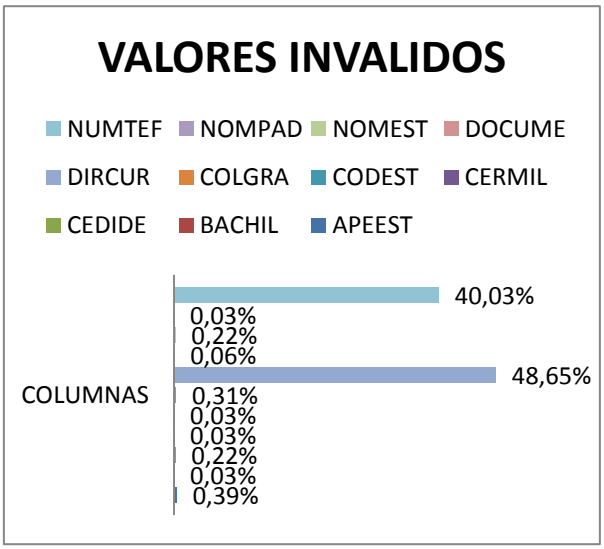


Figura 157 Valores Inválidos
Fuente: Investigador

CERMIL	1	0,03%
CODEST	1	0,03%
COLGRA	11	0,31%
DIRCUR	1750	48,65%
DOCUME	2	0,06%
NOMEST	8	0,22%
NOMPAD	1	0,03%
NUMTEF	1440	40,03%

Valores Duplicados

Tabla 50 Valores Duplicados
Fuente: Investigador

TABLA CESTUD		
Columnas	**Cantidad De Duplicados**	**Porcentaje de Duplicados**
CEDIDE	36	1%
CODEST	32	0,89%
DIRCUR	36	1%
NOMEST	36	1%
NOMPAD	34	0,95%
NUMTEF	29	0,81%

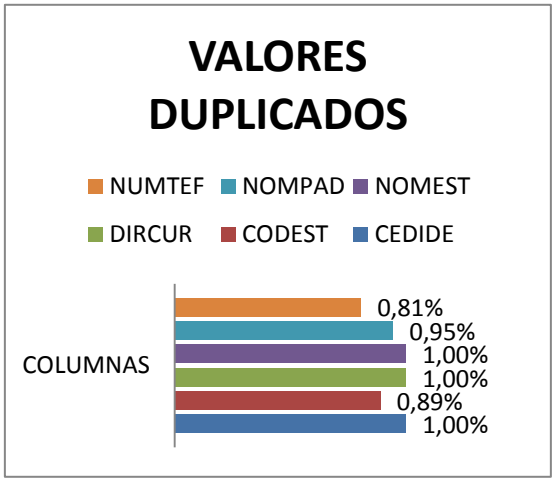


Figura 158 Valores Duplicados
Fuente: Investigador

Análisis de resultados

En las tablas y en los gráficos mostrados anterior mente podemos apreciar el análisis realizado para la tabla CESTUD en la cual observamos varias inconsistencias tales como valores nulos para algunas columnas así como valores inválidos dentro de otras columnas además de encontrar registros duplicados los cuales se mejoraran en el transcurso del desarrollo de este trabajo de limpieza de datos.

- **ESTUDIANTES**

Análisis preliminar de la tabla Estudiantes usando la herramienta DataCleaner.

Valores NULL

Tabla 51 Valores Nulos
Fuente: Investigador

TABLA ESTUDIANTES		
Columnas	**Cantidad de NULL**	**Porcentaje de NULL**
FECING	29	1,27%
strMail	1195	52,27%
strDocumentacion	2286	100%
strCodTit	1127	49,30%
strCedulaMil	39	1,71%
strCodInt	1127	49,30%

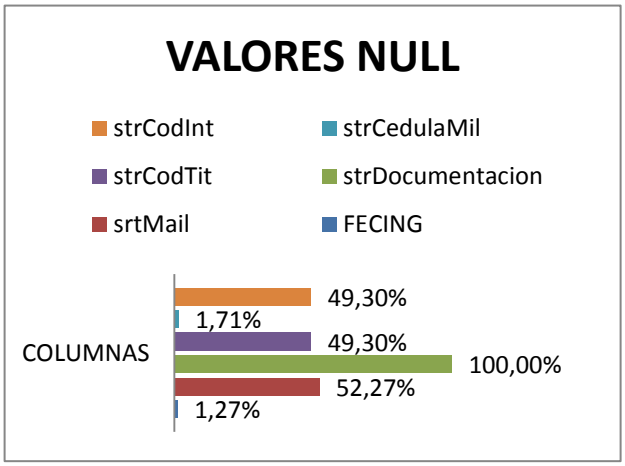


Figura 159 Valores Nulos
Fuente: Investigador

Valores Inválidos

Tabla 52 Valores Inválidos
Fuente: Investigador

TABLA ESTUDIANTES		
Columnas	**Cantidad Valores Inválidos**	**Porcentaje Valores Inválidos**
strMail	11	0,48%
strNacionalidad	1	0,04%
strNombres	22	0,96%
strApellidos	23	1,01%

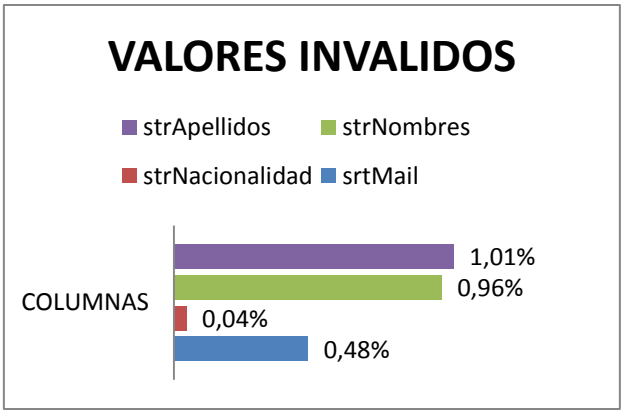


Figura 160 Valores Inválidos
Fuente: Investigador

Valores Duplicados

No se encontraron valores duplicados para esta tabla.

Análisis de resultados

En las tablas y gráficos mostrados con anterioridad para el análisis de la tabla de Estudiantes se puede observar que existen inconsistencias dentro de la tabla tales como valores nulos y valores inválidos dentro de los campos los cuales afectan a la calidad de datos y se procederán a ser limpiados en el transcurso de este trabajo.

- **DOCENTES**

Análisis preliminar de la tabla Docentes usando la herramienta DataCleaner.

Valores NULL

Tabla 53 Valores Nulos
Fuente: Investigador

TABLA DOCENTES		
Columnas	**Cantidad de NULL**	**Porcentaje de NULL**
strCedulaMil	93	93%
strCarnetSeg	95	95%
strDireccion	36	36%
strTel	42	42%
strMail	53	53%
strWww	59	59%
strCodTipoSan	100	100%
strCodEstCiv	2	2%
strTitulos	98	98%
strCargos	98	98%
strCodTipTit	10	10%

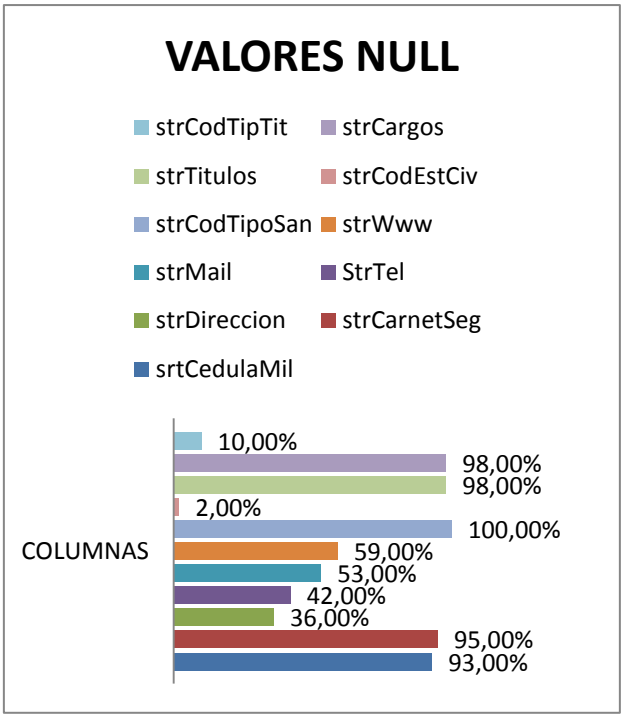


Figura 161 Valores Nulos
Fuente: Investigador

Valores Inválidos

Tabla 54 Valores Inválidos
Fuente: Investigador

TABLA DOCENTES		
Columnas	**Cantidad Inválidos**	**Porcentaje Inválidos**
strApellidos	1	0,48%
strDir	1	0,04%
strTel	3	0,96%

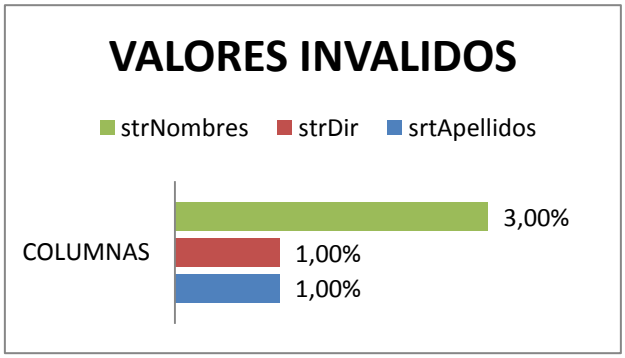


Figura 162 Valores Inválidos
Fuente: Investigador

Valores Duplicados

No se encontraron valores duplicados para esta tabla.

Análisis de resultados

Observando las tablas y gráficos anteriores se puede evidenciar el análisis realizado a la tabla de Docentes con el fin de conocer la calidad de los datos que aquí se presentan, encontrando como resultado valores nulos y registros inválidos los cuales se mejoran en el transcurso del desarrollo de este trabajo.

- **MATERIAS**

  Análisis preliminar de la tabla Materias usando la herramienta DataCleaner.

  Valores NULL

Tabla 55 Valores Nulos
Fuente: Investigador

TABLA MATERIAS		
Columnas	**Cantidad de NULL**	**Porcentaje de NULL**
srtCodigo	0	0%
strNombre	0	0%
dtFechaCreada	0	0%
dtFechaElim	130	58,56%
blnActiva	0	0%

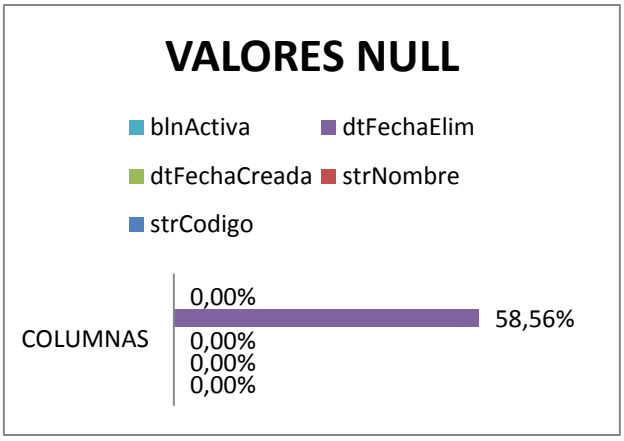


Figura 163 Valores Nulos
Fuente: Investigador

Valores Inválidos

La tabla contiene los valores aceptables esperados según los requerimientos del negocio.

Valores Duplicados

No se encontraron valores duplicados para esta tabla.

Análisis de resultados

En la tabla y en el gráfico mostrado anteriormente podemos apreciar el análisis realizado para la tabla Materias en la cual observamos pocas inconsistencias tales como valores nulos para algunas columnas, los cuales se mejoraran en el transcurso del desarrollo de este trabajo de limpieza de datos, no se presentaron valores duplicados ni valores inválidos en esta tabla.

- **MATRICULAS**

  Análisis preliminar de la tabla Matriculas usando la herramienta DataCleaner.

  Valores NULL

Tabla 56 Valores Nulos
Fuente: Investigador

TABLA MATRICULAS		
Columnas	**Cantidad de NULL**	**Porcentaje de NULL**
sintCodigo	0	0%
strCodPeriodo	0	0%
strCodEstud	0	0%
strCodNivel	0	0%
strAutorizadaPor	0	0%
dtFechaAutorizada	0	0%
strCreadaPor	0	0%
dtFechaCreada	0	0%
strCodEstado	0	0%
strObservaciones	4927	29,70%

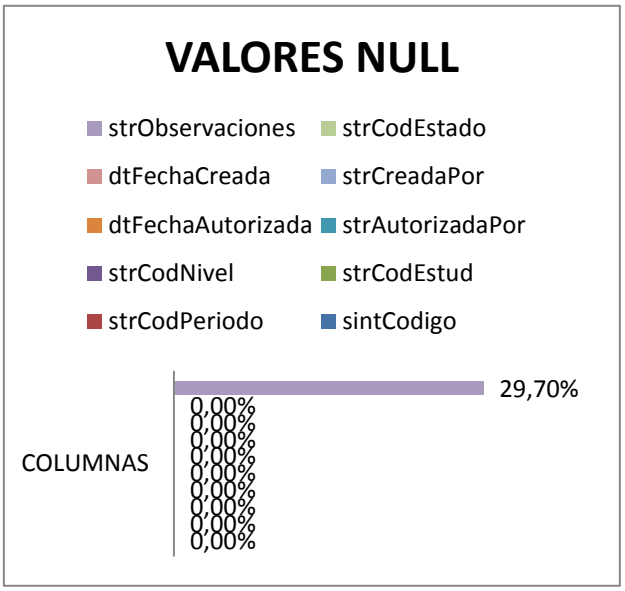
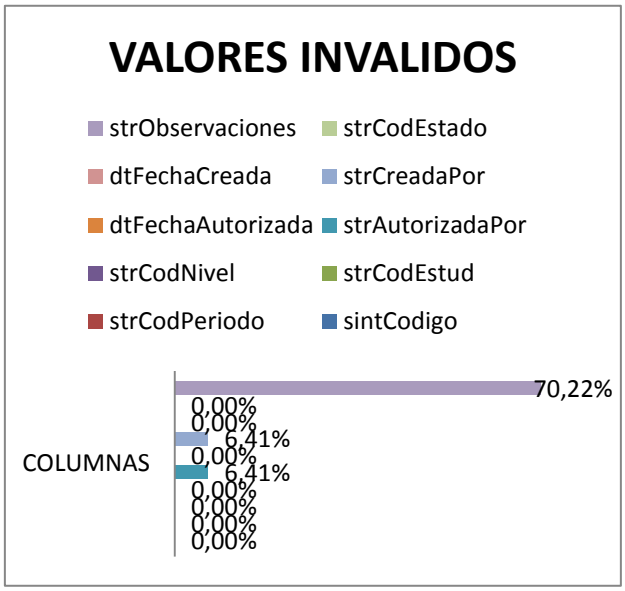


Figura 164 Valores Nulos
Fuente: Investigador

Valores Inválidos

Tabla 57 Valores Inválidos
Fuente: Investigador

TABLA MATRICULAS		
Columnas	**Cantidad Valores Invalidos**	**Porcentaje Valores Invalidos**
sintCodigo	0	0%
strCodPeriodo	0	0%
strCodEstud	0	0%
strCodNivel	0	0%
strAutorizadaPor	1063	6,41%
dtFechaAutorizada	0	0%
strCreadaPor	1064	6,41%
dtFechaCreada	0	0%
strCodEstado	0	0%
strObservaciones	11655	29,70%



Valor

Figura 165 Valores Inválidos
Fuente: Investigador

No se encontraron valores duplicados para esta tabla.

Análisis de resultados

En las tablas y en los gráficos mostrados anteriormente podemos apreciar el análisis realizado para la tabla Matriculas en la cual observamos pocas inconsistencias tales como valores nulos para algunas columnas y valores inválidos, los cuales se mejoraran en el transcurso del desarrollo de este trabajo de limpieza de datos, no se presentaron valores duplicados en esta tabla.

- **EVALUACIONES**

Análisis preliminar de la tabla Evaluaciones usando la herramienta DataCleaner.

Valores NULL

Tabla 58 Valores Nulos
Fuente: Investigador

TABLA EVALUACIONES		
Columnas	**Cantidad de NULL**	**Porcentaje de NULL**
sintCodMatricula	0	0%
strCodPeriodo	0	0%
strCodMateria	0	0%
bytNota1	0	0%
bytNota2	0	0%
bytNota3	0	0%
strObservaciones	905	5,08%

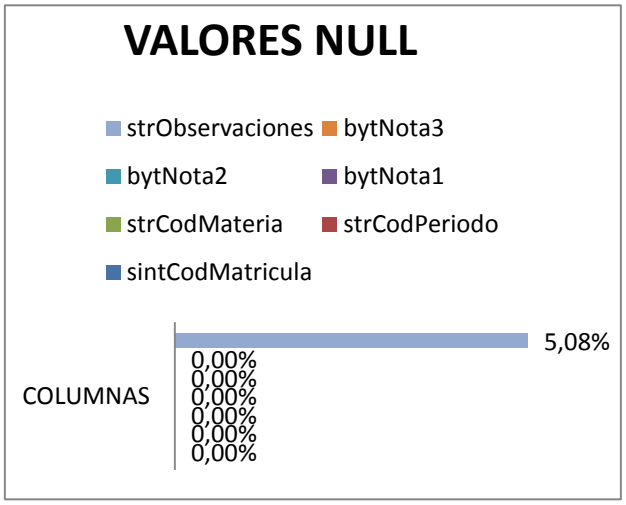


Figura 166 Valores Nulos
Fuente: Investigador

Valores Inválidos

Tabla 59 Valores Inválidos
Fuente: Investigador

TABLA EVALUACIONES		
Columnas	**Cantidad Valores Invalidos**	**Porcentaje Valores Invalidos**
sintCodMatricula	0	0%
strCodPeriodo	0	0%
strCodMateria	0	0%
bytNota1	0	0%
bytNota2	0	0%
bytNota3	0	0%
strObservaciones	16896	94,92%

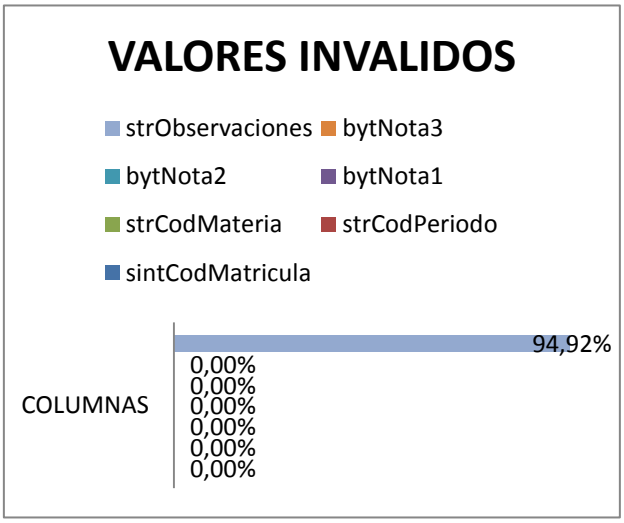


Figura 167 Valores Inválidos
Fuente: Investigador

Valores Duplicados

No se encontraron valores duplicados para esta tabla.

Análisis de resultados

En las tablas y en los gráficos mostrados anteriormente podemos apreciar el análisis realizado para la tabla Evaluaciones en la cual observamos pocas inconsistencias tales como valores nulos para algunas columnas y valores inválidos, los cuales se mejoraran en el transcurso del desarrollo de este trabajo de limpieza de datos, no se presentaron valores duplicados en esta tabla.

- **NOTAS_EXAMENES**

Análisis preliminar de la tabla Notas_Examenes usando la herramienta DataCleaner.

Valores NULL

Tabla 60 Valores Nulos
Fuente: Investigador

TABLA NOTAS_EXAMENES		
Columnas	**Cantidad de NULL**	**Porcentaje de NULL**
sintCodMatricula	0	0%
strCodPeriodo	0	0%
strCodMateria	0	0%
strCodTipoExamen	0	0%
bytAcumulado	0	0%
bytNota	1706	1,72%
strCodEquiv	0	0%
strObservaciones	50513	50,97%

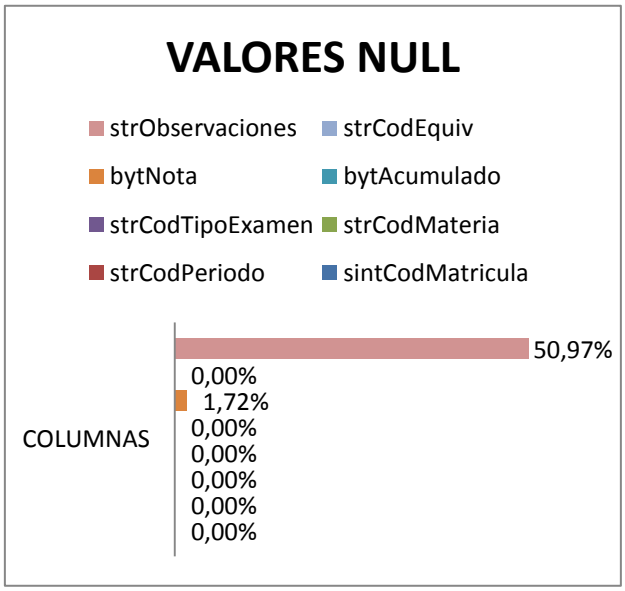


Figura 168 Valores Nulos
Fuente: Investigador

Valores Inválidos

Tabla 61 Valores Inválidos
Fuente: Investigador

TABLA NOTAS_EXAMENES		
Columnas	**Cantidad Valores Invalidos**	**Porcentaje Valores Invalidos**
sintCodMatricula	0	0%
strCodPeriodo	1	0%
strCodMateria	1	0%
strCodTipoExamen	1	0%
bytAcumulado	0	0%
bytNota	0	0%
strCodEquiv	0	0%
strObservaciones	43264	43,66%

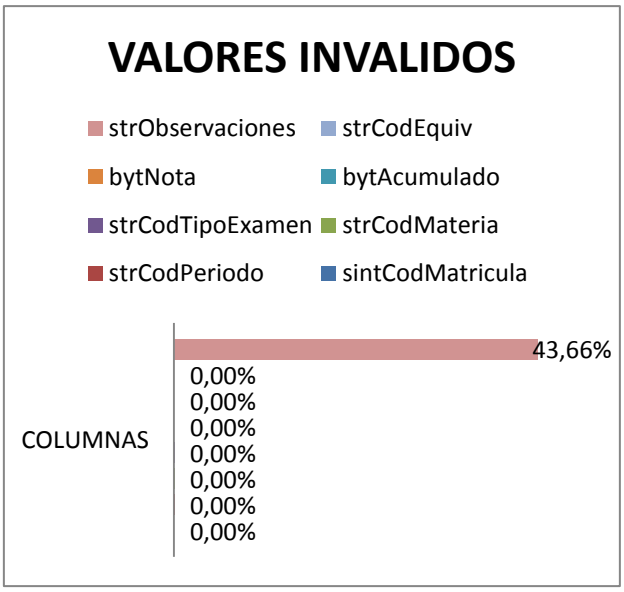


Figura 169 Valores Inválidos
Fuente: Investigador

Valores Duplicados

No se encontraron valores duplicados para esta tabla.

Análisis de resultados

En las tablas y en los gráficos mostrados anteriormente podemos apreciar el análisis realizado para la tabla Notas_Examenes en la cual observamos pocas inconsistencias tales como valores nulos para algunas columnas y valores inválidos, los cuales se mejoraran en el transcurso del desarrollo de este trabajo de limpieza de datos, no se presentaron valores duplicados en esta tabla.

- **PERIODOS**

Análisis preliminar de la tabla Periodos usando la herramienta DataCleaner.

Valores NULL

Tabla 62 Valores Nulos
Fuente: Investigador

TABLA PERIODOS		
Columnas	**Cantidad de NULL**	**Porcentaje de NULL**
strCodigo	0	0%
strDescripcion	0	0%
dtFechaInic	0	0%
dtFechaFin	0	0%
sintUltNumMat	0	0%
strCodPensum	0	0%
blnTransicion	0	0%
blnVigente	0	0%
dtFechaTopeMatOrd	46	82,14%
dtFechaTopeMatExt	46	82,14%
dtFechaTopeMatPro	46	82,14%
dtFechaTopeRetMat	46	82,14%
strCodReglamento	0	0%

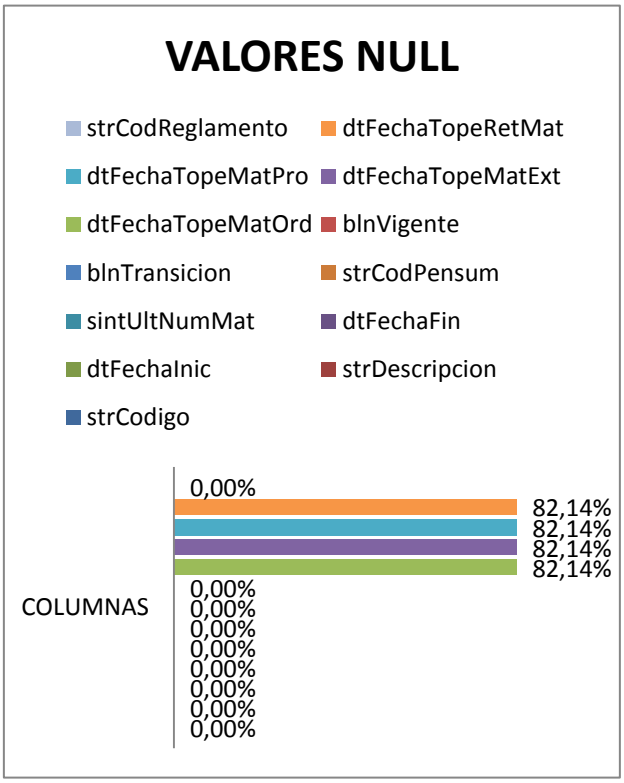


Figura 170 Valores Nulos
Fuente: Investigador

Valores Inválidos

La tabla contiene los valores aceptables esperados según los requerimientos del negocio.

Valores Duplicados

No se encontraron valores duplicados para esta tabla.

Análisis de resultados

En las tablas y en los gráficos mostrados anteriormente podemos apreciar el análisis realizado para la tabla Periodos en la cual observamos pocas inconsistencias tales como

valores nulos para algunas columnas, los cuales se mejoraran en el transcurso del desarrollo de este trabajo de limpieza de datos, no se presentaron valores duplicados ni valores inválidos en esta tabla.

ANÁLISIS DE LAS DIMENSIONES DE CALIDAD

Análisis de la calidad de datos para la tabla CESTUD:

En los datos encontrados en el analisis preliminar para la tabla CESTUD se evidencio:

Parametro de Precisión:

Tabla 63 Precisión
Fuente: Investigador

TABLA CESTUD		
COLUMNA	**Valores NULL**	**Porcentaje de Precisión**
FECING	16	99,56%
NOMPAD	44	98,78%
LUGNAC	39	98,92%
COLGRA	35	99,03%
BACHIL	39	98,92%
FECGRA	239	93,36%
CEDIDE	86	97,61%
CERMIL	1996	44,51%
NUMTEF	2157	40,03%
DIRPER	931	74,12%
DIRCUR	3108	13,59%
DOCUME	3589	0,22%

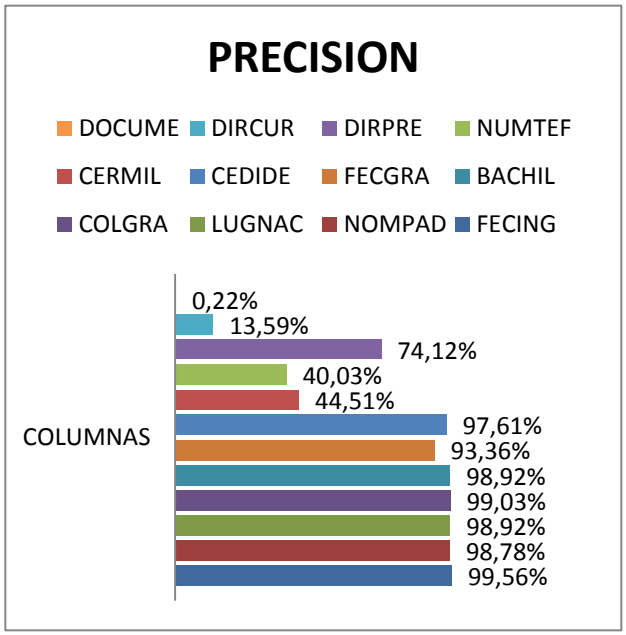


Figura 171 Precisión
Fuente: Investigador

Análisis de resultados para la precisión.

En los datos encontrados podemos evidenciar que en algunas columnas de la tabla CESTUD no se está cumpliendo con la dimensión de precisión en la cual se deberá mejorar para ciertos campos de esta tabla los cuales poseen un alto porcentaje de valores nulos.

Parámetro de Valores Aceptables

Tabla 64 Valores Aceptables
Fuente: Investigador

TABLA CESTUD		
Columnas	**Valores Inválidos**	**Porcentaje Valores Aceptables**
APEEST	14	99,61%
BACHIL	1	99,97%
CEDIDE	8	99,78%
CERMIL	1	99,97%
CODEST	1	99,97%
COLGRA	11	99,69%
DIRCUR	1750	51,35%
DOCUME	2	99,94%
NOMEST	8	99,78%
NOMPAD	1	99,97%
NUMTEF	1440	59,97%

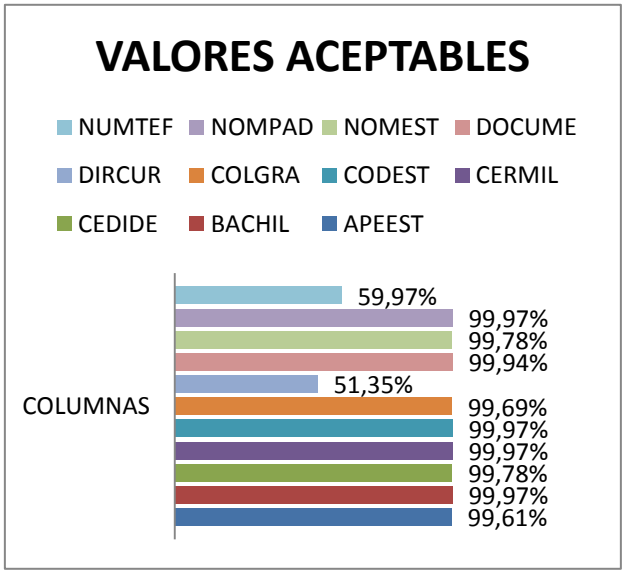


Figura 172 Valores Aceptables
Fuente: Investigador

Análisis de resultados para los valores válidos.

En el análisis realizado se ha evidenciado que en algunas columnas de la tabla CESTUD de la base de datos OASIS los valores ahí guardados no corresponden a la información que se debería almacenar según los requerimientos obtenidos se deberán mejorar los datos que ahí se encuentran.

Parámetro de Duplicidad

Tabla 65 Duplicidad
Fuente: Investigador

TABLA CESTUD		
Columnas	**Cantidad de Duplicados**	**Porcentaje de Duplicidad**
CEDIDE	36	99%
CODEST	32	99,11%
DIRCUR	36	99%

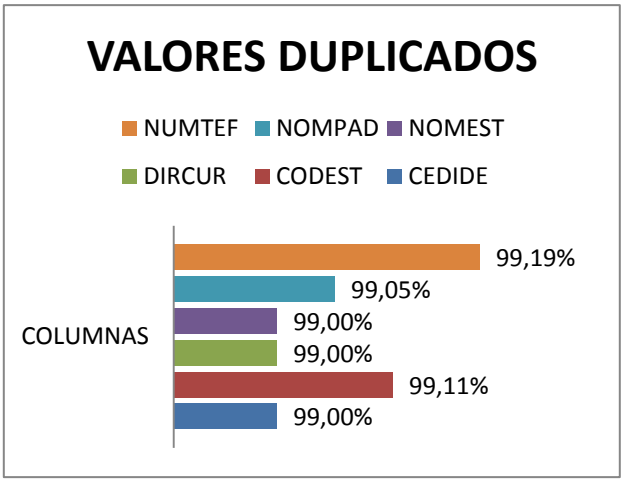


Figura 173 Duplicidad
Fuente: Investigador

NOMEST	36	99%
NOMPAD	34	99,05%
NUMTEF	29	99,19%

Análisis de resultados para la Duplicidad.

En el análisis preliminar de los datos para la tabla CESTUD de la base de datos OASIS en el parámetro de duplicidad se puede evidenciar que no se ha encontrado un gran número de registros duplicados pero que si existen los cuales deben ser limpiados con el fin de mejorar la calidad de los datos para esta tabla.

Parámetro de Confianza.

Aquí utilizamos las formulas planteadas con anterioridad.

Tabla 66 Resultados Finales
Fuente: Investigador

TABLA CESTUD			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
APEEST	99%	99,81%	99,32%
FECING	99%	99,45%	99,18%
NOMPAD	99,05%	98,70%	98,91%
LUGNAC	99%	72,24%	88,30%
COLGRA	99%	99,99%	99,40%
BACHIL	99%	99,36%	99,14%
FECGRA	99%	32,47%	72,39%
CEDIDE	99%	50,08%	79,43%
CERMIL	99%	99,89%	99,36%
NUMTEF	99,19%	99,38%	99,27%
DIRPER	99%	50,00%	79,40%
DIRCUR	99%	87,06%	94,22%
DOCUME	99%	99,46%	99,18%
CODEST	99,11%	99,81%	99,39%

Resultados finales.

Tabla 67 Resultados Finales
Fuente: Investigador

TABLA CESTUD	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	99,03%
CONSISTENCIA	84,84%
CONFIABILIDAD	93,35%

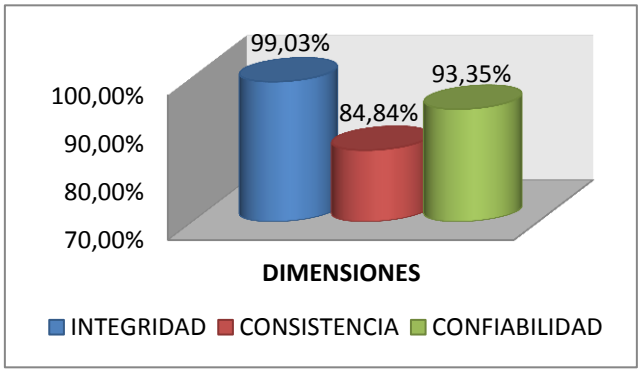


Figura 174 Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla CESTUD se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos pero los cuales se podrían mejorar con el uso de las herramientas dedicadas o esta función.

Análisis de la calidad de datos para la tabla Docentes:

En los datos encontrados en el analisis preliminar para la tabla Docentes se evidencio:

Parametro de Precisión

Tabla 68 Precisión
Fuente: Investigador

TABLA DOCENTES		
Columnas	**Valores NULL**	**Porcentaje de Precisión**
strCedulaMil	93	7%
strCarnetSeg	95	5%
strDireccion	36	64%
strTel	42	58%
strMail	53	47%
strWww	59	41%
strCodTipoSan	100	0,00%
strCodEstCiv	2	98%
strTitulos	98	2%
strCargos	98	2%
strCodTipTit	10	90%

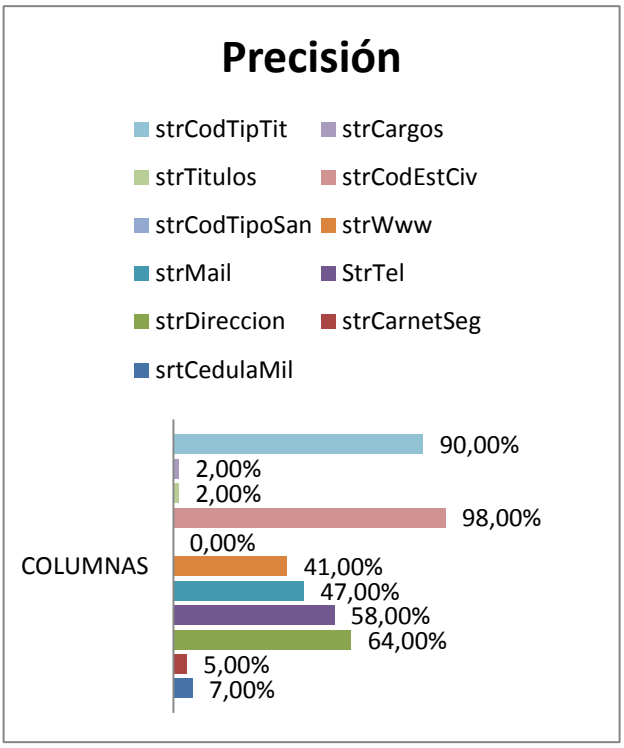


Figura 175 Precisión
Fuente: Investigador

Análisis de resultados para la precisión.

En los datos encontrados podemos evidenciar que en algunas columnas de la tabla Docentes no se está cumpliendo con la dimensión de precisión en la cual se deberá mejorar para ciertos campos de esta tabla los cuales poseen un alto porcentaje de valores nulos.

Parámetro de Valores Aceptables

Tabla 69 Valores Aceptables
Fuente: Investigador

TABLA DOCENTES		
Columnas	**Valores Inválidos**	**Porcentaje Valores Aceptables**
strMail	11	99,52%
strNacionalidad	1	99,96%
strNombres	22	99,04%
strApellidos	23	98,99%

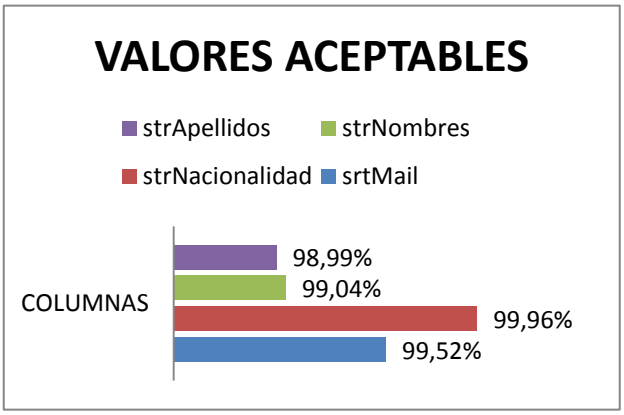


Figura 176 Valores Aceptables
Fuente: Investigador

Análisis de resultados para los valores válidos.

En el análisis realizado se ha evidenciado que en algunas columnas de la tabla Docentes de la base de datos OASIS los valores ahí guardados no corresponden a la información que se debería almacenar según los requerimientos obtenidos se deberán mejorar los datos que ahí se encuentran.

Parámetro Duplicidad

En esta tabla no se encontraron registros duplicados.

Análisis de resultados para la Duplicidad.

En el análisis preliminar de los datos para la tabla Docentes de la base de datos OASIS en el parámetro de duplicidad se puede evidenciar que no se ha encontrado registros duplicados lo cual nos dice que para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Aquí utilizamos las formulas planteadas con anterioridad.

Tabla 70 Resultados Finales
Fuente: Investigador

TABLA DOCENTES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
strCedulaMil	100%	53,50%	81,40%
strCarnetSeg	100%	52,50%	81,00%
strDireccion	100%	82,00%	92,80%
strTel	100%	79,00%	91,60%
strMail	100%	73,26%	89,30%
strWww	100%	70,50%	88,20%
strCodTipoSan	100%	50,00%	80,00%
strCodEstCiv	100%	99,00%	99,60%
strTitulos	100%	51,00%	80,40%
strCargos	100%	51,00%	80,40%
strCodTipTit	100%	95,00%	98,00%
strNacionalidad	100%	99,98%	99,99%
strNombres	100%	99,52%	99,81%
strApellidos	100%	99,50%	99,80%

Resultados finales.

Tabla 71 Resultados Finales
Fuente: Investigador

TABLA DOCENTES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	75,41%
CONFIABILIDAD	90,16%

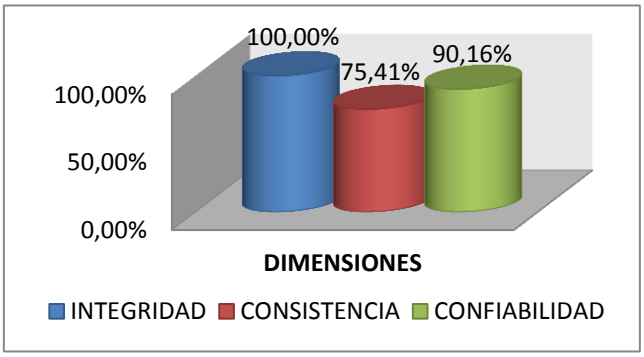


Figura 177 Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla Docentes se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos pero los cuales se podrían mejorar con el uso de las herramientas dedicadas o esta función en especial la dimensión de consistencia.

Análisis de la calidad de datos para tabla Estudiantes:

En los datos encontrados en el analisis preliminar para la tabla Estudiantes se evidencio:

Parametro de Precisión

Tabla 72 Precisión
Fuente: Investigador

TABLA ESTUDIANTES		
Columnas	**Valores NULL**	**Porcentaje de Precisión**
FECING	29	98,73%
strMail	1195	47,73%
strDocumentacion	2286	0,00%
strCodTit	1127	50,70%
strCedulaMil	39	98,29%
strCodInt	1127	50,70%

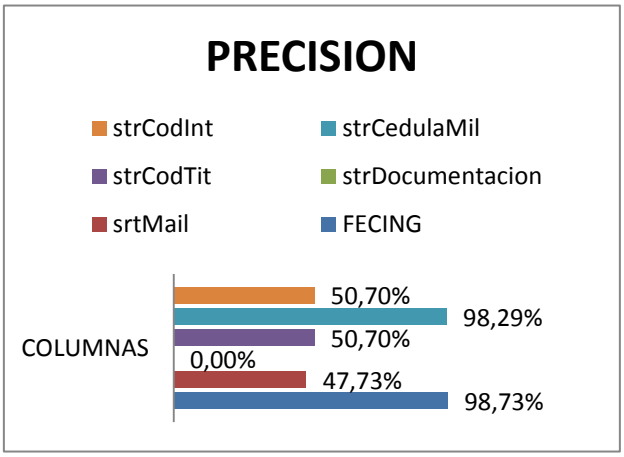


Figura 178 Precisión
Fuente: Investigador

Análisis de resultados para la precisión.

En los datos encontrados podemos evidenciar que en algunas columnas de la tabla Estudiantes no se está cumpliendo con la dimensión de precisión en la cual se deberá mejorar para ciertos campos de esta tabla los cuales poseen un alto porcentaje de valores nulos.

Parámetro de Valores Aceptables

Tabla 73 Valores Aceptables
Fuente: Investigador

TABLA ESTUDIANTES		
Columnas	**Valores Inválidos**	**Porcentaje Valores Aceptables**
strMail	11	99,52%
strNacionalidad	1	99,96%
strNombres	22	99,04%
strApellidos	23	98,99%

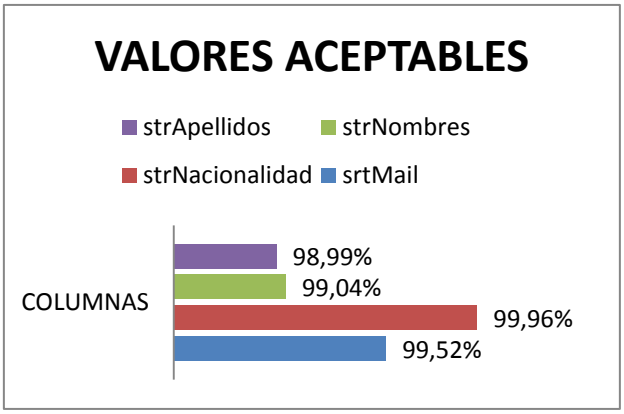


Figura 179 Valores Aceptables
Fuente: Investigador

Análisis de resultados para los Valores Aceptables.

En el análisis realizado se ha evidenciado que en algunas columnas de la tabla Estudiantes de la base de datos OASIS los valores ahí guardados no corresponden a la información que se debería almacenar según los requerimientos obtenidos se deberán mejorar los datos que ahí se encuentran.

Parámetro Duplicidad

En esta tabla no se encontraron registros duplicados.

Análisis de resultados para la Duplicidad.

En el análisis preliminar de los datos para la tabla Estudiantes de la base de datos OASIS en el parámetro de duplicidad se puede evidenciar que no se ha encontrado registros duplicados lo cual nos dice que para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Aquí utilizamos las formulas planteadas con anterioridad.

Tabla 74 Resultados Finales
Fuente: Investigador

TABLA ESTUDIANTES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
FECING	100%	99,37%	99,75%
strMail	100%	73,63%	89,45%
strDocumentacion	100%	50,00%	80,00%
strCodTit	100%	75,35%	90,14%
strCedulaMil	100%	99,15%	99,66%
strCodInt	100%	75,35%	90,14%
strNacionalidad	100%	99,98%	99,99%
strNombres	100%	99,52%	99,81%
strApellidos	100%	99,50%	99,80%

Resultados finales.

Tabla 75 Resultados Finales
Fuente: Investigador

TABLA ESTUDIANTES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	85,76%
CONFIABILIDAD	94,30%

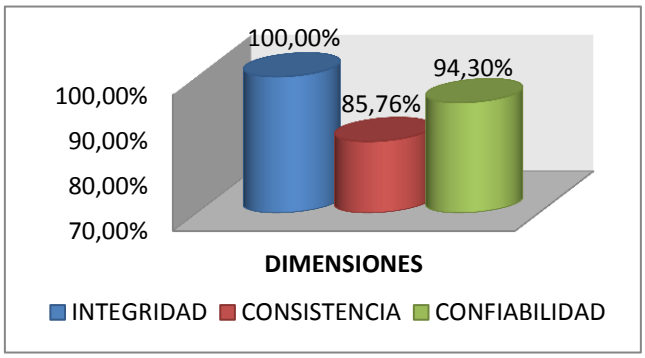


Figura 180 Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla Estudiantes se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos pero

los cuales se podrían mejorar con el uso de las herramientas dedicadas o esta función en especial la dimensión de consistencia.

Análisis de la calidad de datos para tabla Materias:

En los datos encontrados en el analisis preliminar para la tabla Materias se evidencio:

Parametro de Precisión:

Tabla 76 Precisión
Fuente: Investigador

TABLA MATERIAS		
COLUMNA	**Valores NULL**	**Porcentaje de Precisión**
srtCodigo	0	100%
strNombre	0	100%
dtFechaCreada	0	100%
dtFechaElim	130	41,44%
blnActiva	0	100%

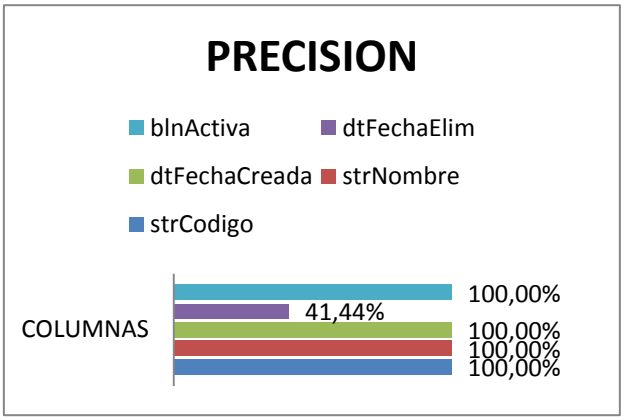


Figura 181 Precisión
Fuente: Investigador

Análisis de resultados para la precisión.

En los datos encontrados podemos evidenciar que en algunas columnas de la tabla Materias no se está cumpliendo con la dimensión de precisión en la cual se deberá mejorar para ciertos campos de esta tabla los cuales poseen un alto porcentaje de valores nulos.

Parámetro de Valores Aceptables

En esta tabla no se encontraron registros con valores inválidos.

Análisis de resultados para los valores válidos.

En el análisis preliminar de los datos para la tabla Materias de la base de datos OASIS en el parámetro de Valores Aceptables se puede evidenciar que no se ha encontrado registros inválidos lo cual nos dice que para este parámetro su cumplimiento es del 100%.

Parámetro de Duplicidad

En esta tabla no se encontraron registros duplicados para esta tabla.

Análisis de resultados para la Duplicidad.

En el análisis preliminar de los datos para la tabla Materias de la base de datos OASIS en el parámetro de Duplicidad se puede evidenciar que no se ha encontrado registros inválidos lo cual nos dice que para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Aquí utilizamos las formulas planteadas con anterioridad.

Tabla 77 Resultados Finales
Fuente: Investigador

TABLA MATERIAS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
srtCodigo	100%	100%	100%
strNombre	100%	100%	100%
dtFechaCreada	100%	100%	100%
dtFechaElim	100%	70,72%	88,29%
blnActiva	100%	100%	100%

Resultados finales.

Tabla 78 Resultados Finales
Fuente: Investigador

TABLA MATERIAS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	94,14%
CONFIABILIDAD	97,66%

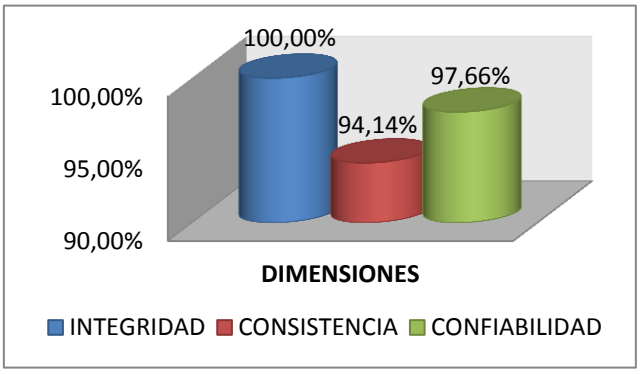


Figura 182 Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla Materias se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos pero los cuales se podrían mejorar con el uso de las herramientas dedicadas o esta función.

Análisis de la calidad de datos para la tabla Matriculas:

En los datos encontrados en el analisis preliminar para la tabla Matriculas se evidencio:

Parametro de Precisión:

Tabla 79 Precisión
Fuente: Investigador

TABLA MATRICULAS		
COLUMNA	**Valores NULL**	**Porcentaje de Precisión**
sintCodigo	0	100%
strCodPeriodo	0	100%
strCodEstud	0	100%
strCodNivel	0	100%
strAutorizadaPor	0	100%
dtFechaAutorizada	0	100%
strCreadaPor	0	100%
dtFechaCreada	0	100%
strCodEstado	0	100%
strObservaciones	4927	70,30%

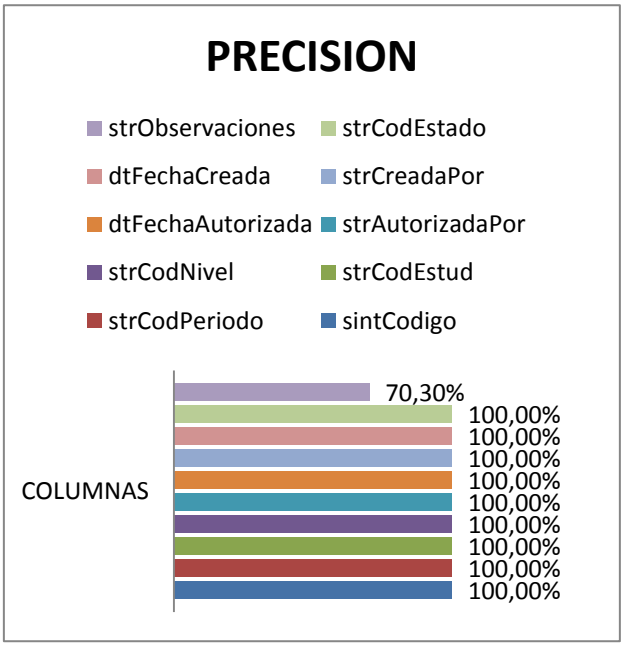


Figura 183 Precisión
Fuente: Investigador

Análisis de resultados para la precisión.

En los datos encontrados podemos evidenciar que en algunas columnas de la tabla Matriculas no se está cumpliendo con la dimensión de precisión en la cual se deberá mejorar para el campo de esta tabla el cual posee un alto porcentaje de valores nulos.

Parámetro de Valores Aceptables

Tabla 80 Valores Aceptables
Fuente: Investigador

TABLA MATRICULAS		
COLUMNA	**Valores Inválidos**	**Porcentaje Valores Aceptables**
sintCodigo	0	100%
strCodPeriodo	0	100%
strCodEstud	0	100%
strCodNivel	0	100%
strAutorizadaPor	1063	93,59%
dtFechaAutorizada	0	100%
strCreadaPor	1064	93,59%
dtFechaCreada	0	100%
strCodEstado	0	100%
strObservaciones	11655	29,76%

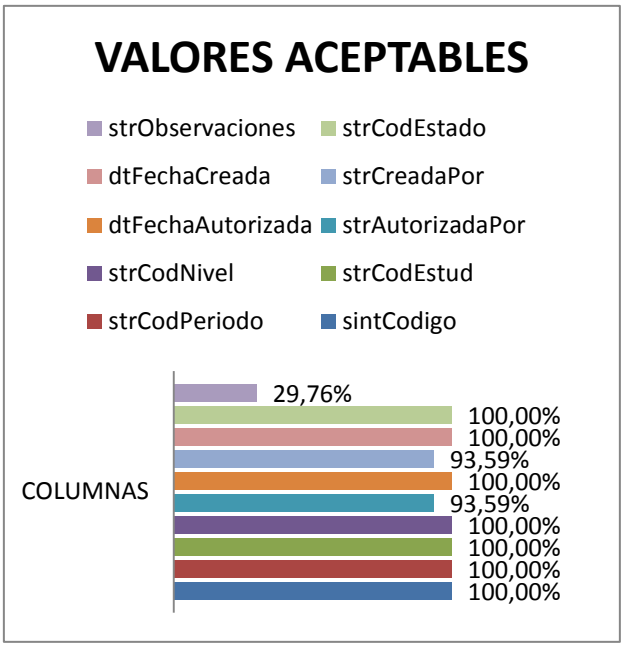


Figura 184 Valores Aceptables
Fuente: Investigador

Análisis de resultados para los valores válidos.

En el análisis realizado se ha evidenciado que en la tabla Matriculas de la base de datos OASIS los valores ahí guardados no cumplen con los estándares establecidos para ciertos campos de esta tabla.

Parámetro Duplicidad

En esta tabla no se encontraron registros duplicados para esta tabla.

Análisis de resultados para la Duplicidad.

En el análisis preliminar de los datos para la tabla Matriculas de la base de datos OASIS en el parámetro de duplicidad se puede evidenciar que no se ha encontrado registros duplicados lo cual nos dice que para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Aquí utilizamos las formulas planteadas con anterioridad.

Tabla 81 Resultados Finales
Fuente: Investigador

TABLA MATRICULAS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
sintCodigo	100%	100%	100%
strCodPeriodo	100%	100%	100%
strCodEstud	100%	100%	100%
strCodNivel	100%	100%	100%
strAutorizadaPor	100%	96,80%	98,72%
dtFechaAutorizada	100%	100%	100%
strCreadaPor	100%	96,80%	98,72%
dtFechaCreada	100%	100%	100%
strCodEstado	100%	100%	100%
strObservaciones	100%	50,03%	80,01%

Resultados finales.

Tabla 82 Resultados Finales
Fuente: Investigador

TABLA MATRICULAS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	94,36%
CONFIABILIDAD	97,75%

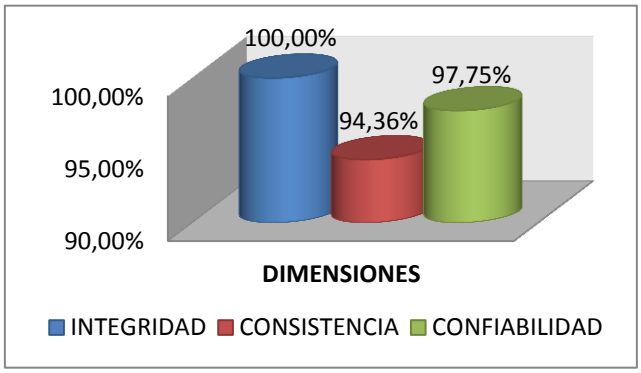


Figura 185 Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla Matriculas se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos pero los cuales se podrían mejorar con el uso de las herramientas dedicadas o esta función.

Análisis de la calidad de datos para la tabla Evaluaciones:

En los datos encontrados en el analisis preliminar para la tabla Evaluaciones se evidencio:

Parametro de Precisión:

Tabla 83 Precisión
Fuente: Investigador

TABLA EVALUACIONES		
COLUMNA	**Valores NULL**	**Porcentaje de Precisión**
sintCodMatricula	0	100%
strCodPeriodo	0	100%
strCodMateria	0	100%
bytNota1	0	100%
bytNota2	0	100%
bytNota3	0	100%
strObservaciones	905	94,92%

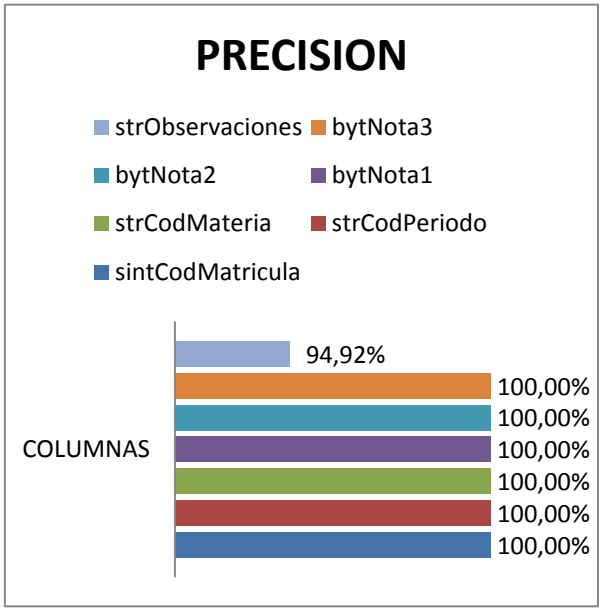


Figura 186 Precisión
Fuente: Investigador

Análisis de resultados para la precisión.

En los datos encontrados podemos evidenciar que en algunas columnas de la tabla Evaluaciones no se está cumpliendo con la dimensión de precisión en la cual se deberá mejorar para el campo de esta tabla el cual posee un alto porcentaje de valores nulos.

Parámetro de Valores Aceptables

Tabla 84 Valores Aceptables
Fuente: Investigador

TABLA EVALUACIONES		
COLUMNA	**Valores Inválidos**	**Porcentaje Valores Aceptables**
sintCodMatricula	0	100%
strCodPeriodo	0	100%
strCodMateria	0	100%
bytNota1	0	100%
bytNota2	0	100%
bytNota3	0	100%
strObservaciones	16896	5,08%

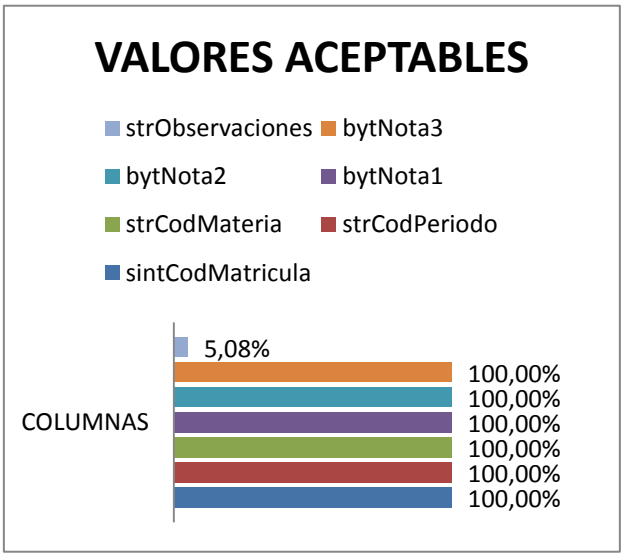


Figura 187 Valores Aceptables
Fuente: Investigador

Análisis de resultados para los valores Válidos.

En el análisis realizado se ha evidenciado que en la tabla Evaluaciones de la base de datos OASIS los valores ahí guardados no cumplen con los estándares establecidos para ciertos campos de esta tabla.

Parámetro Duplicidad

En esta tabla no se encontraron registros duplicados.

Análisis de resultados para la Duplicidad.

En el análisis preliminar de los datos para la tabla Evaluaciones de la base de datos OASIS en el parámetro de duplicidad se puede evidenciar que no se ha encontrado registros duplicados lo cual nos dice que para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Aquí utilizamos las formulas planteadas con anterioridad.

Tabla 85 Resultados Finales
Fuente: Investigador

TABLA EVALUACIONES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
sintCodMatricula	100%	100%	100%
strCodPeriodo	100%	100%	100%
strCodMateria	100%	100%	100%
bytNota1	100%	100%	100%
bytNota2	100%	100%	100%
bytNota3	100%	100%	100%
strObservaciones	100%	50%%	75%

Resultados finales.

Tabla 86 Resultados Finales
Fuente: Investigador

TABLA EVALUACIONES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	92,86%
CONFIABILIDAD	96,43%

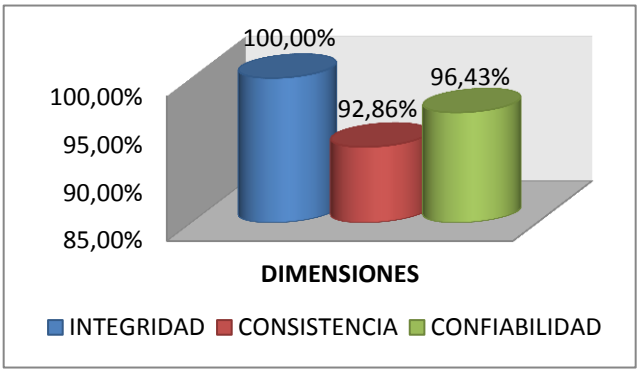


Figura 188 Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla Evaluaciones se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos pero los cuales se podrían mejorar con el uso de las herramientas dedicadas a esta función.

Análisis de la calidad de datos para la tabla Notas_Examenes:

En los datos encontrados en el analisis preliminar para la tabla Notas_Examenes se evidencio:

Parametro de Precisión:

Tabla 87 Precisión
Fuente: Investigador

TABLA NOTAS_EXAMENES		
COLUMNA	**Valores NULL**	**Porcentaje de Precisión**
sintCodMatricula	0	100%
strCodPeriodo	0	100%
strCodMateria	0	100%
strCodTipoExamen	0	100%
bytAcumulado	0	100%
bytNota	1706	98,28%
strCodEquiv	0	100%
strObservaciones	50513	49,03%

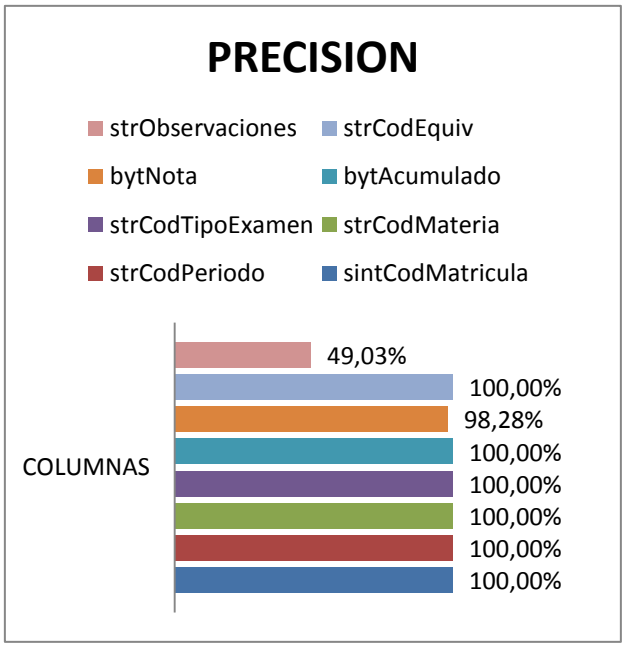


Figura 189 Precisión
Fuente: Investigador

Análisis de resultados para la precisión.

En los datos encontrados podemos evidenciar que en algunas columnas de la tabla Nota_Examenes no se está cumpliendo con la dimensión de precisión en la cual se deberá mejorar para el campo de esta tabla el cual posee un alto porcentaje de valores nulos.

Parámetro de Valores Aceptables

Tabla 88 Valores Aceptables
Fuente: Investigador

TABLA NOTAS_EXAMENES		
COLUMNA	**Valores Inválidos**	**Porcentaje Valores Aceptables**
sintCodMatricula	0	100%
strCodPeriodo	1	100%
strCodMateria	1	100%
strCodTipoExamen	1	100%
bytAcumulado	0	100%
bytNota	0	100%
strCodEquiv	0	5,08%
strObservaciones	43264	56,34%

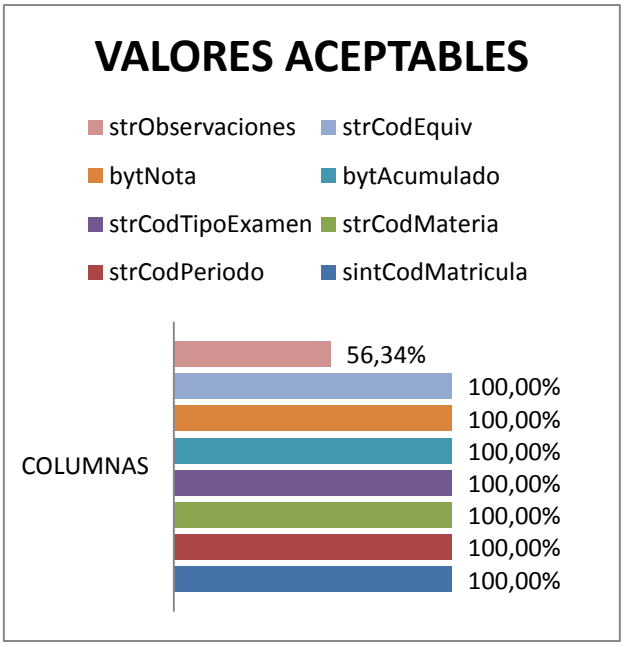


Figura 190 Valores Aceptables
Fuente: Investigador

Análisis de resultados para los valores válidos.

En el análisis realizado se ha evidenciado que en la tabla Notas_examenes de la base de datos OASIS los valores ahí guardados no cumplen con los estándares establecidos para ciertos campos de esta tabla.

Parámetro Duplicidad

En esta tabla no se encontraron registros duplicados.

Análisis de resultados para la Duplicidad.

En el análisis preliminar de los datos para la tabla Notas_examenes de la base de datos OASIS en el parámetro de duplicidad se puede evidenciar que no se ha encontrado registros duplicados lo cual nos dice que para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Aquí utilizamos las formulas planteadas con anterioridad.

Tabla 89 Resultados Finales
Fuente: Investigador

TABLA NOTAS_EXAMENES			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
sintCodMatricula	100%	100%	100%
strCodPeriodo	100%	100%	100%
strCodMateria	100%	100%	100%
strCodTipoExamen	100%	100%	100%
bytAcumulado	100%	100%	100%
bytNota	100%	99,14%	99,67%
strCodEquiv	100%	100%	100%
strObservaciones	100%	52,69%	81,08%

Resultados finales.

Tabla 90 Resultados Finales
Fuente: Investigador

TABLA NOTAS_EXAMENES	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	93,98%
CONFIABILIDAD	97,59%

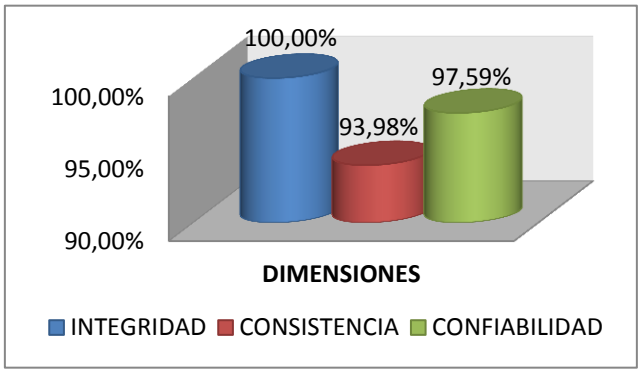


Figura 191 Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla Notas_Examenes se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos pero los cuales se podrían mejorar con el uso de las herramientas dedicadas a esta función.

Análisis de la calidad de datos para la tabla Periodos:

En los datos encontrados en el analisis preliminar para la tabla Periodos se evidencio:

Parametro de Precisión:

Tabla 91 Precisión
Fuente: Investigador

TABLA PERIODOS		
COLUMNA	**Valores Nulos**	**Porcentaje de Precisión**
strCodigo	0	100%
strDescripcion	0	100%
dtFechaInic	0	100%
dtFechaFin	0	100%
sintUltNumMat	0	100%
strCodPensum	0	100%
blnTransicion	0	100%
blnVigente	0	100%
dtFechaTopeMatOrd	46	17,86%
dtFechaTopeMatExt	46	17,86%
dtFechaTopeMatPro	46	17,86%
dtFechaTopeRetMat	46	17,86%
strCodReglamento	0	100%

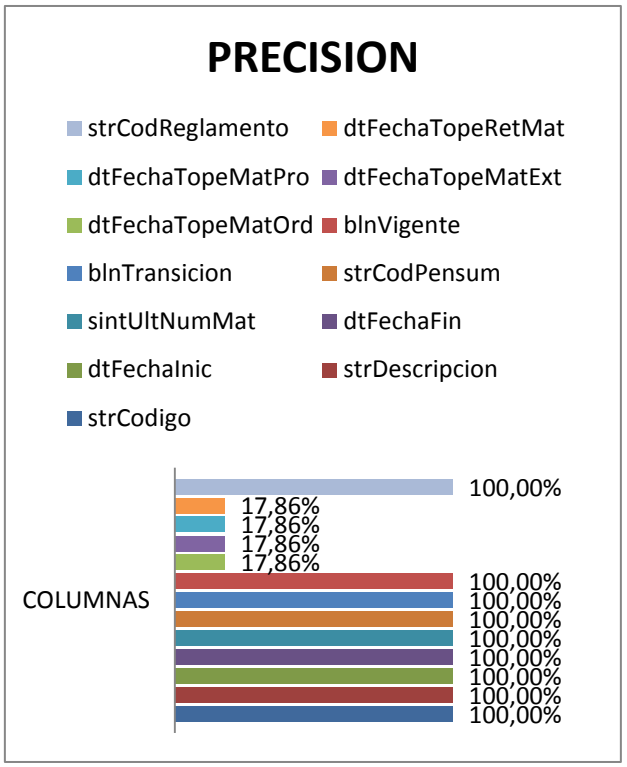


Figura 192 Precisión
Fuente: Investigador

Análisis de resultados para la precisión.

En los datos encontrados podemos evidenciar que en algunas columnas de la tabla Periodos no está cumpliendo con la dimensión de precisión en la cual se deberá mejorar para el campo de esta tabla el cual posee un alto porcentaje de valores nulos.

Parámetro de Valores Aceptables

En esta tabla no se encontraron registros con valores inválidos.

Análisis de resultados para los valores válidos.

En el análisis preliminar de los datos para la tabla Periodos de la base de datos OASIS en el parámetro de Valores Aceptables se puede evidenciar que no se ha encontrado registros inválidos lo cual nos dice que para este parámetro su cumplimiento es del 100%.

Parámetro Duplicidad

En esta tabla no se encontraron registros duplicados.

Análisis de resultados para la Duplicidad.

En el análisis preliminar de los datos para la tabla Periodos de la base de datos OASIS en el parámetro de duplicidad se puede evidenciar que no se ha encontrado registros duplicados lo cual nos dice que para este parámetro su cumplimiento es del 100%.

Parámetro de Confianza.

Aquí utilizamos las formulas planteadas con anterioridad.

Tabla 92 Resultados Finales
Fuente: Investigador

TABLA PERIODOS			
COLUMNA	**INTEGRIDAD**	**CONSISTENCIA** (Precisión + Valores Aceptables)/2	**CONFIABILIDAD** INTEGRIDAD(0,6) + CONSISTENCIA(0,4)
strCodigo	100%	100%	100%
strDescripcion	100%	100%	100%
dtFechaInic	100%	100%	100%
dtFechaFin	100%	100%	100%
sintUltNumMat	100%	100%	100%
strCodPensum	100%	100%	100%
blnTransicion	100%	100%	100%
blnVigente	100%	100%	100%
dtFechaTopeMatOrd	100%	58,93%	83,57%
dtFechaTopeMatExt	100%	58,93%	83,57%
dtFechaTopeMatPro	100%	58,93%	83,57%
dtFechaTopeRetMat	100%	58,93%	83,57%
strCodReglamento	100%	100%	100%

Resultados finales.

Tabla 93 Resultados Finales
Fuente: Investigador

TABLA PERIODOS	
DIMENSIÓN	**TOTAL**
INTEGRIDAD	100%
CONSISTENCIA	87,36%
CONFIABILIDAD	94,94%

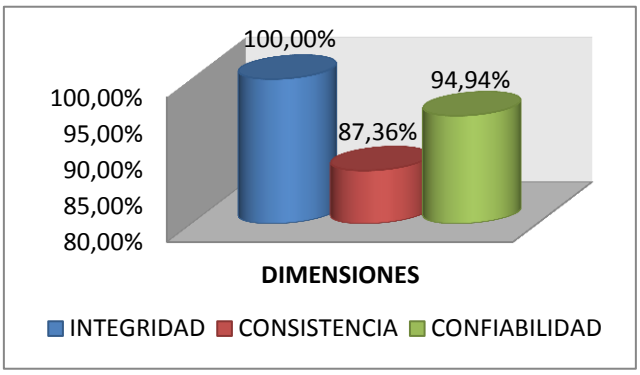


Figura 193 Resultados Finales
Fuente: Investigador

Interpretación de los resultados finales.

Después de realizar el análisis preliminar de la tabla Periodos se puede evidenciar que las dimensiones examinadas tienen un porcentaje alto de calidad dentro de sus datos pero los cuales se podrían mejorar con el uso de las herramientas dedicadas a esta función.